Of course! Here is a summary of the YouTube live session you provided.

This session was the first installment of a new live series called **"MLOps Production Grade Machine Learning Project"**. The goal of the series is to build an end-to-end machine learning pipeline, including deployment and pipeline management.

**📝 Project Overview**

* **Project:** **US Visa Approval Prediction**.
* **Problem Type:** This is a **classification** problem. The goal is to predict whether a visa application will be approved or denied based on various features like the applicant's continent, education, job experience, etc.
* **Objective:** To automate the visa evaluation process, which is often a manual and time-consuming task in embassies.

**📚 Prerequisites & Required Setup**

Before starting the project, the presenter outlined the necessary background knowledge and software setup.

**Prerequisites**

* **Python Programming:** Familiarity with Python, up to Object-Oriented Programming (OOP) concepts.
* **Machine Learning Basics:** Understanding of ML concepts, particularly classification algorithms.
* **AWS Account:** An AWS account will be needed for deployment later in the series.

**Software Setup 💻**

1. **Anaconda:** Used for managing Python environments. You can download it from the [official Anaconda website](https://www.anaconda.com/download).
2. **Git Bash:** A command-line tool for using Git. You can download it from [git-scm.com](https://git-scm.com/downloads).
3. **Visual Studio Code (VS Code):** The code editor used throughout the series. You can download it from the [official VS Code website](https://code.visualstudio.com/download).

**⚙️ Session 1: Project Initialization**

The main focus of this first session was to set up the foundational structure of the project.

**1. GitHub Repository**

A new **GitHub repository** was created to host all the project code. The repository was initialized with a README.md, a Python .gitignore file, and an MIT license.

**2. Project Folder Structure**

A Python script named template.py was written to **automatically generate** the entire project folder and file structure. This approach ensures consistency and saves time compared to creating folders manually. The script creates a modular structure, including directories for components, pipelines, configuration, and utilities.

**3. Environment and Dependencies**

The following steps were taken to set up the development environment:

* A new **Conda environment** was created using Python 3.8 with the command:

Bash

conda create -n visa python=3.8 -y

* A requirements.txt file was created, listing all necessary Python packages, including:
  + Data manipulation libraries (pandas, numpy).
  + ML libraries (scikit-learn, xgboost, catboost).
  + The MLOps tool **Evidently AI** for data drift detection.
  + **MongoDB** driver (pymongo).
  + AWS SDK (boto3).
  + Web framework (fastapi).
* A setup.py file was created. This allows the project's source code (the US\_Visa folder) to be installed as a **local package**. This is crucial for modular coding, enabling clean imports between different parts of the project.
* All dependencies were installed using:

Bash

pip install -r requirements.txt

**🚀 What's Next?**

The next session will focus on:

* Setting up **MongoDB Atlas** (a cloud database).
* Storing the dataset in MongoDB.
* Implementing the **Data Ingestion** component of the machine learning pipeline.

e e e uh hello everyone uh good evening I think I'm audible to all of

you uh hello everyone okay guys so I think I'm audible you can hear me you can see me so now I think we can start with the session so I'll wait uh one more minute uh to join everyone actually uh then I will start with the session for today so let's wait uh one minutes guys so that everyone can join and they can also um learn with me yeah hello uh hello faras good evening hello p uh kishan hello so I hope guys you are doing well yeah good evening good evening HP I don't know what's your name

sir give me the GitHub link for llama index video today um if you check out my GitHub you will get that will just check out my GitHub repository maybe somewhere it is there from there you can download otherwise you can write the code along with me because it's a very simple code I was writing right yeah hi vulp hi okay fine now I think we can start with the session guys uh sir my tab 9 is not working I mean not giving suggestion it should work because uh you have to install that particular extension on the vs code then uh I think

it should work just check it out I think there are some issue maybe installation was not correct you can reinstall it again I think this will work right can you tell me uh yeah because of some personal reason actually I left and uh than thank you thank you for your U uh congrats actually for my wedding yeah hi Aris uh yes I'm doing well what about you yes so now let's start with the session guys so before starting with the session first of all uh I will tell you why you will get the previous video and previous resources so for this just try to visit my channel guys so this is my

YouTube channel called DS with buy if you starts with this particular name you will get my Channel and here if you go to the playlist section so here I created one playlist called mlops production ready machine learning project okay so this particular playlist I have created now if I go to this particular playlist you will see uh yesterday we had a first session called project introduction and setup so if you have missed out uh this particular session so try to go ahead and watch this particular uh recording then you

can uh watch this particular uh session today because every session is related to each other right so if you missed out one particular session so it would be little bit difficult for you to understand the next session so I'll suggest everyone if you have missed out our day one so first of all try to complete this particular day one then uh come to the day to and try to watch this particular session okay yes and where you will get the resources resources wise I already updated in the description let me show you so if you go to the video description so here I already added this

particular GitHub link so I think it is available here so you can see this GitHub link is available here now you can open up this particular GitHub link and you can see all the resources I have shared yesterday okay so everything is available here got it now let me go backit good evening uh Mita all right so I think this part is clear like where you will get the previous resources and all the recording okay so everything is uh like there in the playlist section you just need to go ahead and watch this particular

recording okay yeah so now let me discuss the today's agenda like today what are the things we're going to cover in this particular uh session I think you already saw like today we'll be covering uh this particular database uh setup okay like how we can set up the database and how we can inest the data from the database and apart from that we'll be also learning how we can um like write the exception module then logging module and utility module that means the utility related functionality in our project okay so this thing we'll be covering in this particular session today so first of all I will open up my

Blackboard and here let me write down the agenda I'll be covering today so today's agenda so first thing we'll be doing database setup okay database setup so here database wise we'll be using something called mongodb okay mongod DB Atlas okay mongodb Atlas it's a cloud database okay we'll be using this particular mongodb Atlas and the second thing we'll be doing something called logging module okay loging module we'll be writing this particular logging module I'll tell you what is this particular logging module and why it is required for enter and implementation okay then the third thing

we'll be discussing about something called exception module so we'll be learning why exception is required and why you have to write your own custom exception okay for your code then we'll be also discussing about something called utility module uity module so I'll tell you what is utility module that means the functionality you will be using frequently in your code so you can write those functionality in the utility module okay instead of writing again and again so this is called utility module all right so this is what

actually we'll be covering today in this particular session now here um what would be the main actually um I mean uh concept like whenever I will be investing the data from my database so what would be the main concept there so the main concept would be let's say this is my database let's say this is my database my mongodb Atlas database okay so this is my database so this is my mongodb database so first of all what I have to do I have to set up this particular database definitely Because unless and

until youone not set uping the database you can't store the data there now here I will be having some data let's say here I'll will be having my Visa data okay my Visa data so this particular data where you will get you can download this particular data from the Kagel website okay Kagel has this particular data so you can download from the kagle website even I will also share with the with you this particular data I'll share in GitHub so from there you can also download this particular data now the first thing what I have to do okay the

first thing what I have to do I have to make the connection in my mongodb so we'll be making the connection in the mongodb and we'll be push the data to the mongodb that means we'll push the data we'll push the data to the database that means in my mongodb okay then what I will do okay then what I will do I'll just write another component called Data injection okay data inje data in so what this data in will do it will go to the mongodb that means to my database and it will fetch the data from the database okay it will face

the data from the database okay it will face the data from the database and here I will get my data that means again I will get my something called Visa data okay Visa data so this is what actually we'll be doing as an experiment first of all so everything will be writing on the Jupiter notebook today okay Jupiter notebook we are not going to write write any kinds of modular coding today okay modular coding we'll be doing once our experiment part is clear if it is working fine there is no error that time actually we'll be writing our modular coding but before starting with the modular coding I always suggest to my let's say learner

first of all try to do the notebook experiment okay first of all try to do the notebook experiment and if this particular notebook experiment is working fine then try to convert that particular notebook as a modular coding that means end to endend implementation okay that is the idea I hope this part is clear all of you right yeah so uh to work with the mongod DV first of all let's set up our account and there first of all I have to create the cluster and I have to do some additional setup there then I'll show you this particular connection and how we can push this particular data to the mongod and how we can fit that

particular data from the mongod itself every experiment I'll show you in this particular session itself okay so I think guys you are ready you are ready with your system if yes just give me a thumbs up in the chat guys all of you so give me a thumbs up in the chat or any kind of emoji you can send so that I can um get a confirmation you are ready and uh you can code with me let me know guys so instead of joining the session with your phone try to join from the system itself so that you can also do with me and I will share each and

everything uh in my GitHub so no need to worry you can pull the changes and you can get the update okay in your code as well so give me a like um any kinds of emoji in the chat any kinds of thumbs up okay anything f is a vector database okay so it's a machine learning project we are doing so here Vector database we are not going to use so this is what we'll be using whenever we'll be creating the generative way application very go okay I hope you got it so give me a thumbs up guys give me a thumbs up if you're ready so give me a thumbs up quickly

all right so now I think uh we can start so what I can do uh just a minute I can minimize this particular screen and I can go to the Google go okay so here first of all guys what you have to do you have to search for One name okay you have to search for One name called uh you have to search for One name called uh mongodb okay so let me search this particular name mongodb okay you can see that U DB okay mongod DB so I will share this particular link with you I think that would be fine so let me first of all um copy this particular link you can

see this is the link guys account. mongod dev.com okay so this is the link you have to visit so let me share this particular link with you so what I can do I can open up my code yesterday code so here I will open up my visual code Studio so now let me close these are the tab okay so here in the rme file I think I can add this particular link database mongodb okay mongodb database so here is the link guys okay now I think everything is fine um just a

minute now I can comment the changes so updated the link all right so link is updated now you can refresh my GitHub so let me share you my GitHub link as well here is the GitHub link okay so from here you can download this particular um code and everything okay so everything is up to date now I can refresh and I can show you so everything is up to date now just

visit this particular link and here try to create an account so what you can do you can use your Google account okay you can use your Google account either you can also directly write your email so what I will do I'll use my Google account and just try to quickly log in this particular account so this is my account I'll Sate so if you don't have the account so it may ask some information you can fill up and you can create this particular account all of you so previously I have some project here so that's why it's showing but for you it would be completely empty okay it would be completely empty so don't need to worry

so see guys previous I'm having some of the project here see I already created some of the project so for you it would be completely empty so what you have to do guys so here you can see one option called database okay so click on the database uh section here I can click on the database or I think I can click on the overview yeah overview I think it's fine overview and uh overview it's fine now here I will click and uh here you will see one option called new project okay just click on the new project section so

we'll be creating a first of all new project then we'll be creating the cluster inside that okay so here you have to first of all give the name so let's say this is the US uh Visa okay US Visa project okay we are doing US Visa project so this is my project name fine now what you have to do you have to keep everything as default no need to change anything uh simply click on the next button okay so there there is a next button you have to click here fine now it is asking for the project owner uh give your member access

see if you're working as a team and if you want to give the permission to other members what you can do you can invite them okay with the help of email address so that they can also use your uh cluster let's say they are contributing with your application what they can do they can use this particular database they can push the data they can fix the data they can do everything okay based on your permission you will be giving here but I'm not like giving any kinds of permission I'll be doing individually this particular project okay that's why I'll just simply create this particular

project yeah so project has created now it is telling just create a cluster okay what is cluster cluster is the main thing actually so that will store your data okay so here you have to uh select a server there are different different server like aw server gcp server okay so you have to select a server and there you have to create a cluster so let's create a cluster here so I'll just click on create cluster now it is uh telling just give the name of the cluster so I'll keep the default name cluster zero you can give

any kinds of name it's completely fine and see guys if you're using like mongodb Atlas okay so you will get a free tire access okay we will get a free tire access so there you can store 500 uh I think see this is the free tire one you can store 512 MB okay you can store 5 512 MB more than that actually you can't do it okay more than that you can do it and if you if you have more than that actually uh you need let's say more storage that time what you can do you can use their uh this particular service

called M10 so it will charge you per hour 0.08 and there is another service they're providing for serverless okay so here you will get up to 1 tbyte space okay it will do the auto scaling and all and I think u based on some let's say negotiable uh what you can do you can set the price okay like how you are going to pay okay pay as you go I think you can do it so if you're doing for the company that time you can go with this particular service and if you are doing some project okay individually but you

need more stories that time you can go with this particular service but if you're only exploring if you're creating the PC that that time you can go with this particular mzero service okay this is a completely free service so you don't need to pay anything here got it and if you're using uh free service if I'm using free service so everything you should keep as a default okay no need to change anything so provider I will be selecting AWS region I'll be selecting Mumbai F South one because this is the nearest one so we can easily uh like

interact with our data okay in a very fast manner please take a session on how to get a job in the market in the data science and analytics see to get a job first of all try to learn the skill set because skill set is required if you have the proper skill set definitely you can get a job okay and try to prepare your resume in a good way try to mention some good projects there then keep applying for the company okay but many people what they do they don't learn the skill set properly they don't have the good project they keep on applying they

don't get a call okay that is the main thing even I will also give you the suggestion how to create a proper resume how to mention the project in the resume okay how to discuss the project in the interview so it would be very much clear at time okay in this particular uh YouTube channel only I'll be bringing some of the YouTube live okay on top of this particular concept right yeah so these are the setup uh I'll be keeping as default now what I will do I'll just create this particular uh create this particular cluster okay

so there is a button called create deployment so let's click here now it is asking um to solve one puzzle actually bus so I have to select the bus here now let's do the next now motorcycle now see my class St is getting created so it will take some time let's wait once this cluster is ready then we can uh do the configuration okay so now it is asking uh connect to the cluster username and password so here I'll set the username see make sure you are setting this particular username and password and you are remembering this because you need it later on whenever

you want to make the connection with your cluster so I'll keep this particular username as YB and the same name I'll use for the password okay so don't share this credential with anyone else guys otherwise they can also use your cluster okay I'm showing because I will delete after this particular session on that's why I'm showing here now everything I will keep like that yeah so antiy antiy now database user I have created now choose a connection method so it is telling add a connection IP address so let's go to the security

so here I will add one IP address I will add one public IP address so from all the operating system I can access this particular uh cluster so here I will allow access from anywhere I will give this particular permission and I'll will confirm it yeah thank you thank you um D sh yeah so now you can see uh my IP address is set now what I will do um I'll go back and I'll just do the connection method right now yeah so now I'm not going to save it so now it is telling how you are going to connect this particular um let's say uh cluster so here you can use different different service you can also use

compass compass is a service uh from the mongodb mongodb has one software called Compass okay with the help of compass also you can connect otherwise you can also do the Shale connection mongodb for vs code connection you can do Atlas SQL connection you can also do okay but I'm not going to do these are the connection because simply I'm going to use the python code here so with the help of python I will do the connection so there I'll select this particular driver option okay now here driver wise I'll be

selecting the python because I'm going to use Python programming language if you're using any other programming language you can select any other programming now version wise you have to select this particular 3.6 or later don't use any other version because I saw other version will give some of the issue so that's how you will be using this particular 3.

6 or later version because this is the stable version driver so far now uh once you select it you cannot connect user security permission you canot connect here set up your security permission in the first step security permission is already set up I think H just a check just a minute guys go back okay now I think it is done because you have to create the user first of all then you will be able to do it now I'll choose a connection method driver 3.6 all later now see this

is going to be my connection string okay so this is going to be my connection string so now what I have to do uh it's showing the password yeah I'll copy this particular connection string and I will go back I'll go back to my code okay so here I'll be creating one folder first of all let me create a folder here so I'll name this particular folder as notebook see guys everything I'm doing in front of you only everything I'm doing live so sometimes actually uh there would be some issue because because uh I know that live coding is like uh very much difficult and it's fun

also like if you're getting any bug we are resolving in front of you only so that you can also get to know okay how to resolve any of issue so live coding is always fun guys okay yeah it's not a recording session so I'm not cutting anything in front of you I'm doing each and everything okay yeah so in the notebook folder I'm going to create a jupyter notebook here so let's create a jupyter notebook so I'm going to name this particular notebook as mongodb okay mongodb demoore demo dot IP ynb ynb okay

this is The Notebook I'm going to create because as I told you first of all we'll be doing the notebook experiment and if this particular notebook experiment is working fine then we'll be doing the modular coding so first of all let's see how we can connect with our mongodb Atlas okay and how we can push the data there so here is my notebook so first of all I'll select the cardal here so let's like the cardal so I created one python environment I think you remember called uh Visa okay US Visa I think yeah so this is going to be my environment I'll

select this particular environment and here I'm going to run one print statement if this particular print statement is working fine that means my kernel is ready now you can see it is connecting with my kernel you can also use your separate jupyter notebook it's completely fine but I'm using my jupyter notebook inside my visual code Studio okay because visual code Studio provide this particular functionality to you now see guys it is printing okay that means everything is fine okay everything is fine there is no issue okay with this particular C now here I'm going to paste

my connection string so this is going to be my so here I can paste my connection string and uh don't share this connection string with anyone guys I'll remove after this particular session that's why I'm showing in front of you okay so this is going to be my connection string so if you share this connection string so people can access your cluster that means database cluster okay now here this is your username this is your password the username and password what we have set right so with the help of this particular connection string we'll be able to make the connection with our database okay so now

what I will do I'll keep this particular connection string just for my reference so that later on I can use it yeah yes you can uh mention this particular project in your resume very good okay because this is a real time project you are doing it's not a dummy project okay and it's completely end to endend we are doing just a minute fine now what I have to do uh first of all I have to show you the data set okay that first of all I have to get the data

set because I'll be loading this particular data set and I will push this particular data set to my mongodb that means here is the architecture I have drawn already I think you remember first of all I have to get the data set and I have to push this particular data set to the mongodb okay mongodb Atlas then I get this particular data okay so to get the data what you can do you can simply search U um easy Visa easy easy Visa data okay kagle so this particular data set is available here easy Visa data set so this is the data set URL let me

show you yeah so this is the data set guys you can see Cas study uh sorry case ID continent education U uh of employee has job experience or not required job training or not the same features I showed you right in that particular PP so everything is available here so this is going to be my data set now you can download this particular data set okay simply you can download this particular data set so let me give you this particular Link in this particular readme file so here is my readme file or U yeah so here I can just mention data link okay data link so from

here you can download this particular data set okay either what I can do I can uh simply give you my data okay so I already downloaded this particular data so what I will do I'll keep this particular data inside notebook folder so just let me move this particular data just a minute Visa data set okay so this is going to be my data guys okay the CSV file I have already downloaded here yeah so visar data set you can see the CSV file I have already downloaded okay and I kept it inside my notebook folder now this particular data set is available here okay now I can open this particular data set you can see this is

the comma separated value okay yeah you got the connection error I'll tell you okay how to resolve the connection error first of all let me complete that particular quote Yeah so now what I have to do guys first of all I have to load this particular data okay so how to load this particular data so for this we can use pandas Library I think you remember yesterday we set up our requirements and there we installed the pandas Library okay pandas Library we installed now with the help of pandas we'll be loading this particular data now first of all let's import this particular pandas so

import pandas okay as PD all right then I will load this particular data so here I'll just write DF is equal to PD do pd. read CSV okay pd. read uncore CSV okay so inside that I have to mention the location of my data so this is my location and the name of the data is visad dat.

CSV so here what I can do I can copy the name I copy the name and here I'll will just try to mention it here okay visad dat. CSV now if I just do simply DF doad it will load my data see guys it has load my data so here you can see K ID continent education of employee has job experience has required training number of employee here of estabish reasion of uh employment privilege wage unit of uh wage full-time position and case status so this is going to be my target value so by that this particular Visa is uh certified or denied that means whether this particular Visa is accepted or not accepted okay this is going to be my target variable and everything is my X

variable that means my independent variable okay I hope this part is clear all of you all right so now next thing what I have to do I have loaded my data now the next thing um first of all let me um see the shape of the data so to check the shape you can write DF do shape so here you will see the number of records and number of column so here we are having 2,480 rows in this particular data that means this many of application we are having and we are having 12 column here okay 12 column and what are the 12 column we are having you can also printor the columns so here you can write DF do columns okay DF do columns

so it will list down all of the columns actually you are having in this particular data set right yeah now one thing guys if you want to okay if you want to um store something to the mongodb okay if you want to store something to the mongodb so first of all you have to convert this particular data to a format called dictionary okay dictionary so you have to convert to a dictionary So currently my data it is present in a CSV file okay CSV file CSV file is a comma separated value because there we are having table okay we having table and inside that we are having our data now there is a

converter I think you know inside pandas called uh called two dict okay two dict with the help of two dict what I can do I can convert any against of CSV file to dictionary format okay so what is the dictionary format so dictionary format is having first of all it is having the key okay is having the key then it is having the value okay it is having the value so with the help of key and value pair we have to store our data the mongodb okay so this is the functional functional actually technique of the mongodb this is the function of mongod

DV okay that's how mongod store that data but if I'm talking about myql or any other data so they stored the data in a schema okay schema format they stored the data so that's why myql data is a uh it's a relational database okay that means it's a uh it's actually structured database but whenever I'm talking about mongodb it's unstructured database because it can also store unstructured data okay unstructed data means let's say there is no fixed columns and row okay so this particular data also it can store okay that's why this particular dictionary format is

required okay we'll be storing our data in a key value format so first of all let's convert our csb to dictionary so for this here I'm going to write DF dot uh two dict okay DF do2 dict so this is the converter and here you have to mention one parameter called oriented okay oriented is equal to record okay records so now here I'll be storing everything in my data variable now if I show you my data so you'll see the complete data right now see all of the data has been converted to the dictionary now you can see this is my my column has become my key and uh you can see inside whatever

I'm having that means case ID this is going to be my key and and these are my value this is going to be my key and these are my value okay you can see one by one all the record okay one by one all the record actually I'm having now if I show you this particular length of this particular list so you'll see that I'm having 2,480 the same number I'm getting uh whenever I was calculating the shape of the particular data right that means everything is working fine only what I have done I have converted my csb2 I have converted my data csb2

dictionary format so that I can store to the mongod okay and that is what mongodb suggest this is not my actually technique so this is what actually mongod suggest usually right yeah so you can let me know guys if it is working for you you can um code with me along with me yeah so now what I have to do I have to uh write the main logic so now I will store my data uh I store my data where I will store my data to my mongodb okay so for this I need uh some of the things actually first of all I need my DB name first of all I'll decide my DB name so dvore name that means the

database name okay data first of all I have to create a database inside that particular class chart and inside database I have to create a collection okay so how your mongodb works so let's say first of all I have created a class CH here I think you remember we created a class chart so let's say this is my entire class chart so inside class CH first of all I have to create a database let me take a new color so let's say this is my database so cluster name is I think you remember clust name is it was cluster zero okay cluster zero now we have

created a database okay so let me Define that database here so database name is here I can give the database name called US Visa okay usor Visa let's say this is my database name so it will become like that US Visa us \_ Visa so this is my database name now inside that I'm going to create a collection okay I'm going to create a collection so inside collection all of the record will store okay all of the record will store that means all of the data actually you have converted to the dictionary all of the data will stored inside the collection okay so let me Define one variable here so I'll open it

up and here I can write collection name okay so collection collection underscore name is equal to so here I can give the collection name as uh Visa data okay Visa \_ data so that means inside this particular collection inside this particular database it will create one uh collection called Visa uncore data okay so that is how actually your mongod works first of all you have to create a collection cluster inside cluster you have to create a database inside database you have to create a collection and inside collection you have to store your data okay that is what you have to

do now how to store the data for this I need a connection string I think remember we collected one connection string as well so let me just Define so connection okay connection underscore URL okay connection URL is equal to so here I'll copy this particular URL I copyed I think you remember this particular URL I'll copy and here I'm going to mention it here I'm going to mention it here okay so these are the variable are required if you want to do the connection with your mongodb Atlas all right now I think yesterday remember we installed one package called P I told you with the help of P

will be making the connection with our mongodb with our database so now let's import the P here so here I will just write input P Pi now simply I'll just create one uh client here so I'll just write client is equal to P dot uh client so there is a method called client inside that you have to pass the connection string okay connection URL connection URL okay so with the help of this particular connection URL it will create the client now once we created the client I have to create a database here okay inside client I have to create a database so here first of

all I'll be create a DB name that means database and this is going to give me one object called database so data base okay database object I get here now once I created the database inside database I have to create a collection I think remember first of all I created the cluster inside that I have to create a database now inside that I have to create a connection okay collection so let's create the collection so here I'll just write dataor base and inside that I'm going to create a collection name okay so collection name

and this is going to give me an object called collection so collect collection is equal to my database now simply what I will do I'll just store my data okay I have to store my data uh just a minute [Music] guys e [Music] sorry I think my voice is not Audible

am I audible guys can you hear me okay fine so now I'll store my data here so how to store the data so for this simply you just need to write collection okay collection do insert menu okay insert insert menu so I have to insert my data the data actually I have uh converted to the dictionary so this particular data I have to store here so here I'll give the data data and this is going to give me my record object okay so re is equal to my data now if I execute this particular line so it will restore my data to my mongod DP let's

see see if you go to the mongod right now so you'll see that it has created the database as well as the connection and it is storing the data one by one let me show you so I'll just click on done now if I go to the database and if I click on the browse collection yeah you can see that US Visa database is created and inside that my Visa data has created that means my collection is created now inside that you can see all of the data is getting stored okay now I think execution is completed let me check yeah so guys you can see execution is completed that means it has successfully stored my data

okay it has successfully stored my data why it has successfully stored my data inside my mongodb Atlas okay inside my mongodb cluster okay that means this particular step I have done successfully that means there is no error now what I can do I can use this particular code snipp it I can use this particular technique to write my modular coding whenever I'll create the pipeline I'll use the same concept okay to um I mean sorry data I have already stored but whenever I'll face the data so I'll use this particular code init only like how to make the connection with my client right how to get the record we'll

be using this particular concept only there that is the idea now see we have successfully stored the data now if I want to show the data that means if I want to F the data so how to do it for this you can write one line of code so here um you can write one line of code called uh Records is equal to collection do find okay so it will find all of the record inside the collection whatever you are having and it will give you one object called record object okay so this is going to be an object so let me show you so this is going to be a Pongo

object now if you want to show the data what you have to do you have to write a fold Loop so here let me write a fold Loop so I'll just write for for i j in I'll do the enumerate operation so that I can see the number and inside that I'll pass my records okay now here I'm going to print my uh number of uh item as well as my data so here I can take a string now first of all I'm going to print my I value that means the number okay 0 1 2 3 for this particular number and side by side I also want to print my data okay that means my record now if I

execute see it will F all of the data one by one see so this is my first record this is my second record third record so on okay you can see and one unique object IDE would be created by the mongod itself this particular uh column we we don't have in our data if you check this particular column if you check this particular CSV file we don't have this particular feature so this particular feature would be automatically created by the mongodb okay by the mongodb it would be automatically created so this is the unique ID so with the help of this

particular unique ID it just differentiate the each and every records okay in the mongod okay that is what actually happens because let's say you are having same uh same record that time what will happen it will replace right but if you are creating this particular unique ID that means you are separating that particular record okay that is why this particular column would be created fine now see one by one it is fetching all of the data so let's wait uh it will F all of the data then what I can do I can convert this particular record to uh my

data frame again okay so it is fetching uh almost more than 2ou uh 25,000 records so it's taking some time okay done you can see execution is done now see this is the like dictionary format right now so now I'll just try to convert this particular dictionary to CS uh uh data frame format again so how to do it so here I'll just write uh uh this particular I'll import this particular um pandas so here I'll just call this particular method called data frame and inside that I'm going to pass this particular list as my collection.

find collection. fine okay and this is going to be give my data frame now if I show this particular DF dohe you'll see my data has uh my data will become data frame e so that's right it is loading after that I will see that it will convert everything as a data frame

so in between if you have any query guys you can ask me I'll try to reply all of your answer let me know any query or any issue you are having you can let me know and I'll push this particular code okay no need to worry I'll push this particular code if you have any question guys you can ask me and please guys uh give me a like okay give me a like in this particular live okay give me a like in this particular live and if you haven't subscribed to my channel just try to subscribe to my channel guys because I

saw like many people are watching my content but they are not doing the subscription okay so please uh give me some motivation so that I can bring different different content okay in front of you so that you can learn something at the end from my side okay so please do the Subscribe guys and please like like the I mean uh hit the like button okay all of you so please uh hit the like button and do the comment here whatever issue you're having you can let me know we want fine tuning playlist from you sir definitely I will do the fine

tuning also first of all let me complete this particular playlist server timeout error okay if you're getting timeout error in that case what you can do you can use this particular code snippit see this particular record has been converted to my data frame again okay now let's say I don't want this particular ID so what I can do I can write another line so whenever uh this is the line guys so from this particular DF I'm just removing this particular ID column you can see drop drop I'm doing now if I

print it again if I print my DF do here again now you'll see that this particular ID is not there okay now it's my completely raw data okay I imported in my mongodb right and if you're getting this particular timeout error guys so you can use this particular Cod Sate let me show you so there is a package called certified you have to use that time so let me just mark down yeah so this is the code snipp you can use okay I have uh given as a comment I have given as a markdown so so later on you can use it fine who is getting this particular error let me

check I think uh Shafi sadik you are getting this particular no not sadik I think v v you can check this particular code okay I think this will resolve your issue please uh complete oops part in more detail for example keybo net is keyb flow monitoring t uh I will try to cover also okay in the future sadik all right so that's how actually guys we can uh push our data to the mongodb atlas uh to the our database and we can f the data I think you already saw that demo okay successfully you are able to do it and it is already available in my mongodb Atlas you can see in Cloud it is available okay in

Cloud it is available all of my records okay now what I have to do in the next session whenever I'll will do the data in part that time I'll tell you how to write the entire pipeline entire Pipeline and how to F the data okay from here as of now we have only done the notebook experiment here fine now let me give you this particular code so but before that first of all let me remove my connection URL I'm not going to share my connection URL with you guys so what I can do I can um copy this particular URL because

tomorrow I need it right tomorrow I need it tomorrow I have to um complete this particular project so what I can do I can save it somewhere I think that would be fine just a wae guys let me save this particular connection somewhere and please try to use your uh connection string okay don't use my one it's better to use your one that should be my suggestion so now I can remove this particular connection stream from here okay so use y URL okay so here I have commented out and here also I'll just try to remove

this particular line this particular line I'll just try to remove all right now let me save and let me push the changes to my GitHub so that all of you you can get my codee so here I'll just simply update mongodb demo and so I'm getting different error operation failure UNS supported op query command inserted client driver may not require an update um maybe you missed out something please try to check from scratch again provin because it is working working for me when it is working for other people also okay so please try to check from your end maybe you can remove that particular cluster again you can delete

that particular cluster again and you can create a new cluster okay that way actually I created all right okay so mongodb is collected from Alo I think yeah great so fine we have successfully done the mongodb connection even I have already shared the code you can check out my GitHub so it is available yeah mongodb demo added now here is The Notebook folder where is The Notebook folder just a minute yeah you can see the notebook folder inside that you will have the data as well as the notebook okay so you can execute from your end fine

yeah so now what I will do uh I'll just try to uh show you the next um topic called this one uh this one is completed uh database setup is completed now let's do the logging module then exception module then utility module okay it's like very simple logging means like uh whatever print statement you are doing in your code uh instead of doing the the print statement in the terminal you can save in a file okay you can save in a log file because what happens whenever you production add this particular code okay in the Cloud Server there you will be getting any kind of terminal so let's say if some error

occurs in your application if some uh if some failure is there what you can do you can download that particular log file and you can immediately check okay what is the issue happen inside your board okay that's why logging is very important so to create the log what I will do there is a logger folder I created I remember inside US Visa so there is a folder called logger okay so there is a Constructor file we created let me open this particular Constructor file okay now see logger is very simple I created a basic logger for you so this is the

custom logger okay this called custom logger so here is the logger guys okay so I'm using this particular logging module only okay I'm using this particular loging module only from the python so this is the pre-built login loging module you don't need to install it okay on top of that I'm creating my loging string and I'm creating a log folder inside that I'm creating a file and you can see I'm creating a file then I'm just logging the information okay that's it now you can ask me sir how to write this particular loging string okay how to write this particular loging

custom loging so for this you can refer this particular website called python just a minute python um python login okay python login so there is the official U documentation of python you can see the login so there they have mentioned different different login you can do it custom login see you can either get a function you can also use uh this one and if you just go below see different different format is there different different format is there okay so I referred this particular documentation I created this particular

loging string okay that means I created this particular custom log for you all right so so this is what actually you can use so the log form it would be like that first of all it will save your um you can see that first of all it will save your date and time okay just to create the log file because inside the this particular folder it will create a time stamp folder okay inside that it will save the log okay it will save the log now let me execute this particular code then I'll tell you okay how it will look like I think that part would be

very much Clear okay instead of explaining here and this is the logging format first of all it will save the aski time that means the current time stem you're executing this particular code then name of the file you are write uh like running this particular code okay and the label name okay label name that means whether it's information label log or or say the bug level log okay it will show you here all right then the message okay message of the log like what particular message actually want to save here so now let me

show you the demo so what I can do here I can create a file or I can use this particular file called demo now let me import this particular loging first of all here so let me import so here I can import like that so from uh US Visa let me select my environment it's Visa okay yeah so from uscore Visa then inside that I'm having one uh folder called logger logger then I'm going to import this particular login so import login okay import login now you can log any kinds of message so let me just log one message here so I'll just write login

do info and here you can give any message let's say Welcome to our custom log okay Welcome to our custom log now let me save and if I execute this particular demo. Pi you'll see this particular log message would be created so first of all let me activate my environment so P activate Visa let me clear now now I'm going to execute this particular demo file okay demo file demo. Pi now see demo.

Pi I have executed now you can see automatically this logs folder has created okay inside logs folder you can see one time stamp file has created so this is the current time stem and current date if you can observe this is the current time stem and current date okay the current time stem and current date I have executed this particular file okay now inside that the log message actually I have given it is saved here Welcome to our custom log and you can see the current time stamp it is running from the root directory and this

is the information label log okay this is called logging that means instead of printing the information what I can do I can log it so that I can download this particular file later on and I can see what happens inside my code okay so whenever we'll do the uh like pipeline implementation that time I'll tell you okay why you have to use this particular loging functionality and all okay as of now we have just created this particular um code okay later on we'll be using fine so this is what actually login that means your custom login you can complete this particular login like that now similar wise will be writing our exception that means if some

exception is occurring inside my code okay so how to check this particular exception so for this let me write one custom exception so here what I will do I have already written this particular exception so this is what actually custom exception okay you can see so here I'm using this particular exception package okay from the python only on top of that I'm doing the in init operation okay on top of that I'm doing the inherit operation okay I think you know inheritance okay inheritance in object ored programming that's why I told you

op concept is required okay because here I'm going to use op concept okay so you can see here I've written two function uh one function called error message detail so this particular error message detail what it will do with the help of C package it will try to figure out in which line in which file it was it is having the error okay automatically it will figure out just a minute guys I got one call e e okay so sorry guys uh I just got a call sorry so now let's continue so here you

can see I have written one function this particular function we try to figure out where uh it is having this particular error and all and this is the uh class I have written okay inside that actually uh I'm just calling this particular function okay I have written now what I have to do guys uh this particular uh exception Let's test okay how it will work so for this what I will do I'll go to the demo file and here let me first of all import this particular exception so here I can just write from US Visa okay US Visa Dot uh

exception okay us visa. exception input US Visa exception okay I think you remember this is the name so this is the class actually USB is exception we have to import this particular uh class here now as of now let me just comment this particular line and here what I can do I can just simply write one code so here I can give a try except block so try so inside that I will do one operation a is equal to let's say 2 divided by 0 so definitely it will throw me an error right definitely it will throw me an error now let's see whether it is giving me the proper error or not so here what I can do I can write this

particular statement in exception block so accept uh exception okay exception as e now inside that I'll raise my okay US Visa exception and inside that I'm going to pass my error message as well as my system package okay s Ys so I have to import this particular system because as I told you with the help of system package it will identify okay where it got the error now let me save and let me show you one thing so let's say if I'm executing this particular logging info again okay what it will do instead of replacing this particular log file it will create a

different timestamp okay let me show you so I'll open up my terminal and again let me execute this particular demo. Pi see guys it has created a different timestamp okay for the login and for the exception you can see this is the detail exception I'm getting so it is telling file file demo. Pi that means inside demo.

Pi line number nine okay line number nine you can see so here line number nine okay I got one error called zero division error okay zero division error I got okay so this is the proper exception message actually I'm getting getting okay so this is the proper exception I'm getting here fine this is the proper exception I'm getting so that's why actually we use this particular custom exception okay instead of using any module and all we just write our custom extion and again how to write the custom exception you can visit this particular website so exception okay extion in

Python so here you will see the official documentation and they have already written how to write the proper exception Z with the help of different different technique okay you can follow okay so again I referred this particular website and I created this particular custom exception for you okay so both I will share with you so whenever you are creating different different project guys you can use the same template okay uh the template I have shared with you no need to change anything only the change you have to do the name the name I have written now USB exception you can change with other name that's what

actually you have to do that's it fine yeah so here guys this is the login exception now let's write the utility module so inside uh US Visa there is another file I created called utils okay now inside that I having main utils okay now what is utils actually I as I already told you the functionality will be using frequently in your code okay the functionality will be using frequently in your your code so this is called utility functionality that means let's say here I having some EML file I think remember I'm having some EML file

okay now let's say if I want to read this particular EML file I have to write a function called read EML read EML definitely I have to write a function right read EML function def read ml function so whenever I need to read the EML I need to call this particular function let's say I want to read I want to read 10 times okay I want to read 10 times this particular EML file so what I have to do I have to again write this particular function again and again 10 time 10 time I have to write okay inside my code but it's not a recommended way okay it's not a recommended way so what

I have to do I have to create a file called main utils okay maincore utils okay do PI so inside that I'm going to create this particular function Main utils and whenever I required it when okay whenever I required it I'll just try to import from this particular file let's say from Main UTS import read EML okay and whenever I read it whenever I need it I'll just try to use this particular function there okay so this is called utility related function that means whatever function you will be using frequently in your code that function instead of writing in your code you can keep inside a file called utils

file okay that's why we'll be using this particular utility model got it now let me show you some utility related function I'll be using in this particular project so I already listed down all the functions so let me just copy paste so main utils so here are some utility related function guys okay you have to Us in this particular project you can see that uh we are having uh first of all we have imported some of the libraries and we are also using exception as well as the login as I told you why this exception as login is required okay that's why we're using see first of all

I'm using rml that means if I want to read any kind of EML file you can see that I'm having some EML file I'll be using this particular function if I want to write any EML file okay that means I want to save any information in EML file so that that's time actually I can use this particular function called WR file and it's a simple logic guys okay I'm just opening this particular file I'm writing it it's just a simple code if you search on Google you will get this particular snippit okay uh so you don't

need to write from scratch you can use this particular same template as as it is now if you want to load any kinds of pckl object any kinds of binary object you can use this particular function if you want to save any data to nump array you can use this particular function okay if you want to load any nump array you can use this particular function if you want to save any object in a pickl format you can use this particular function if you want to drop any column you can use this particular function okay so these are some utility related function I have created in this

particular project okay the function actually I'll be using okay the function I'll be using frequently in our code got it I hope this part is clear guys all of you just give me a confirmation yeah so now let me push the changes so I'll save all the file and I'm going to push the changes so here I can write um utils add

it see the changes so it's done now if I go back to my GitHub now I think it's it's updated already here I think my connection is slow I don't know okay see everything is up to dat now you can get the change guys and you can execute from your uh system and you can let me know okay whether it's running or not why to download the data from your

git uh you can see it is available inside the notebook folder okay so notebook folder it is available the data so from here you can download hello timeout error it should resolve because um I already checked it is working and I'm not sure maybe there was some issue with your uh string the connection string you created maybe you missed out something please try to check again all right so fine guys I think this is all from today so today we have covered so many things yeah so done for today you can see like uh we have covered exception as well as the utility moduel

as well so tomorrow uh what I will do guys tomorrow I will start with the component implementation that means but before starting with the component implementation first of all we'll do the notebook experiment that means all the project we'll be doing inside our jupyter notebook first of all okay that means the Eda feature engineering OKAY model training everything we'll be doing in our jupyter notebook then we will try to convert this project project to our endend pipeline okay so this is what actually we'll be doing and I will schedule the next class guys you can check out in the live section so you'll

see the next schedule like when I will schedule this particular class I think mostly tomorrow tomorrow I think at the same time we'll be Shing the next next class here got it can you tell me I'm facing issue commit the code after updating the GitHub what is the issue we are getting simp just try to clone the repository and try to comit the code okay first of all create the repository yeah abishek we are using mongodb database okay you can check out our session from scratch yeah so let me know guys how was the session how was the session was it helpful was it helpful guys let me know

if everything is fine uh if you're getting so you can give me a quick thumbs up in the chat how was the session guys so should I bring these kinds of project in future for deep learning NLP or any other take let me know and please guys try to like this particular video and try to share this particular video with your friends and family and if you haven't subscribed to my channel so please do the subscription okay yeah so let's uh complete uh 12K guys 12,000 sub subcriber quickly uh let's uh I mean let's let's have a Target guys okay I think uh in some days actually we'll

reach uh 12K subscriber okay yeah fine so from tomorrow I'll just try to simplify the things in a easiest way I'll just write uh uh I'll just give you detail actually notebook like all the comments should be there okay all the explanation should be there so that whenever you are referring this particular notebook in future so it will help you a lot okay to get the understanding and all so I'll keep everything ready for you guys only I just need your uh uh this one uh dedication just put your dedication try

to do with me okay and definitely we'll be able to learn each and everything okay so yes guys this is all from today's session uh I hope you like this particular session so thanks for joining the session guys bye everyone take uh take care good night everyone bye thanks

Show thinking

This session was the second part of the "MLOps Production Grade Machine Learning Project" series. The focus was on setting up the database, ingesting data, and creating essential, reusable code modules for logging, exception handling, and general utilities.

**📝 Agenda for the Session**

The presenter covered four main topics:

* **Database Setup:** Setting up a cloud database using **MongoDB Atlas**.
* **Logging Module:** Creating a custom logger to track events and errors.
* **Exception Module:** Building a custom exception handler for detailed error reporting.
* **Utility Module:** Writing reusable functions to avoid repetitive code.

All initial development was done in a **Jupyter Notebook** to experiment and validate the code before integrating it into the main project's modular structure.

**☁️ Database Setup with MongoDB Atlas**

The core task was to take the project's dataset (a CSV file) and store it in a cloud database.

**Workflow**

1. **Data Source:** The **US Visa dataset** was downloaded from Kaggle and loaded into a pandas DataFrame.
2. **Database Choice:** **MongoDB Atlas**, a cloud-based NoSQL database, was chosen. It stores data in a document format (similar to JSON), which is flexible for various data types.
3. **Setup Steps:**
   * A new project and a **free-tier cluster** were created on the MongoDB Atlas website.
   * Security was configured by creating a **database user** (with a username and password) and allowing **network access from any IP address** (0.0.0.0/0) for development purposes.
   * The unique **connection string** was copied, which is required for the application to connect to the database.
4. **Data Ingestion:**
   * The pandas DataFrame was converted into a list of dictionaries, as MongoDB works with document-style data.
   * Using the **pymongo** library in Python, a connection was established to the MongoDB cluster.
   * A new database (US\_Visa) and a collection (visa\_data) were created programmatically.
   * The entire dataset was inserted into the collection using the collection.insert\_many() command.
5. **Data Retrieval:** The session concluded by demonstrating how to fetch the data back from MongoDB and load it into a pandas DataFrame using collection.find().

**⚙️ Core Project Modules (Logging, Exceptions, & Utils)**

To build a robust, production-ready application, three key utility modules were created.

**1. Custom Logging Module**

A custom logger was implemented to replace standard print() statements.

* **Purpose:** In a production environment, you can't see terminal output. Logs save important information, events, and errors to files for later review and debugging.
* **Functionality:** The logger automatically creates a logs directory and saves log files named with the current **timestamp**. Each log entry includes the time, log level (e.g., INFO, ERROR), and the specific message.

**2. Custom Exception Module**

A custom exception handler was built to provide more informative error messages.

* **Purpose:** When an error occurs, a standard Python traceback can sometimes be vague. A custom exception provides more context.
* **Functionality:** It uses Python's sys module to capture the exact **file name and line number** where the error happened, making it significantly easier and faster to debug the code.

**3. Utility (utils) Module**

This module serves as a toolbox for common, reusable functions.

* **Purpose:** To follow the **DRY (Don't Repeat Yourself)** principle. Functions that are needed in multiple places are written once in the utils file and imported wherever required.
* **Example Functions:** The presenter added functions to:
  + Read and write **YAML files**.
  + Save and load objects using **pickle**.
  + Handle **NumPy arrays**.

**🚀 Next Steps**

The next session will move on to the core machine learning components, starting with notebook experiments for:

* Exploratory Data Analysis (EDA)
* Feature Engineering
* Model Training

e e e uh hello everyone good evening I hope I'm audible to all of you just let me know in the chat

uh hello everyone good evening okay fine I think uh you can hear me you can see me uh hi good evening HP um good evening Michael all right so now I think we can start with the session but I will wait one more minutes uh so that everyone can join yeah good evening kishan good evening everyone guys sir please provide collab link for jna projects for the reference purp please I think it is available in my GitHub let me just show you so if you go to my uh GitHub profile so make sure you do the follow on uh the GitHub so that uh whatever let's say new repository I'll be creating so you'll get a notification

all right so here I think uh there is a repository I created called uh complete generative AI course on YouTube okay so inside that uh all the materials is available so see today I think I uploaded the Falcon video so it is available here okay so inside Falcon folder so this particular uh ipb file is available s okay you are doing amazing job in weekend good effort yeah thank you yeah today link also it is available here inside the Falcon folder so just do check it out or else I can share you this particular repository link uh just a minute

so here is the link so so many content guys you will see in my YouTube channel you can see related jna related uh end to end project okay related deployment related uh mlops so all kinds of tutorial are there but going forward I will try to create more detail tutorial I'll just pick up real time example like I will be doing different different projects with the help of these are the tools okay so that each and everything would be uh clear in your mind fine so this is going to I mean um this is going to be completely unique content

guys so make sure if subscribe to my channel and uh hit the like button and please try to share this particular Channel with your friends and family so can uh so they can also get to know like this kinds of content is going on okay in this particular channel so okay guys so I think we can start with the session but before that let me uh show you where you will get the previous resources like the previous class resources so here here I created one uh playlist in my YouTube channel you can see if you go to the playlist section so here is the playlist called mlops production ready machine learning

project so if I go to this particular playlist so you can see uh today is day three till day two everything is up to date so if you just open this particular video and if you just go to the description section so here you will see this particular GitHub link okay so here actually I'm committing all of the code all of the source code okay throughout the entire session so please try to check and uh try to Al also Fork this particular repository also you can start it so that you can um I mean save this particular repository okay in your

account all right so first of all let me show you the agenda today I'll be covering so let me just connect my writing paper and here I can uh open up my blog board how many days are left to complete the full project so we'll be doing seven days 7even Days uh live actually in seven days actually try to complete everything fine so now here uh today actually what we are going to do so today I think uh so far what we have done guys I think we have completed the introduction uh and the setup part then um yesterday I was doing the

database setup database setup as well as the uh like uh data inje operation I was doing so today I'll just try to do the explor data analysis that means Eda part of this particular project okay Eda sorry so it should be Eda then we'll be doing the feature engineering then we'll be doing the model training OKAY model training everything we be doing on Jupiter notebook okay jupyter notebook everything will be doing because of before starting with our actual implementation so what I feel like if I do everything in the jupyter notebooks it would be easy for me to write the

code later on right and definitely Eda um is kinds of things actually you can do in the python scripting file so definitely for this you you should use this particular jupyter notebook or collab notebook whatever you are using because we have to um we have to analyze our data we have to load the data we have to plot different different graphs and all okay and from that particular graphs we can get the Insight like what I have to do based on that actually we'll be performing the feature engineering technique okay if you know the machine learning life cycle so I think you know Eda feature engineering

model training uh evaluation these are the very important uh step we usually perform right so that's why uh in entn Project also you be following the same uh strategy I hope this part is clear okay so I already prepared guys uh some of the notebook for you and it is completely in detail I have commented each and everything so I'll share this particular notebook with you so that you can use this particular notebook you can use this particular template for the other project as well every I have written as a automated way so there actually I'm not writing any against of

manual code so whatever things I was doing in a automated way only okay I'll tell you how I've written this particular entire code all right so today I'm not going to write the code from scratch because I already prepared the notebook and it took lots of time from my side right to prepare the entire content so definitely I'll explain the things because it's like very easy at the end if you know the machine learning if you have done the ADF feature engineering you already know okay what is edn feature engineering so no need to explain something new here okay only try to show you that strategy you can follow

whenever you are working with an oft data set okay what are the things what are the key component you have to keep it uh in your mind okay that is what actually will'll be teaching in this particular session all right and this is super important session guys if you don't understand this particular session and if you don't do this particular work so definitely you can't create any kinds of good product you can't create any of good model so to create a good model EDF feature engineering is super important concept um I will suggest to everyone right yeah

yeah so if anyone uh um having any doubt you can let me know or everything is fine guys uh you can let me know if everything is fine just please give me a thumbs up in the chat or any kinds of emoji you can send so that I can get a confirmation you are ready and I can start with the uh like discussion okay for today's agenda let me know guys if everything is um fine from your end just uh give me any kinds of Emoji guys in the chat yeah sakit uh we'll try to start the jna as well first of all let's complete this

because I'm individual person so I can handle one Topic at a time right so allow me some time definitely we can do each and everything great so now uh let's uh go to our uh visual code Studio guys so you can open up your visual code Studio guys and uh what I will show you guys um see I have updated two notebook here so one is like Eda US Visa ipynb and another is like feature engineering and model training ipynb so two notebook I have added here so in this particular notebook we are having the entire code you are having the entire analysis and

everything so we'll try to discuss this particular notebook today so if you want to get this particular notebook I can comit these are the changes in my GitHub so let me just comit quickly so here I can write Eda and feature engineering added okay for today so I can commit and I can send the changes so yesterday I completed the mongod setup okay mongodb demo and I think you remember this particular notebook so as of now I'll close this particular notebook it's not required see my code comet is done now if I go

back to my GitHub Hub and here is my repository and Let me refresh this particular page so you can see this particular comet is added and this notebook is also available inside the notebook folder so please try to get this particular notebook and open up uh in your system okay and you can understand along with me so let me share you this particular link one more time so here I have shared this particular Link in the chat guys okay we can open up from here only so now let me open up this particular expor data analy this

notebook first of all Eda okay Eda I think you know guys Eda let me know how many of you are familiar with Eda Concept in machine learning let me know if you are doing machine learning project so Ed should be clear okay Eda should be clear like what are the things we usually perform inside Ed let me know if you are familiar with machine learning I hope this Eda I mean things you have done before exploited analysis without that actually you can't uh I mean do the machine learning things and if you see this particular notebook I have uh like uh documented in a detailed way each and everything I have

written so if you're referring this particular notebook in future also U so easily you can understand okay what you have done and that's how you have to do the documentation that's how we have to do the documentation that's how we have to write the report whenever you are gaining some insight from your data okay that's how industry project happens it's not like that you are only importing the data you are analyzing the data okay you are dropping the column it's not like that you have to Pro you have to document in a proper way so that other

people can understand because it's not like that you are only person you are working in a team there would be thousands of people right so definitely you have to document each and everything if we upload the data on the mongodb through the compass still I need to write the code for the uploading the mongodb no you don't need to do it so if you're uploading from the compass itself it's completely fine but here we are not installing the compass because it's additional tool so everything doing with the help of python only right as a developer we have to use the programming language always yeah so here if you see guys this

particular notebook so here I have mentioned the life cycle of a machine learning project I think this thing you already know like what is this particular life cycle and all right so first of all we uh what we do we just understand the problem statement in any machine learning problem statement we first of all try to understand the problem statement like what we are going to solve right then with respect to that what we have to do we have to collect the dat data that means we have to do the data collection

okay data collection then after collecting the data we have to perform the Eda part that means exploor data analysis which we'll be doing right now right explo data analysis then after doing the exploit data analysis we have to start with the feature engineering technique inside feature engineering you can uh do the data cleaning if uh some issue with your data you can do the data cleaning you can do pre-processing the data okay then you can perform the model training operation once model training is done you can choose the best model out of it okay that's how uh your life

cycle has uh happened okay in the machine learning field okay so this is the common step for all the project all the regression classification classing any kinds of project you will be doing you have to follow the same pipeline okay same life cycle always all right now you can see the about this particular data set about this particular problem statement here I have written I think no need to explain again I have already discussed this particular problem statement now here you can see the problem statement now I think you can see the data collection part data wise I have collected the data from the kaggle itself I already told

you from where you can get this particular data it is available in this particular uh URL that means it is available in this particular Kagel website and the data name is easy Visa data set okay from here you can download the data set and it is having more than two 25,000 U 25,000 rows and 12 columns almost 25480 rows and 12 columns okay I already showed you this particular data right then here what I'm doing guys first of all I'm importing some of the necessary Library so let me execute one by one so here I can select the cardal so this is

my Visa cardal I have selected now let me execute this particular cell so here if you see I'm putting pandas numai M plotly cbon okay plotly so these are the library I'm importing because going forward I'll be doing the export dat analysis I have to visualize my columns and everything right so for this these particular libraries are required all right then what we are doing first of all guys we are first of all loading the data because you can see that this particular data set is available in my notebook folder so I'm loading loading this particular data called Visa data.csv okay with the help of pandas

read CSV I'm loading this particular data so let me execute this particular line you can see my data has loaded and this is the data look like okay you can see this is the data interface so it is having some of the features as I already told you like case ID continent education um of employee has job experience or not okay and so on so now can anybody tell me uh which is the irrelevant column here like which is not useful in our uh analysis can anybody let me know uh just see this particular columns okay and let me know okay but by

seeing that particular column you can easily understand okay this particular column is not required for this particular analysis I can directly drop this particular column let me know guys just see this particular data uh in a clear way and let me know which column is not required I can drop directly so please uh reply quickly in the chat so that I can um have a like inter this session with you yeah case ID right case ID is not um continent I think continent is required but if you see the case ID case ID is not required

because it's a unique idea at the end right it's a unique idea at the end so I can directly drop this particular case ID column got it yeah so we'll be dropping this particular column later on whenever we'll be doing the feature engineering task as of now we are just gting this inside from the data like we just try to understand the data okay what my data is able trying to tell me first of all I have to understand the data based on that I have to decide which which particular feature engineering task I have to apply whether I have to apply power transformation whether I have to apply uh let's say

label encoder where whe whether I have to apply one or encoder okay whether I have to do the scaling or not whether I have to handle the imbalanced data set or not everything I'll be designing after performing the Eda part okay Eda part in this particular data set itself right yeah now if you see the shape of the data as I told you it is having 25,0 480 rows and 12 columns you can see is the count and you can see the summary of the data as well so you can use this particular method called describe so it

will describe the entire data you can see the count of each uh numerical column here and this the mean I think you know the mean right in statistics it is available mean standard deviation then 25% uh percentile uh quantile 50% quantile 75% quantile right max value and mean value I hope this particular uh table is clear all of you this particular uh summary is clear because this is the statistical summary okay statistic maybe you know already right but is mean median mode okay 25% percentile 50% percentile 70% percentile okay max value

you already know plotly means like it's a library okay it's a library with the help of plotly we can visualize our data okay it's like a m plotly when cbon you can either use plotly M plotly cbon it's up to you but we are using all of them so that I can show you like some of the amazing graph okay that's all let me know this particular graph is sorry this particular table is clear or not yeah I think this part is uh this part is clear because these are the very common step we usually perform in any project right then you can also do the info info check that means information

check of your data so here you can see like uh this is the entire information of my data so here I'm having 12 columns okay different different 12 columns I'm having and whether it is having null value or not so there is no missing value you can see so all of the data is having the same count and you can see the data type actually most of the data type is a object type okay you can see most of the data type is a object type so some of the column is having floting data some of the column is having integer type data okay so that's how I can get a quick uh actually idea of my data set like whether my data set is

like numerical data whether it's a categorical data whether I'm having some missing value or not okay so if you want to get a quick kinds of information you can get with the help of this particular info okay info command and you can see the data size also around 2.3 megabyte more than 2.

3 megabyte okay it is using in my memory space all right and this is the inside count floating column you are having one integer column you are having two object type column you are having nine yeah isal do sum you can also do but this will give you the quick summary okay quick summary of your data isal will be also doing in later on right later on will be doing in now exploring the data okay first of all I have to explore the data for this uh what I have done guys you can see here first of all I have separate out of my numerical features and separated out my categorical features now let me know guys what are the columns are numerical

what are the columns are categorical I think you know the difference between numerical and categorical right I think you know the uh categorical and numerical column so whatever columns is having object type okay whatever the columns is having object type this is a categorical features and what are the columns is not having object type that means it's not object type it's a floating or integer type those call it as a numerical features so see I have written a uh list comprehension here inside that I'm just writing one condition okay first of all

I'm going through all the features then I'm checking the feature data type if this data type is not equal to object that means it's a numerical feature otherwise if it is a object that means it's a categorical features that's how instead of manually picking this column what I am doing I'm writing one automatic script okay automatic script with the help of automatic script I'm easily able to um like differentiate my numerical as well as the categorical features okay so that's how you have to write a smart code always guys okay no need to do like so many effort in your

code no need to write so many code so many lines of code just think about how you can automate the process most of the time because at the end automate with automation will help you a lot okay in future it's just a one time effort you are doing now you can use the same template for other project implementation as well only you have to change up your data that's it okay so that time you don't need to look for what are the columns or numerical what are the columns for categorical you can execute this particular script it will

automatically differentiate your numerical and categorical features okay this is called Automation and this is what actually smart coding okay less effort I am doing very smart work okay now you can see I'm printing my numerical features I'm printing my categorical features okay now if I execute this particular line so here you can see we have three numerical features the numerical feature is number of employee year of establish and privileged boy okay and here you can see we are having nine categorical features Cas ID continent uh education of

employee has job experience not and so on so these are my categorical features all right yeah now uh what we'll be doing next next we are checking the proportion of each each actually categorical features okay the proportion okay the proportion means the percentage okay will be seeing so if I execute this particular code you can see that um K ID K ID every actually um you can see category is having uh 0.

okay that means almost same right almost same white is almost same can anybody tell me why it is almost same because it's a unique number right it's a unique ID it's it doesn't have any reputation okay that's why this particular uh percentage is like very less okay this particular percentage is very less that's why this particular case ID I can remove because there is no meaning in uh this particular case ID in my analysis okay I can directly remove this particular case ID now you can see the continent proportion so Asia is having more actually present in the data you can see 60 6% okay 66% around Asia is appearing

in the application okay then after Asia Europe is there 40% then North America okay then South Africa Africa Oceania okay that's how you can get some quick understanding of your category like what are the categories appearing more inside your data okay that's how you can also get to know the outliers you can see o ocean ocean actually occuring very less time in your data that means this particular uh category is outliner in my data okay yeah that's how you can understand so similar wise for the educ education of

the employee Bachelor is having like more application than uh if I open this particular um thing in a text editor I think this would be clear now you can see Bachelor is having more then Masters then high school then doctorate okay then has job experience or not so yes uh that means whoever having job experience so their their application is more you can see uh 58% then no then required job training uh 88% of no and only 11% actually you can see for the uh yes okay so that's how you can see different different percentage different different

proportion of the each category I hope this part is clear guys how to check the proportion so this is the formula you can apply this is the formula you can apply you can do the value count normalize is equal to true then multiply with 100 it will give you the proportion now after analyzing till here what is the inside I get okay what is the inside I get you can see the inside I have written so case ID have a unque values for each columns which can be dropped as it is not uh it is there is no use of this particular uh this particular case ID because there is no

importance okay in this particular analysis then continence column is highly B uh biased towards the Asia because Asia application was more we can see Asia application was more okay then hence we can combine the other categories from a single category then unit of wage you can see seems to be an important column as most of them are yearly contract okay you can see this particular column it is mostly year contract so that's how actually you can whatever inside you are getting for you from your analysis you can write down here okay you can write down here and if you're sharing this particular notebook

with any other developer they can get to know okay so that's how you have analyzed the data and this is what actually data is trying to tell you okay that's how you can do the feature engineering task you can do the model training task okay that's why it is super important to the documentation always okay whenever you are performing Eda task yeah then uh we'll be applying first of all one amazing anal analysis called univariate analysis are you familiar with univariate analysis guys let me know univariate analysis are you familiar with univariate

analysis what is univariate analysis it mean univarate analysis referred to an analysis of one variable okay that means uni means one the purpose of univ analyis to understand the distribution of the values for a single variable that means we'll pick one single variable and we'll plot the particular variable okay and we'll try to understand the distribution of that particular variable okay this is called univariate analysis only one column analysis okay there are other type of analysis are available like B variate analysis is there by variate means we are analyzing of the

two variable okay there is another analysis available called multivariate analysis that means we are U taking more than two variable okay this is called multivariate analysis okay so that's how I think you have learned this thing univariate multivariate bivariate right whenever you used to learn the basics of machine learning maybe you have covered this particular part okay we are using the same thing here okay maybe you can recall that particular concept now let's do the uniate analysis on top of the numerical feature right now so here I have written one M plot Li code so I'm

just um just taking all of the numerical features okay one by one you can see in the fold loop I think I'm taking one by one and I'm just plotting okay I'm just plotting the uh actually you can see that I'm plotting the distribution okay I'm plotting the distribution you can see now by seeing the distribution guys what you can observe here one very major problem I got in this particular distribution can you tell me what is the problem here what is the problem guys if you see this particular distribution number of employee are of establish and this particular privilege

W priv privilege W is like little bit good but this particular two columns are having this particular problem I think you know the skewness right skewness problem left is skewness right is skewness okay this happens because of the outlayer yeah skewed yeah it's a skewed actually okay yeah so we have to also handle this particular situation so can anybody tell me any solution for this particular problem what is the let's say technique I can apply to get a normal distribution can some normal distribution

uh there are some technique I think you know we can apply uh I think power transformer right uh sorry power power transformation power transform transformation we can apply to get a normal distribution kinds of output right apart from that there are so many technique you can apply but I will be using power transformation technique here because SP power transformation is like uh one of the amazing technique I can use proportion means the percentage okay percentage like log transformation is there Z score is

there Z statistics is there right there are different different techniques are there but I'm going to use the power transformation technique now you can see the inside number of employee columns looks more SK skewed but it is may be because of the outlets okay then you can see like different different inside I got from this particular analysis that's how you have to note it down each and everything after doing each and every Ed you have to note it down that particular insights okay yeah so how is this particular notebook guys um will it help you in

future whenever you are doing any kinds of project that's how I prepared this particular notebook guys for all of you and it took lots of time from my side right to prepare this particular entire notebook I could have shown the uh like direct part like I could have loaded the data and I can write each and every line in front of you but it won't be helping you later on okay that's why I have documented each and everything so that if you're referring this particular notebook after one year also it will help you a lot to understand like what

is happening inside this particular code okay I hope this particular notebook will help you and that's how you have to do for other other project as well whatever project you are doing there is a good practice actually so it will give you the entire story of the data at the end if you share this particular notebook with non- coder as well th those who don't know how to code they can also uh like like slightly understand what is happening here okay so fine so now um You can execute this particular sale it will work I already executed so I'm not going to execute

again now let's come to the categorical features so I have plotted all the categorical features okay in a different different actually subplot you can see so here is my categorical features and you can see the um actually distribution of the categorical features here here so as I told you Asia is having more count than Africa North America okay then for the education of the employee so you can see bachelor's degree are most then Master's then high school then doctorate then job experience yes is more appearing in that particular data than no that's how you

can see different different column distribution okay it's a histogram actually I'm plotting at the end right sorry it's a bar bar plot I think I'm plotting at the end okay so that's how you can do the univariate analysis okay for the categorical features and you can see the inside so request job training uh unit of w then full-time position then continent columns have um actually single category dominating okay so you can see the single category is dominating everywhere okay then you can see the in the rest of the columns are balanced you can see some of the columns are still balanced

okay we don't need to part from this particular um actually technique okay in this particular column now once univarate analysis is done guys I will start with the multivariate analysis you can see so again I'm just uh separating out my discrete features and uh continuous features I think you know the difference between discrete feature and continuous feature guys anyone having any idea see if you don't know let me give you uh in a actually easiest definition so see if you just calculate the length of each

and every unique features okay that means each and uh every unique category in that particular features and if it is less than 25 if it is less than and equal 25 it's a discrete features okay but if it is more than if it is more than 25 that time I will consider it's a continuous features okay this is the simple logic you have to apply then you can differentiate your discrete features as well as the continuous features with the data changing over the time how do we keep up all the rate retraining and continuous EDF on the data see Eda you can't add in the pipeline Eda is what actually you do

individually it's out of our pipeline okay you will start from the feature engineering okay now see after executing this particular cell you can see we are having zero discrete features and we are having three continuous features in this particular data you can see number of employees year of stabes and privilege voice okay that's how you can easily get the discrete variable as well as the continuous variable means some Comm means many no that means the count count of each category if it is less than 25 it's a discrete otherwise it's a

continuous see um that's how you can understand let's say there is a number let's say we are having um how I can uh let's say there is a number let's say one 2 one uh let's say again two two now let's say um you can also call it as 1 3 3 okay three now if you count okay if you count this particular number let's say if it is uh three is appearing let's say Okay three is appearing let's say more than uh three times more than three time it is appearing more than three times then two is appearing more than three times okay that's so one is appearing more than three times that we can consider

sorry less than less than three times that that that time you can consider actually uh it's a discrete feature actually it's a discrete features and let's see if there is a number like 1 2 3 4 1 5 6 7 8 okay that means unique value that time you can consider it's a continuous actually variable okay yeah continuous also decimal point can be also considered it's 1.

2 1.3 okay and so on okay decimal point can be also considered so you can simply search on Google you will get the entire definition okay what is the discret features what is the continuous features so now we'll do the multicolinearity checking actually in the numerical features what is multicolinearity I think you know if this is a COR it is having correlated with the target video or not okay so for this actually what I can do I can apply one um technique called uh chai square test chai square test have you heard of

this particular test guys chai square test in statistics maybe you have learned this particular technique called CH square square test let me know have you heard of this particular name chai square test hypothesis testing right hypothesis testing Nala hypothesis right yeah so with the help of CH Square test we can see the multicolinearity of our data okay so see this is the library we can use called Spacey so in side space we are having statistical module and I can imploy this particular chai square test here now this is the simple

code I'm just loading all the categorical features okay and what I'm doing I'm just checking okay whether it is rejecting the N hypothesis or whether it's accepting or fail to reject the N hypothesis okay now see simply you can understand like that you can see after applying this particular logic all the columns is getting reject n hypothesis reject null hypothesis reject n hypothesis only required job training is telling fail to reject the N hypothesis okay now here I have written the report you can see here required job training fails to reject null hypothesis which

means it doesn't correlate with the target column that means what are the columns actually is not getting correlated with the target column this particular column will give you failed to reject the N hypothesis otherwise rest of the column will be reject to n hypothesis okay that's how you have to understand whether this particular column is getting correlated or not with my target column this is only for the categorical column okay this is only for categorical column guys you can see here I'm doing for the categorical features not the numerical features for numerical features you can

use correlation okay correlation will help you to do it okay I think you know the cor relation is one function we are having inside pandas so c o r r okay correlation so it is available inside pandas I think you know that with the help of correlation um method you can check the correlation of the data that means the features okay you can see that uh this the correlation that's how you can calculate the correlation you can see one means it's a 100% correlated and less than one means it's a less than 100% okay that's how you can understand otherwise you can also plot the graph

and all it is also possible here now let's see the null value whether we are having a null value or not you can see all the column is having um it's not having null value okay it's completely fine now let's load the continuous feature again so you can see this is my continuous features okay now here we have applied one distribution called uh you can see this is the distribution we call it as I think box sport okay we have applied box port on top of it you can see the box port and when we apply the box port guys I think you know that box port we apply if you want to check

the outliers okay you can see after applying box port you can see the outliers okay you can see the outliers so this is the number of employment sorry case status that means this is my target variable and you can see the outline okay then number of employees you can see this particular this is the numerical that's why it's showing the distribution like that so you can see the denied certificate is less less than certified certificate sorry application then you can see that particular case status sorry year of stabiliz and case status okay so that's how you can see

this particular box port different different box port okay now you can see the distribution of the previl which like denied application is much less than your certified application okay so these are some quick actually uh I mean analysis you can perform on top of the data okay whatever things you have learned in your machine learning statistics okay everything you can use here now you can see the initial analysis report so you can see number of employees has many outlines which can be handled in the feature engineering now you can see of stabil is left squid and some of the outl below lower bound of

the Box Bo that's how we have uh reported all the initial analysis okay initial analysis whatever initial initial analysis we have done we have reported each and everything here okay now we can visualize our Target features okay you can see this my target features and we can visualize visualize this okay so this is like a p graph we are implementing so you can see with uh this particular uh M plot leave we are plotting this particular pip plot okay you can see the PIP plot and you can see certified application is much than your denied certified that means most of the application got certified 66%

application got certified only 33% application was denied okay yeah that means again it's an imbalanced problem you can see certified people are high then your denied denied people so again it's an issue okay it's an issue we can give this particular data for the model training so what I have to do I have to handle for the imbalance data okay so to handle the imbalance data guys which technique I can use can youbody tell me in machine learning to handle imbalance data what are the technique we can use there are some technique I think you know

smart yes vit you are right we can use Smart okay yeah we'll be also using smart smart to handle this particular imbalance problem right yeah resampling you can also do but smart would be better we can use Smart okay yeah so there are like thousands of technique you can apply but again it's up to you which one you'll be using as a data scientist you have to decide so here smart was working fine that's why I was using smart here so again you can see the target variable imbalanced why what is imbalanced data so imbalanced data means uh there are types of the

data where the target classes has no un um uneven distribution of the observation denied values has more count than the certified values in the data set okay yeah now we'll be doing some grouping okay we'll be doing some grouping we'll try to see that this particular continent okay continent has any impact on the Visa status of not that means we are taking individual column okay see what we are doing let's say this is my target features okay so here are we are having lots of features okay and at the

end we are having Target featur which is nothing but I think status right it's a case status yeah it's a case status that means whether this particular application is got certified or denied so case uncore startus okay so what we are doing we're picking up one column and we are just trying to see okay we're just trying to see the case status whether it is having some impact or not okay on the target variable so first of all we'll be picking this column then this column then this column okay this column That's How we'll be picking

different different column see here we're doing first of all we are picking up the continent okay continent and we're checking the uh checking the impact on the case status now you can see we are doing the grouping operation in pandas so I think you know the group by okay what group by does usually now after applying this particular group by operation you can see this is the continent this is my Cas status okay now Africa okay from Africa you can see certified people proportion is 72% and denied people proportion is 27% then

Asia certified people you can see 65% then denied 34% from Europe certified people are uh you can see uh 79% and denied people are 20% okay that's why you can see the all the proportion okay all the proportion from different different continent okay from all the proportion from different different continent okay and by seeing that particular continent okay by seeing this particular actually table you can observe okay so this is my continent and from this particular particular regon people are getting more certified and people are getting denied okay now let's say you are from okay you

are from Africa okay you are from Africa and you are applying for the uh US Visa okay you are applying for the US Visa now there definitely there is a high highp probability your application will be certified because you can see based on my data okay based on my data the data we have collected here it is telling from Africa people those uh actually who applied for the Visa they got more certified actually around 72% % people okay I hope this part is clear all of you and see everything is automated way I'm writing guys okay there is no manual things I'm doing so that's how you can

also follow the same template now you can see the percent you can see the count you can see the Asia from U this is the denied this is the certified you can see certified people are then your denied people okay so you can see the different different count operation even if you can plot the distribution you can also do it so this is the code for the distribution so here I'm using the M plot Li and this is the percentage you can see this is the percentage you can see okay so similar wise I have written the report you can see as part the chart Asia application applied more than the other countries

okay 43% of the certified application are from the Asia so you can see the report okay the entire report now after that we are picking up another column called education of employee and we again see the impact on the cas status now you can see those who are Bachelors they got more certified okay than the doctorate and high school and Masters okay that means Bachelor people okay th those who have completed their bachelor's degree if they're applying for the US Visa they're getting more certified okay you can see here then actually you can see doctorate then High

School okay then Masters again you can see the count and again you can see the percentage okay the same thing the same thing we did for the first the same thing we are doing for the other as well now we are see uh checking with respect to the job experience you can see those who are um like have doesn't have no job experience they're getting certified around 56% those who are getting U job experience they're getting 74% okay you can see that again the same graph like the count and the percentage okay and this is the report

the final report actually I got now I'm picking up another column called job require job training on not again on the case status you can see yes certified 66 uh sorry no certified 66 and yes certified uh 70 uh 67% and again this is the count and this is the percentage okay and this is the report so for uh this particular column as well we are doing the employe me okay so you can see see the uh this is what this is the numerical column we are taking this particular number of employees on the case status okay you can see this is the gra so denied this

is the percentage of the denied and this is the certified that means almost same fine and this is the inside then we are applying for the wage actually unit of wage on the case status you can see hourly basis those we're applying for the hourly basis they denied got at this perc certified got this person monthly basis weekly basis and basis okay you can see we can use the reset index after grouping you can do it so many things okay I'm seeing yeah now you can see the percentage as well sorry count and percentage okay and this is the final report so for the uh Reon of employment

as well okay we are doing the same thing okay we're doing the same thing okay and this is my final report and this is the distribution of my privilege wage okay that means the salary you can see this is the salary that means uh you can see this is the certified uh this is the salary range and den is the Sal salary range okay so privilege wage based on the uh education based on the privilege wage you can see Masters uh this is the privilege W Bachelor this is the privilege voice and high school this is the privilege voice okay and this is the doctorate privilege voice you can see

this is the distribution of the box and this is are the outliers actually okay so simply you can download this particular notebook and you can execute from your end and you can read this particular report you will able to understand okay what is happening here fine so that's how we have done different different analysis based on the different different column you can see the same thing we are doing okay at the end you can see the air of estabish and this is the distribution and this is my final report I got guys okay you can see the case status column can be dropped as it is an ID required job

training columns can be dropped as as it is doesn't have much impact on the target variable okay number of employe previl Bas columns are have the outlier quent columns is a few unique Valu which is very less count which can be made as the outliers sorry as others Target column case status is imbalanced can be handled before the model building okay so these are the final report I got now with respect to that I have to apply my feature engineering task okay because I got to know okay there is a imbalance problem there is a distribution problem okay there is a column I don't need I

can drop it okay now we'll be applying for the feature engineering okay feature engineering t then we'll be doing the model training all right so this is what actually our exploit data analysis part we have to do okay this is the entire explor analysis part and each and everything I have documented in a such a way so that it will help you a lot okay in future for the understanding purpose now let's open the next notebook guys called feature engineering and model training you can see this is the feature engineering and model training so again I'm not going to execute guys you can

execute from your end because I already executed and everything is working fine for me me again you can select the cardal select the your environment and you can execute this particular notebook it will work fine okay yeah now let's see how we can perform the feature engineering task you can see uh I'm importing some of the library first of all then we'll be doing the data cleaning that means handling missing value dropping duplicates value okay check the data if I understand the data now again what I'm doing so here you can see I'm first of all checking

the missing value whether it is having some missing value or not this is the logic I'm writing you can see just to get the percent I'm writing this particular code now you can see there is no missing mow this particular variable is completely empty just a minute okay now I'm checking the duplicates value okay whether it is having some duplicates value or not you can see there is no duplicates value okay yeah now I can directly drop this particular case ID because as I told you case ID can not be used so I can directly drop

it you can see I'm dropping this particular case ID now feature extraction we'll be doing so this is my data so you can see first of all what what I have to do there is a number of year actually we are having year of estabilish of the company okay now if I want to get the company age so guys what I can do if I want to if I want to get a company age let's say this is my company year of estabish 2017 now if I want to get the company age yeah thanks ransing so let's say if I want to get the let's say age of the

company so what I can do I can get the current current actually year okay today's year current year so from this particular year today's year that means the current year what I will do I'll just try to um like do the subscri substract okay from the previous year now here I will get my current company age okay you can see this is my current company age so here I have done this particular things current company years company age so now you can see this is my company Edge okay company Edge yeah now I can drop this particular year of estabilish column because it's not required because I got my company

Edge okay then type of features you can see so this is my numerical features how to get the numerical features this is the logic I already showed you in the edia notebook as well okay yeah then categorical features these is my categorical features because it is ending with object okay this is my categorical features discret features again again I showed you how to collect this discrete features and as well as the continuous features okay so all the features I collected one by one now I have to do the split okay X and Y split so here what I can do I can first of all

drop this particular y from this particular X and I can uh store this particular y that means this is my independent variable this is my dependent variable this is my target variable okay now here where uh what you can do manual encoding you can apply because you can see the target target features okay it's a categorical it's a categorical but I can directly pass this particular category to my model so what I have to do I have to convert to the numerical representation okay I have to convert to the number so here I can apply one logic so whenever I'm having

the denied it will be assigned as a one and whenever I'm having certified it will assign as a zero that means if your output is one that means this particular application is getting denied and if your output is zero that means this particular application got certified okay that's how we can do the encoding operation okay with the help of np.

wire operation okay now you can see that all the certified and denied got converted to my numerical representation okay this is called encoding technique okay manual encoding technique you can apply now once encoding is done I can apply the feature transformation because as you can see some of the columns okay some of the columns is like squid it is having squid problem so to get the I mean to let's say overcome this particular squid problem what I can do I can apply something called uh you can see the squeez okay squeez count you can see with the help of skewness you can

see the skewness count we can apply something called power transformer okay to check the skewness to like overcome the skewness okay you can see the power transformer so there is a method called EO Johnson okay EO Johnson so EO Johnson is a method inside power transformation those who have learned the statistics I think they know okay what is Yo Johnson okay there are so many method you can use but yo Johnson is like so popular one okay now you can see uh I'm applying on top of this particular company is and number of employee okay you can see this is the company ede company ede and

number of employee because this particular privilege Bas is almost good I don't need to apply the power transformation here but this two columns is having problem so I'll be adding this particular two column here you can see now once you fit it and again if I convert to the data frame then again if I like plot this particular variable right now you can see company is and my number of employees is getting normally distributed okay this is the power of power transformation okay yo Johnson Technique we can apply to get the um actually normal distribution from the

skewness data okay I hope this part is clear all of you so this is what you have to apply in the feature engineering guys okay feature engineering is what you are changing the data okay you are playing with the data and EDM means you are only analyzing the data yeah R seeing you cut now if you if you check the skes again you will see there is zero skewness in your data right now not zero actually some of the skes is still there but it's far better than before right yeah now you can see the feature encoding and scaling technique okay for this I'll be using uh first of all I'll

get my numerical features you can see the numerical features now for the preparing for the column Transformer okay see now what we have to do okay what we have to do first of all I have to apply one ordinal columns okay that means some of the columns is having category it is having category if you see this particular column is having category like has job experience or not yes or no request job training yes or no full-time position yes or no education of the employee like Masters Bachelors okay so if I'm having this particular column guys if I'm having this

particular column let's say I'm having yes no so which encoding technique I can apply guys one hot one hot would be better all or ordinal encoding would be better tell me one hot encoding would be better or ordinal encoding would be better let me know guys um not one not exactly because if you apply one not so what will happen guys at the end so one not will create like that now so let's say this is my yes this is my no so whenever it is getting yes it will be one it will be zero okay whenever it is let's say uh no it would be zero it is one that means

you are creating two Dimension at the end two dimensional Vector you are getting creating again in machine learning Dimension is a issue okay so I can use the ordinal encoding so ordinal will apply some of the number let's say it would be zero it would be one only okay in one dimension we can handle this particular situation yes that's how you can think about okay which one would be better so I can apply one ordinal encoding so that's why here have written ordinal encoding I have to apply here okay o r means ordinal encoding okay now one o

means one H encoding okay there are some current L continent unit of wage okay regions of employment if you see these are the columns is having more category okay that time I have to apply one coding okay now some of the columns I have to apply the transformation that means the power transformation okay Power transformation like a number of employee and Company is I think I saw you one experiment right because this is this qess now if I want to perform all the jobs in one chance guys what I can use I think you remember we are having one function

called Escalon pipeline okay escalan pipeline I can use if you have learned escalan maybe you know what is escalan pipeline right so pipeline means let's say you want to perform one not encoding you have to perform orinal encoding you have to perform transformation okay transformation so what you can do you can create a pipeline so let's say this is my pipeline okay in the first stage it will perform the ordinal encoding second stage it should perform the one encoding third stage it will perform the transformation okay and fifth stage it

will give you the data okay and what are the columns you have to do so here you have mentioned already you can see this is the one ordinal encoding this is the one encoding this the transformation so everything you can perform inside a pipeline object okay now you can see I'm initializing the pipeline and here first of all I'm ining the standard scalar W encoding ordinal encoding okay and this is the complete Pipeline and I'm using Y Johnson meth me okay for the power transformation and here I'm writing the column transformation object so first of all I have to do the one in coding on top of this particular column o column

then I have to perform the ordinal encoding okay I have to perform on top of ordinal column then I have to perform the transformation on top of the transformation column okay then I have to do the scaling operation that means entire data I have to do the scaling operation just to get the gradient descent okay quickly just to reach the global Minima quickly because whenever your gradient desent would be calculated it be quickly happen Okay that's why I'll be doing the stand escaline technique okay so that's how you can create a pipeline object okay you can create a pipeline object escalan pipeline object instead of doing this

thing individually then I'm fitting the data you can see and after fitting the data so this is the numerical representation of my data this is the final representation okay now this particular data you have to use for the analysis okay you have to use for the analysis that means you have to use for the model training now I can start the model training but before that I have to handle the balance problem so here you can see I'm using smart Okay smart and with the help of smart actually I'm U doing the sampling okay I'm doing the sampling and I'm trying to handle my

imbalance problem okay here I'm handling my imbalance problem now once imbalance problem is handled now you can do the T speed operation it's just a simple T speed we are doing okay then after that we are initializing the model you can see we have import different different model as I already told you we are not going to use only one model okay we are not going to use one model we'll be using list of the model okay we'll be using list of the classification model OKAY classification model we'll be using now we'll be using all the model one by one and we'll be training our data and we'll see we'll just log the

accuracy okay we'll loog the accuracy in a table okay then we'll try to see which model is working better we'll picked up this particular model again we'll do the hyper parameter tuning okay hyper parameter tuning okay then after hyper parameter tuning we'll decide which one I have to select this particular model okay so that is what actually we are doing instead of using one model we are importing all of the model as well as the accuracy score Matrix and all okay then we are creating one evaluation

function you can see evaluate classifier so inside that we're calculating the accur accuracy F1 score prision recal Roc AOC card and we are returning it okay one by one now here if you see here is a model object I've created and this is the model different different model random for decision trees classifier gradian boosting classifier logistic regulation X boost CAD boost SVC that mean support Vector classification addab boost classified so this is my model list okay now here we have function we have

written uh called here we have function we have written called evaluate models okay now here I'm giving my X data and Y data as well as the model list you can see the model list I'm giving and here I'm doing the T split of my data and what is happening I'm writing one full loop and first of all I'm getting the model one by one okay and I'm doing the feat operation on top of my data and whatever Matrix actually I'm getting after doing the calculation okay that is the evaluation calculation you can see we are doing the evaluation calculation

here evaluate calculation we are doing we are getting our recall model training AOC R okay then we are also getting accuracy code we just saving in a we're saving in a you can see we're saving in a table okay we're saving in a table even we are also printing in in the terminal okay we're printing in the terminal okay then once I got all the accuracy I'm saving in the data frame and I'm returning the report now now see if I execute this particular line this particular code like evaluate model and if I give my data and if I give my model see this is my final result I got so

this is the random forest model and this is the model performance so you can see this is the training set performance so accuracy 100% F1 score 100% Precision 100% recal 100% EOC 100% And for the testing data accuracy 95% F1 95% prision 95% 96% 95% Okay so that's how you can see different different model decision trees for accuracy you can see and now I can open in a text editor you can see all the model performance okay can we uh just get the model for the personalized parameter you can use it but you can also use random s CV now you can see the gradient boosting logistic regression now that's how we

can understand okay which model is working better so I can plot this particular table I created that means the data frame I created now see this is my base module result and here is the final result so Random Forest accuracy 95% then car neighbor okay then car neighbor you can see it's 94% then ex exib classifier it's 93% okay now see I will be selecting only two model two to three model not the entire model I can see random forest and K nearest neighbor is working better than any other model what I can do I can select this particular two model and I can do the hyper parameter tuning on top

of the two model because I'm not sure if I'm applying hyper parameter tuning my K NE might work better okay than my random Forest because I'm not sure I have to apply first of all then I have to decide now you can't decide I have to use the random forse because still hyper parameter tuning is left after doing the hyper parameter tuning you will be deciding whether you will be using this particular model on not okay so that is what we'll be doing later on see so we have selected the exib Boost model so we have selected three model I think random first kist and exib okay see this is the

exost parameter this is the random first parameter this is the kist parameter and we are using random SAR CV and we are doing the hyper parameter tuning here you can see okay hyper parameter tuning so kold validation is happening after that we are getting the model parameter that means the final result we are getting the final result you can see okay then again we are training the model by random Forest K neighor and X boost okay now this is the final result I got guys now tell me which model is working better random Forest exib boost or

k definitely K neighbor is working better but previously I saw random forest was working fine now kest neighbor got 97% accuracy random Forest got 70 sorry 95% accuracy and exib got 94% accuracy that's why I told you hyper parameter tuning is required you can't decide after training the model first of all apply the hyper parameter tuning then decide okay which model you'll be using now now you can get to know okay K neighbor is working better here okay so I'll be selecting this particular model as my final model that is what I'm doing

kers neighbor I picked up then again I'll just train the model and I saw the accuracy score and you can see it is doing the very less mistake okay than your previous model okay that this is how actually entire end to end implementation happens guys okay it's not like that you are selecting one model you are fitting the data it's not like that okay as a data scientist you have to do the all of the experiment then you have to decide which one you'll be using okay for the product C itself so you can execute this particular code guys I'm not executing because again it is going to take time

hyper parameter tuning text time right so you can execute because we are doing the live streaming so I can't wait okay so you can execute in your system okay it will work fine I already executed the output I got that's why I'm showing you okay in front of you now you can see this is the my final analysis I got best model is K NE which is having 97% accuracy okay 96 because previously I'm getting 96 okay now it's fine I think 97% I okay yeah imbalanced we have solved now so me already yeah so this how actually we have to do the feature engineering as well as the model training okay after doing the

exploited data analysis part now in the next class we'll be converting this particular entire notebook experiment as our modular coding okay that means we'll be writing the entire pipeline okay you can see different different component we have added inje transformation validation evaluation Pusher we'll be writing each and every component and before that we have done the notebook experiment and this particular notebook experiment has you has given you the entire clearcut idea what we have to do throughout the entire project okay so tell me guys uh how is

this particular U like uh project how is this particular notebook and everything is it fine like uh are you able to learn some new topic yes or no let me know apart from whatever you have learned so far whether it's a new concept or not tell me because I know that you have done so many machine learning project in your career but the thing is like I just wanted to show you some unique unique things okay unique way you can do the project great guys how was the session let me know if you like the session so please give me a like guys please give

me a like you just like the session and please try to subscribe to my channel guys if you haven't subscribed please try to subscribe uh let's try to hit 12,000 subscriber okay yeah so next class we'll be doing the data ination component datation component will be writing from Scrat okay yes and make sure you are ready with the uh this one actually um yeah mongod setup because from the mongod itself we'll been justing the data where you will get this code I have already updated my code in my GitHub so this is my GitHub so here every code is

available in the notebook folder it is available uh Kumar abishek so let me share this particular Link in the chat and please try to execute guys this particular notebook so only attending session W be helping you unless and until you are not doing the practice from your end okay so once you execute the code from your end once you try to read everything whatever I have written things would be much clear I know that okay so yes guys this is all from this particular session for today uh I hope you like this particular session now I think we can uh end the session and if you have any quer you can

ask me otherwise I think I can end the session I'll take some of the doubt then I will I think end the session okay you can let me know guys when will C langen session start whenever we'll start the jna session that time we we can start the langin because it is the part of J okay fine I think uh okay fine I think um we can conclude a session guys thanks for joining again and uh please try to support my channel so that I can bring lots of content for you with that guys uh bye everyone good night uh take care everyone bye thank

Show thinking

This session focused on the experimental phase of the machine learning project, covering Exploratory Data Analysis (EDA), feature engineering, and model training using detailed Jupyter Notebooks. The goal was to understand the data and find the best-performing model before building the automated production pipeline.

**📝 Agenda for the Day**

* **Exploratory Data Analysis (EDA):** Analyze the dataset to uncover patterns, identify problems like outliers and imbalances, and generate insights.
* **Feature Engineering:** Clean, transform, and create new features based on the findings from the EDA.
* **Model Training:** Train multiple classification models, evaluate their performance, perform hyperparameter tuning, and select the best model for the project.

**📊 Exploratory Data Analysis (EDA)**

The presenter walked through a comprehensive EDA notebook, emphasizing the importance of documenting insights along the way.

**Key Findings**

* **Irrelevant Features:** The **case\_id** column was identified as a unique identifier with no predictive value and was marked for removal.
* **Skewed Data:** Univariate analysis of numerical features (like number\_of\_employees and company\_age) revealed they were **highly skewed**. This indicates a need for data transformation to help machine learning models perform better.
* **Imbalanced Target Variable:** The target column, case\_status, was found to be **imbalanced**. Approximately 66% of the applications were "Certified," while only 33% were "Denied." This imbalance must be handled to prevent the model from being biased toward the majority class.
* **Feature Correlation:** A **Chi-square test** was used to check the correlation between categorical features and the target variable, confirming that most features had a significant relationship.

**⚙️ Feature Engineering and Model Training**

Based on the EDA insights, a second notebook was used to prepare the data and train the models.

**Data Preparation Pipeline**

A **scikit-learn Pipeline** was used to streamline the feature engineering process, which is a best practice for production code.

1. **Handling Skewness:** A **Power Transformer** (using the Yeo-Johnson method) was applied to the skewed numerical columns to make their distributions more normal.
2. **Encoding Categorical Features:**
   * **Ordinal Encoding** was used for features with a clear order (e.g., education level).
   * **One-Hot Encoding** was used for nominal features with no inherent order (e.g., continent).
3. **Handling Imbalance:** The **SMOTE** (Synthetic Minority Over-sampling Technique) was used to create synthetic data points for the minority class ("Denied"), resulting in a balanced dataset for training.
4. **Scaling:** StandardScaler was applied to all numerical features to bring them to a similar scale.

**Model Selection**

* **Multi-Model Evaluation:** Instead of choosing a single model, a wide range of classifiers were trained and evaluated, including Random Forest, XGBoost, Gradient Boosting, and k-Nearest Neighbors (k-NN).
* **Hyperparameter Tuning:** The top-performing models were selected for **hyperparameter tuning** using RandomizedSearchCV to find their optimal settings.
* **Best Model:** After tuning, **k-Nearest Neighbors (k-NN)** emerged as the champion model, achieving the highest accuracy of approximately **97%**.

**🚀 Next Steps**

With the experimental phase complete and the best model identified, the next session will focus on converting these notebook steps into a modular, automated **machine learning pipeline** using the project structure created in the previous sessions.

e the easiest way to make content is to use an AI the easiest way the easiest way to is to use an a

the easiest uh hello everyone good evening I think I'm audible to all of you just give me a confirmation in the chat the easiest way to uh hello everyone I think guys uh you can hear me you can see me uh just uh give me a confirmation whether I'm audible to or not uh yeah hi suras sir I want to learn develop zat to B chatbot using large language model so in my I think YouTube channel I already have one playlist on top of it let me tell you so there is a playlist I created called chain late I think I think it is available inside video section there is a video I created

called chain lit let me show you huh so this one how to build llm uh app super fast with chain lit so there I showed this particular zatu chatbot kinds of application SAS so you can check this particular video okay yeah so uh yeah hi everyone I think uh now we can start the session and today I'm not going to teach you anything so today I just kept this session for the question and answering so here I'll be taking some live doubt if you are having any kinds of problem with this particular application the application we are implementing or any kinds of doubt you

are having you can ask me in this particular uh live session so I'll try to clarify each and every of your doubt okay because what I feel like only teaching is not a main thing okay um in between we also should interact with our audience so that I can get to know okay if they are facing any difficulty level or not and if uh they need something from my side okay I can try to provide that particular content as well okay so that's why I kept this particular session guys uh and I'll be keeping uh this kinds of session more in future so whenever I'll be teaching something in

live definitely there would be some of the session for the doubt resolving and question answer okay these are the things we'll be doing so uh I want to start learning mlops currently I'm Miss data scientist what skills should I start with so if you want to start with mlops first of all uh just try to uh good with the machine learning deep learning whatever let's say you are interested in then you can start with the mlops because mlops is nothing but just a tools and Technology at that so whenever you are implementing something you are adding some automation tool okay with your

project okay this is called mlops Pipeline and all okay I got I think you got your answer sirum then uh uh yes C I'm audible anur thank you so helloy I need to understand how to perform data versioning and the model versioning with the complete endn project so if you want to do the data versioning at that time you can use this particular tool called DVC okay DVC that means data version controlling so this particular tool will help you to do the data version controlling okay you can see this particular tool okay this is the data

version controlling tool okay apart from that there is another tool you can use called kro I think so kro is one of the uh open source framework for the data science code you can see that you can also use this particular tool okay this is the alternative of this DBC and if you want to vers on your model that time what you can do you can use this particular ml flow okay ml flow can also track your model it can also register your model okay with a different different version so you can use this particular tool and in my

YouTube channel I already created the content related ml flow you can see ml flow for data science so there uh each and every video I have already discussed if I show you so let's say introduction to MLF flow setting up the mlflow server okay how to set up this ml flow server on AWS even I also show you one end to end implementation okay with a realtime demo so maybe it this particular uh playlist will help you to understand this particular things okay the things you are asking I hope this part is clear fine so guys if you don't know we

have started one amazing playlist called mlos production ready machine learning project okay so this is the playlist you will see so here we are doing one end to-end implementation okay we are doing one end to end implementation of a machine learning project with the help of mlops integation so you can see this is a completely production ready project we are implementing so we have completed till day three and we have completed like expert dat analysis fature engineering and training on Notebook experiment and today is the doubt session I'm taking only because I just wanted to make sure everyone is having

um like clearcut understanding or not so tomorrow or in the next session what I will do I'll just try to start with day five there will be implementing the actual component that means the pipeline we'll be creating the data in transformation validation OKAY model training model valuation even we'll also see like how we can do the deployment of this particular entire application on the cloud platform so each and everything we're going to discuss so make sure you just try to follow this particular playlist and if you're new to my channel so what you can do guys you can subscribe to my channel and please

try to like this particular video and try to share as much as you can okay let's try to hit guys uh 12,000 subscriber in my channel okay as soon as possible because if you are motivating me definitely I will bring lots of content okay for you only so there are so many things in my pipeline I'm planning to bring up in this particular Channel okay as a community with free of cost guys so here um uh yes audible okay hello I just want to ENT project okay I have clarified this particular answer thanks also if possible if you can build

uh aent and function calling rag on custom connected yeah whenever we'll be starting with the jni definitely we can implement this particular project as well can you explain the hypothesis testing in the detail in the feature uh selection yes I can explain so let's say this is my notebook I think you are asking for this particular um hypothesis testing let me show you so here is the code we are doing this particular categorical features yeah so this this is what I think you are asking right so here see what we are doing we are applying the multicolinearity detection that means what is

multicolinearity like like colinear like I mean relationship between the two columns okay or multiple columns this is called multicolinearity okay I think you know that so let's say we have done the multicolinearity in the numerical features okay in the numerical features we already know with the help of correlation plot we can easily get to know but what about the categorical features because category is not a number it's an word actually it's a one on individual actually uh individual actually different different and

so how to calculate this particular corelation okay categorical on top of categorical column for this we'll be using this particular chai uh chai 2 square sorry yeah chai to continency I think this is the name uh yeah chai square test so this particular test we can apply okay this particular test we can apply and we can get to know whether this particular column is having U like any relationship with the target variable or not okay so here I did this particular hypothesis testing and it is telling continent is like getting

rejected null hypothesis that means there you can see here that means it is having some correlation with the target features okay it is having some correlation with the target features but whenever it is giving fail to reject null hypothesis that time it is telling there is no cor correlation with the target featur here you can see I have already reported here requir job training is fils true null hypothesis which means it doesn't correlate with the target columns that means whenever you are getting reject null hypothesis it is correlated with

the target column and whenever it is not having the relation that means there is no relation okay with this particular Target column and if there is no relation you can remove this particular column okay it's not required for our analysis okay I I hope this part is clear right now so yeah hello AR is mobin so in this course can you explain the kubernetes and CD pipelines for okay yeah so CU Nies will be following in a different project so this is what actually we are implementing something basic okay so we'll be also

doing kubernetes uh and any other mlops tool okay we'll be also integrating in this particular um I mean uh Channel also okay don't need to worry so why did you use chai square test for the checking multicolinearity in your Ed notebook initially we should vif test for the checking multicolinearity see there are so many ways suit so it's not like that you only need to use the vi testing for the multicolinearity there are so many things you can apply okay if you know the statistics I think you know there

are so many things you can apply so that is what I'm using this particular SI Square test because it is having one library in the scipi okay I can easily get to know that particular information whether it is having some relation with our Target column or not okay instead of applying the vif test you can also apply vif test it's it's up to you okay there is no issue with that uh but you are doing great job thanks Aris uh even you should also do the great job guys so I want to see your uh like great trans transition and all

uh if you are learning something um okay from my channel so it would be really proud for me okay so I'm able to teach you something at the end okay so if this channel is helping you so definitely I'll be more happy with that so please try to put your dedication on guys try to learn as much as you can and definitely uh if you have the proper dedication you can do anything okay tell me one playlist for the statistics to the study sir okay for statistics maybe you can use any kinds of playlist what whatever you feel like it's good you can go with that particular playlist okay there there won't be any kinds of suggestion from my side because all the

content are good I have seen all of the content over the YouTube so it's just uh like your choice which particular exploration you will prefer okay so right try to see two to three more playlist okay and you can decide which one you will go ahead with okay yeah but don't try to learn the entire statistics because it's a completely bigger field uh whatever actually concept is required re to learn the data science only just try to learn that particular concept only okay that will help you a lot because you have to do the Smart Learning so don't dedicate

much time on a specific topic instead that just try to learn whatever you need okay and try to um be I mean uh active on the Practical implementation always if you are good with the Practical implementation so uh definitely will get a good opportunity at then but if you are if you want to go with the resource position and all if you want to let's say bring something new for the audience what you can do you can also explore the theoretical understanding in a deeper way you can understand each and every mathematical concept okay this will help

you are a lot whenever you are doing any kind of research but whenever you want to be a developer you want to develop something I'll suest try to more focus on the Practical implementation and definitely theory is required try to understand the theory whatever you required for that particular implementation okay don't need to go in depth because again if you're going in dep so it will take lots of time to cope up that particular concept even people will also get confused like sir I don't know this particular mathematics okay it is having some issue so that's why I always suest try to understand as a high

level and try to implement okay in a lowlevel way okay that is what I will be sesting to everyone I'm from uh Telecom domain how to prepare resume for the data science transition can you arrange a session data science resume preparation definitely I will bring a session on so it is in pipeline I'll will be taking like interview preparation session then resume discussion session okay in this particular Channel only so I think uh you can wait for this particular session okay yeah even I will also share you some of the uh like sample resume template so that you can also follow

whenever you are preparing your own resume that time sir can you explain the cross attention in the Transformers uh see this is not uh uh like today's agenda like today we are not implementing any kind of Transformer and all whenever we're doing um NLP kinds of let's say community session that time definitely we'll be discussing this part okay Charan okay so please ask guys if you have any kind of question you can ask me today or if you having any kinds of issue you can let me know

and guys this is the project we are implementing so this is the GitHub link I think you already know so let me share this particular link again uh so from here you can get all of the code and please try to start this particular repository guys uh because I can't see only two people has start so definitely you should uh appreciate something if you're getting right so I'll uh expect guys you will do the star here you'll also Fork this particular repository okay okay so I appreciate for an interview today for uh okay I appeared for an interview today for jna

interviewer didn't ask about jni instead as the solve uh use cases I just want to know if anyone is using apis for llm then why ask the random question see it depend again upon the interviewer what kinds of question they may ask but you should prepare with all the content okay if you're going for an interview uh it should be related to gni it should be related to other topic as well okay definitely you should give some proper answer there okay but uh mostly I can say people will ask related the jni question but I don't know why they are asking related random kinds of question okay yeah so uh how to delete the data diff

how to okay how to Det the data diff so this is the agenda for the next class V we'll be learning the data diff detection with the help of evident so there is a tool called evidently AI so with the help of that we'll be detecting the data da okay in our data so this is one mlops tool we'll be using in this particular project API yes used for um using the llm model right API yes used what using llm what kinds of API you are talking about because there are uh kinds of model uh they use API to interact with the U let's say uh user

there are some of the model it will interact directly let's say uh if I'm talking about open source lar language model which is available in the hugging face okay hugging face you can see hugging face is a platform so it will give you different different large language model okay so let me Zoom this my screen little bit so you can see this is the Hing face and it is providing different different large language model so these are the model are open source and if you want to use this particular model you have to download this particular model in your system then you can use it but there are some model let's say Jin is there okay jini is

there then we are having something called open AI model is there okay so these are the models are API based do means uh if you want to use this particular model you have to use their uh API API key to access this particular model that means you will be hitting the request in their server and it will give you the response you don't need to download this particular model okay okay that's how you can use it uh can you try to build a uh mlop production ready machine learning project on realtime data and historical data this time using apis yeah real time

data wise the same thing the data ination you have to do let's say the data ination we'll be doing uh in this particular project we'll be doing from the uh database but the datation you have to do for the API for the API management let's say there are so many API provider okay API provide data API provider data API there are so many data API provider you can see so you can use one of the API and you can fish that particular data from the API itself okay only the data injection pipeline you have to change there nothing else uh how to write the detail uh rme

related to the projects uh what we are uh things we have to in the redme file maybe we are also managing the rme file you can see that's how you have to manage the proper commands and all uh the proper uh setup each and everything you have to mention let's say if I'm sharing this particular repository with someone else they should get to know okay so that's how we have to install this particular uh project in our system that's how we have to mention proper the command and uh some of the description related to your project if you want to

add some images and give file you can also add it here that's how you can manage so I think remember there is a file called readme.md so here you can write each and everything and it will reflect here okay it will reflect here so anyone having any question guys you can ask me let me know guys can I build a jni project the user will give image based on the image the J

model want to give the medicine to the C can we can build the model gen projects to the user will give the image based on them image model want to give the medicine I didn't get your uh question Shan just write in the easiest way hi buy can you uh include feature store in any of your project yeah feature store YC uh then what is Feature Feature is nothing but uh whatever artifacts you are generating okay this is called features you can uh store it in a three bucket and all okay it's up to you even your data should be also

considered as a feature okay if you want to develop uh I saw some graph like structure in a rme file yeah so graph also you can add in the rme file uh so there is a uh rme file tutorial you will get on Google okay you will see that how to write the proper rme file there with a markdown and each and everything and there is some tool actually you can use for the rme reader so rme Doo okay so this is the uh website you can use to write a beautiful readme file let's say if I do the get started now see here you can uh select different different let's say I want to take the

title then uh acknowledgement okay appendix author see right hand side you will get the preview that's how you can use this particular tool and you can write a proper readme file okay beautiful readme file uh I have mlops interview in today's how should I explain my project can you give a Barlow and sequence structure so that I can explain so see first of all try to explain the problem statement I always say um like you have to explain the problem statement you are trying to solve and based on that whatever solution you have designed okay the solution you have to discuss and the

tools and Technology you are using why you are using that particular tools and Technology that's how you have to start your project okay uh mlts I have been following you especially your jni videos yeah thank you even jni also will try to see in this particular Channel because previously I used to channel I used to take the jni in a different Channel Channel even uh I also used to take some courses okay yeah now we'll be doing in this particular Channel as well any more query guys please uh ask me and if you want this kinds of session guys you can let me know whether this doubt session is helping or

not what about MLF scort that we are going on uh see I'm not sure actually I already left the company so I won't be able to uh answer your query did that yeah maybe uh he will cover can we build the jni project uh the user give skin progran bacteria image first on the image the J model want to give to the medicine to the C yes you can also do it for this you have to design that particular uh let's say architecture okay that particular image architecture and all you have to so for this you can use vertex AI platform there you will be

getting uh Vision model mostly okay can you uh create some videos on how to deplo ml models in data breaks um let's see uh I'll just try to note it down okay I'll see that I want to uh know what are the things that we need to know inside jni so that I can say that I'm ready for the jna engineer job see in jna what are the things you have to know okay like see gni first of all you have to know the basics of gni okay what is gni exactly why you have to use gni basics of NLP is required because if you want to understand the text and all if you want to understand the entire llm model or not so you have to learn the

basics of NLP inside NLP you have the RNN architecture G architecture lstm architecture encoder and decoder attention mechanism okay so try to learn these are the things then after that you can start with the text preprocessing and word embedding technique okay once it's done what you can do you can do some projects actually handson projects it is available on the hugging face platform you can do it okay then try to start with the hugging face platform so in hugging face also you are having different different model with different different task let's say you want to do uh let's say you want to do um translation task so select the

translation and these are the model related to the translation now let's see if I open any kinds of model okay see code is also available so code snipp it also available here I think somewhere they have given the code uh if I go back so somewhere they have also given the code snipp you can use see code snipp is also there even they have also provide the tutorial let's if I go to the spaces uh so there you will get the tutorial I think somewhere okay documentation then inside that you are having the task and here is the different different tutorial let's say

you want to learn the translation task so here is the translation if I click here so they're having their official tutorial you can also follow their official tutorial you can also learn okay then once hugging face is clear you you can start with the open a how to use open models and all okay you can uh use this particular open you can learn this particular open after that you can start with something called Vector database okay Vector database so Vector DB is one of the like uh I mean important Concept in jna you have to know after that you have to learn something called Lang chain okay Lang

chain this is a gni framework you have to use then you can also learn something called Lama Index this is the alternative of L Lang chain so both you should know I think in current market okay then once you covered the Lang chain Lama index everything okay so you can explore some open source large language model so there are some open llm is also available open llm uh so you can see open llm there are so many open LM is available uh you can see that particular leaderboard and you can explore different different open llm okay you can see like how we can load this particular llm how we can f tune that

particular llm okay how to create custom application on top of it then you should learn something called rag retrieval augmented generation okay rag in jna so this is the uh retrieval augmented generation you have to learn so how to create the rag application and all okay that's how you can um explore each and every toping okay after one by one okay so if you know these are the thing I think I I think it's enough um then you will be able to create any kinds of GNA based application okay uh check out once sir um hi I have your mail address

please increas of the queries I want to connect you through mail you can send me a mail if you have my mail uh buty can you please cover some mlop TOS like uh yeah definitely I'll try to cover [Music] uh what I want to check Rohit check out once I didn't get okay I think I have clarified all of your query guys uh let me know if you have any other query [Music] okay so uh how to evaluate rag so for this you can use ragas okay ragas there is a package you can use

ragas uh ragas GitHub so this is the framework you can use to evaluate rag application ragas okay you can explore how to evaluate the rag VI so tell me about Google pal is the completely free yes Google pal is a free so I think I've created a tutorial on top of Google pum you can also watch this particular recording see Google pum open llm but it is uh like depricated uh nobody use Google pom right now so everyone will use the JY right now okay uh how to evaluate your rag application or LM models I told you you have to use something called ragas okay

ragas uh will help you to the rag evaluation technique I'll create some content on Rags like how to evaluate your rag Pipeline and all you can see evaluate framework for your retable augmented generation that means rag pipeline even this is like more popular application the last comit was 5 days ago very active research okay you can see the contributor and all so you can use this one so all the code snippit they have given every tutorial is available even there they have their beautiful documentation also you can also follow

this particular documentation to learn like how to evaluate your rag pipeline can you suggest me some free hosted llm like Google pal uh you can use jmin that means Google AI studio so ji API from Google a studio you can yeah you can use Okay Google AI for developer so it will give you the free API as well okay so see from here you can get the API yeah so I think I have clarified all of your doubt guys I think uh no more query so uh next class guys we'll be start we'll be continuing with this particular project again so we'll try to

implement the component first of all we'll start with the data in and we'll see like how to detect the data diff okay with the help of evidently then we'll try to complete this particular project okay um side by side so yes I think now we can end the session uh and if feel free to guys comments here um if you are having any of Doubt even I will also conduct this kinds of session in future so if you have any kinds of doubt and question and all you can drop me um message here I'll try to clarify all of your doubt uh how to use your ml flow

repo with uh different ml models just try to watch this particular playlist you will get the idea okay how to use the MLF flow with different different take tag okay let's say machine learning de learning computer vision everything will learn from here only no no fine guys so please uh guys try to support my Channel Try to subscribe to my Channel Try to like this video try to share okay as much as you you can and uh um I think we'll be learning so many things from here and please try to join the next session guys next session is super important so there we'll be learning the pipeline creation we'll be

doing the modular coding and each and everything okay until here uh just try to be with the setup try to be with the like ready project whatever things I have showed you if you are getting any kind of error you can let me know okay and uh I think next class we'll try to start with our main uh pipeline integration yeah so uh yes guys thanks for joining the session I hope you like this particular session with that guys bye everyone uh good night take care have a nice day bye everyone thanks

Show thinking

This session was a special **live Q&A and doubt-clearing session** for the ongoing "MLOps Production Ready Machine Learning Project" series. Instead of a coding tutorial, the presenter took questions from the audience regarding the project, MLOps, career paths, and Generative AI.

**💬 Key Questions from the Session**

**MLOps Project & Tools**

* **Q: How does the Chi-Square test work for feature selection in the notebook?**
  + **A:** It's used to check the relationship between **categorical features** and the target variable. If the result is "reject null hypothesis," it means the feature is correlated with the target and is likely useful. If it's "fail to reject," the feature isn't correlated and can be considered for removal.
* **Q: How do you handle data and model versioning?**
  + **A:** For data versioning, you can use tools like **DVC (Data Version Control)**. For model versioning, experiment tracking, and registration, **MLflow** is an excellent tool. The presenter mentioned having a dedicated MLflow playlist on his channel.
* **Q: How will you detect data drift in this project?**
  + **A:** This will be covered in an upcoming session using the MLOps tool **Evidently AI**.
* **Q: Will Kubernetes be covered?**
  + **A:** Yes, but in a separate, more advanced project in the future.

**Career & Learning Path 📝**

* **Q: I'm a Data Scientist. How should I start learning MLOps?**
  + **A:** First, ensure your fundamentals in machine learning and deep learning are strong. MLOps is primarily about learning the tools and technologies to automate and productionize the ML lifecycle.
* **Q: How should I explain my project in an interview?**
  + **A:** Follow this structure:
    1. **Problem Statement:** Clearly define the problem you solved.
    2. **Solution Design:** Explain the solution you architected.
    3. **Tools & Technology:** Discuss the tools you used and, most importantly, *why* you chose them.
* **Q: How can I prepare my resume for a data science role?**
  + **A:** The presenter will hold a dedicated live session in the future to cover resume preparation and interview strategies.

**Generative AI 🤖**

* **Q: What's the learning path to become a GenAI Engineer?**
  + **A:** The recommended path includes:
    1. Mastering NLP fundamentals (RNN, Attention).
    2. Becoming proficient with the **Hugging Face** ecosystem.
    3. Learning to use API-based models like Gemini and OpenAI.
    4. Understanding **Vector Databases**.
    5. Mastering frameworks like **LangChain** and **LlamaIndex**.
    6. Learning about **RAG (Retrieval-Augmented Generation)**.
* **Q: How can you evaluate a RAG application?**
  + **A:** You can use a specialized framework called **Ragas** for evaluating RAG pipelines.

**🚀 Next Steps for the Project**

The next session will resume the hands-on coding, starting with the implementation of the project's core components (data ingestion, transformation, validation, etc.) as a modular pipeline.

 for uh hello everyone good evening I hope I'm audible guys to all of you just let me know in the chat okay guys I think I'm audible you can hear me you can see me so I will start the session within one minute uh let's

uh let's wait for 1 minute guys so that everyone can join and we can start the session uh yeah hi hi everyone so guys I think you are doing well and uh sorry for the delayed actually I was uh uh having some work actually so that's why I couldn't able to continue the live when I think you saw that uh I just uh um like did one setup at my home I have taken one new system at my home that means the desktop so that's why I was doing the entire setup and uh now currently I'm um I'm live actually uh with the particular system so that's why I had to do the setup for all the things actually I

usually do in my previous system right so that's why actually it got delayed so there is a video guys I think I have posted uh if you if you want to see that like what kinds of system I have uh uh like uh built you can see this particular video uh so the specification and recommendation I shared because many people are asking like uh sir what would be the best configuration PC for the data science right uh or let's say machine learning deep learning whatever we are doing so that's why I thought let's create this

particular video so whatever specification I'm using so everything I have shared there you can go ahead and watch that particular recording okay you'll get the entire idea okay so now I think we can start the session guys so before starting the session first of all let me tell you uh um uh why you will get the previous video and previous resources so here you can see I have created one playlist called mlos production ready machine learning project so if I open this particular playlist guys you will see we

have completed till day four and this is the day five today so we have completed till I think uh experted analysis feature engineering and training on Notebook experiment all the resources I have shared in that particular uh description section if you open that particular description section you will able to see this particular GitHub repository okay so this is the GitHub repository guys so from here you can download all the codes all the material everything I have shared in that particular session itself okay you can

see whatever live link I shared everything it is available here so if you have missed uh missed out that particular session guys what you can do first of all try to complete this particular recording then try to join this particular uh like session otherwise uh some of the things you you will you otherwise some some of the things actually you won't be able to understand that's why that would be my suggestion first of all try to complete till day four and try to uh watch this particular day five okay because today

we are going to start with our modular coding so uh definitely you need the previous knowledge to implement this particular entire uh pipeline okay so if you're ready guys you can give me a quick uh yes in the chat or you can give any kinds of emojis so that I can start with let me know guys if you're uh if you're ready guys all of you just give me a quick yes thumbs up anything guys in the chat so that I can get to know uh you are ready and I can start the session okay great so now first of all I will open up my Blackboard and let me write down the agenda like what are the things

we're going to cover in this particular session okay so today is the day five so today is the day Five Guys the day five and we have completed till our notebook experiment like we saw the entire project notebook experiment like how we can um uh do the Ada part how we can do the feature engineering part how we can do the model training and everything we have saw right on The jupyter Notebook so from today actually we'll be starting with our uh component implementation the first component we'll be implementing today called Data

injection data injection okay data injection component will be implementing so uh what is data injection exactly the data injection is nothing but it just a way of getting the data from the database I think you saw that I was storing the data where I was storing the data in the mongodb class TP okay in the mongodb database and this is the cloud database so there actually I was storing my data you can see this is my database called e Visa inside that I have created a collection called Visa data and inside that I have stored all

of the data actually I was having in that particular CSV file right so now what I have to do guys in the notebook I showed you how we can F this particular data from the database like now what about the component whenever I'm writing modular coding okay whenever I'm writing model coding so how to write this particular component there okay that is what actually we'll be explaining in this particular component today that's why I kept this particular name is as a data inion right and with that actually we'll be also learning with that we'll be also learning so this is the uh

additional thing we'll be learning today uh called uh data D data drift okay with evidently EV dently okay evidently so evidently is nothing but it's a mlops tool ml ml Ops tool okay if time allows I will try to discuss this evidently today otherwise we'll be discussing this particular evidently tomorrow okay tomorrow because tomorrow actually I'm going to use this particular evidently without uh like pipeline so I think it would be better to discuss this part tomorrow because today I just wanted to discuss this particular data in s and I want to explain the entire code in

easiest way so that everyone can get to know okay like how we are going to write the entire code okay whenever we'll be also implementing some other component as well okay if you understand the data this so you can also understand something called Data validation so let me just um just a minute guys let me just reduce the size H so you can also understand data validation you can also understand data transformation transformation you can can also understand model training you can also understand model evaluation OKAY model Evolution and you can also understand

something called Model Pusher model Pusher okay so everything you can understand because the coding structure will be following okay the coding structure will be following so everything will be same okay for all the component so that's why I want to discuss this particular data inition in detail so that everyone will have the clearcut understanding on the data inition that's why I will be discussing this particular data D with evidently tomorrow okay tomorrow I'll keep a separate session and we'll try to see how we can detect the data diff with the

help of one amazing mlops tool called evidently I hope guys this agenda is clear all of you let me know if this agenda is clear with all of you so let me give you a quick uh revision like how we can inest data I think you remember I created one graph previously as well so for the data in what I have to do let's say this is my uh hi sir I'm uh currently working on your NN projects uh deep learning using MLS I get an error for Ure uh what is the error you can let me know um I'll try to give you the solution uh Vu uh definitely I think I'm going to conduct uh class tomorrow as

well okay you will see that particular schedule from my YouTube channel only yeah uh Jo thank you so much uh for your valuable um feedback okay so now what I can do guys I can show you the entire like architecture like how we are going to in data see uh for this actually what I can show you um I can show you one high level diagram first of all so just a minute guys okay so see this is the like high high level diagram so let's say this is my uh let's say this is my mongodb this is my

mongodb database so this is my mongod DB database here I can write DB database and here I'm having my data which data guys I'm having I think you know that we are having one data called US Visa okay or I think the name was Visa data right Visa data Visa data. CSV so this particular data is available okay this particular data set is available inside my uh inside my uh mongodb okay mongodb database now what I have to do guys what I have to do so I have to write one component I have to write one component called Data inition data injection okay data injection so

data injection is going to be my component okay okay so this is going to be my first component in a pipeline because we're going to create a pipeline after data ination my data validation will run after data validation my um this uh transformation will run after transformation trainer will run then evaluation will run then Pusher will run so as you can see data data actually inje data injection is my first first component it is my first component of this particular pipeline so you can see this is a complete pipeline it's a

pipeline okay it's a pipeline and data inition is the first component of the pipeline okay so we are completing this particular first like component first of all so what I have to do guys I think you saw that I have written one script with help of that I will just connect with my mongodb I'll just connect with my mongodb okay with the help of connection string I think you saw that we just collected one connection okay connection string okay connection string from the mongod DB server itself I think you remember um we connected one connects to

the string so let me I think you remember otherwise I can show you just a minute let me open up my code editor so this is the code guys and here in the mongodb demo so here I think I mentioned that particular connection string somewhere so here see this is the connection URL but I removed that because uh I just wanted to keep it a secret I I don't want to share with you so because otherwise other people will access my account right that's why I kept I just deleted okay from here okay but after some time I'll show you this particular connection string because we also need to set up this connection string inside my environment variable

okay so make sure you have your connection string if you don't have try to check out this particular video the video actually I have taken so database setup and project utility so there I showed you how we can set up our database and how we can get that particular connection string okay so with the help of connection string I will just make the connection with my database now what I have to do from this particular database I have to fetch the data I have to fetch the data okay I have to fetch the data

F the data okay so then it will give me my Visa data it will give me my Visa data Visa data. CSV okay because again I will get a like dictionary kinds of output because you can see your mongod stored your data as a dictionary format you can see it's a dictionary format so I just need to convert this particular dictionary to um data frame and that particular data Fame I will try to convert in a CSV file okay in a CSV file because the same thing I have done inside that particular notebook as well you can see so first of all what I was doing I was reading my data I was converting this

particular data to the dictionary format okay then I was pushing this particular data in my mongodb okay mongodb cluster and so this is the connection actually I was doing then after that I was fetching the data so then after fing the data I was converting this particular data to the data frame and this particular data frame I was removing this particular ID column then this then this particular data frame what you can do you can save as a CSV file okay you can save as a CSV file okay uh so I do I do have an error while

connecting the mongodb from Office network uh it seems it's a fair well issue the mongodb address should not be you can try from your system that would be uh better V okay instead of using the office system at test case ID did you mean theet regx which project I'm not able to sh uh Arad actually I'm not sure like which project you are executing uh is there any other way to handle that particular issue uh you can use your own system uh your own network maybe you can do it SS handshake bu with the trying to mongodb

python client see for me it is U like connecting even for everyone also they're able to do it okay but I'm not sure why it is getting uh this particular error for you okay maybe you can check with your system it will work fine so this this is the like high level understanding of our data in okay so we'll just try to like make this particular architecture in a low level and we'll try to understand how we're going to quote this particular example so for this what I have done guys I have created one uh flowchart so this flowchart is available inside your repository as well let me show you so if

you open up your uh if you open up your uh code repository so this is the code repository guys so inside that I just created one uh flowchart folder and inside that I kept this one uh this particular flowchart called Data inion PNG okay okay data in PNG now if I open it up so you'll see the entire flowchart the entire lowlevel flowchart okay I have I have kept here so now see this particular flowchart will help you to understand this particular flow like how it is ingesting the data from the mongod and to inest the data what are the

component what are the configuration it needs what are the variable it needs right so everything I have mentioned here okay everything I have mentioned here so that's how also you just need to prepare your flowchart any project you are doing just try to prepare the flowchart so that if you are referring this particular project after 6 month after one year just by referring that particular flowet you can easily get to know okay that's how my component is working that is the best practice guys I can um like I can tell you because I saw that people

will start with the project they will directly write the code okay so it's not a like I mean efficient way to write a code first of all try to prepare the documentation proper documentation then try to write the code it would be easy for you to understand like what is happening okay so that's how what I have done guys I have created this particular flowchart for you so let me open it up again so this is the enter flowchart guys you can see this is the lowlevel flowchart I have created so here what I'm doing guys you can see this is my mongodb server okay this is my mongodb

database and from here what I'm doing I'm just just inting the data for this I'm using one data access okay data access I'm using and to access the data that means to face the data I need some I need some helper variable I need some helper variable I need some helper constant you can see I need the data invest directory like after investing the data why it will keep the data in which folder it will keep the data in my local machine so these are the variable these are the variable okay I have to pass as a constant I'll tell you

whenever I'll write the code what is constant what is entity what is um like artifact entity I'll discuss each and everything as of now just try to understand if I want to face the data and if I want to keep somewhere in my local system I need a folder okay that particular folder name is my data injection directory in that particular like variable I mention the directory the inside the directory I I'm going to store my data okay then I need the feature store file path because you can see after ingesting the data let's say

this is my data in config it has triggered it will inest the data from the mongodb that means it will initiate the data in after that it will export the data as a feature store that means what will happen it will face the data as a like dictionary it will try to convert as a data frame that particular data frame will be converted as a CSV file and this particular CSV file it will store in a feature store folder so there would be a folder called feature ReStore in the artifact okay artifact is my data ination directory inside that there would be a folder called feature restore inside feature restore this

particular data would be saved okay then what I have to do in the data I think you you know that in machine learning first of all we do the tenter split operation okay first of all we have to do the tent split operation then we have to do the further processing like whatever feature engineering whatever Ed we are doing everything we can perform not Ed actually feature engineering you can do after doing the T split okay but for Ed actually you can keep the entire data so you can see guys I will do the T

split operation after doing the T split operation what will will happen it will give me my training set as well as the testing set so this particular training set as and testing set I will keep inside a folder called ingested folder so one folder would be created called artifact inside that my feature would be there inside my like root USV file would be saved and after doing the 10 T speit I will create another folder called ined inside that I will keep my train.

csb and my train. csb so this is the entire idea of this particular flowchart okay so that's how my data injection will happen Okay so I I think you got this particular I think you got this particular understanding how things are working and this is the entire flowart guys I'm sharing with you so you can open it up in your system and you can try to understand okay how things are working now we'll try to refer this particular flowchart and we'll do the coding with respect to that only okay so now let's go to my code editor so this is my code editor guys let me close this particular tab now here first of all

what I have to write guys first of all I have to write the workflow okay like how we're going to update this particular file because here you can see there are lot so many file there are so many files and now one question would be in your mind like which file I need to change first and which file I need to change last okay so that's why first of all let's try to understand the workflow see workflow wise what I have to do guys so inside US Visa you can see there are so many folder constant configuration

component entity exception okay logger everything so exception logger we have already written we don't need to write it so first of all let's try to write the workflow like which file I need to change at the very first okay so artifact means the like generated output let's say after running the code whatever output it is generating I have to save somewhere that is called artifact folder well go okay now here let me just comment it out so this is my project workflow so work workflow so in that particular workflow so first thing we be executing sorry the

first thing will be updating called constant okay constant okay constants this particular constant folder we we have to update first of all okay then the second thing we have to update uh called U config or here I can just write U entity entities okay entity enti okay entity now first of all let me update this particular constant so let me open this particular constant uh folder and inside that you are having one file called underscore uncore inore dop because this is your uh this is your uh like uh this your Constructor file inside that particular Constructor file I'm going to mention my constant I just

landed to this particular lecture are we also covering the data engineering here no big data we are not going to cover has okay big data um we are not going to cover here yeah I already created the virtual environment for this particular project in my first video you can check it out see that okay now inside this particular constant I have to mention these are the variable that means my data interation directory feature store file path okay training file path testing file part training split ratio collection name everything have to mention so I already created all the variable let me show you

guys how this particular variable will look like so this is going to be my constant so these are the constant actually I need so these are the common constant so let me show you so these are my common constant okay these are my common constant you can see so here what I have written guys you can see here I have written uh my database name so what is my database name okay what is my database name I think you you know that this is my database name if I go to my mongodb this is my database name called US Visa so that's why I kept this particular data name as US Visa you have

to give the same name the name actually you have created in your mongodb first of all try to check the name you have created and give the name here database name okay then you have to mention the collection name so you can see this is the collection name inside US Visa I'm having Visa data okay Visa undor data so this is my collection name okay this is my collection name so here you can see inside actually collection name I have mentioned this particular Visa undor data so this is going to be my connection name okay now mongodb URL key that means the mongodb URL that means that collection string

okay the collection string I collected okay the collection string I collected okay from the mongodb now what I will show you guys how to um mention this particular mongodb URL so for this what I can do I can first of all take that particular connection string so let me just copy just a minute I kept somewhere this is my connection string yeah so you can see that guys uh I think I can open this particular file and at the last I can mention it so this is my connection string guys you can see this is my insire connection string okay with the help of that we'll be making the connection with our mongodb now I

have to export this particular connection string as my environment able okay then this particular uh variable will able to sorry this particular variable will able to load this particular connection string from my environment variable I think you know that what is environment variable so in Windows actually will get one option called environment variable so if you start here environment variable so you'll see this particular option environment variable okay so here what you can do you can click on this particular environment variable and here you can create a system variable to create the system variable what you can

do guys you can create a new you just click on new then after that you just need to give the variable name so the variable name should be mongodb URL so I'll copy this particular name and here I'll just try to mention it so this is the mongodb URL okay now you have to define the variable value so value is nothing but this is the value the inter connection string so I'll copy okay and here I'm just trying to mention this particular uh connection string okay now once it is done just click on okay and see it has saved inside my environment

variable okay now with the help of operating system package I can load this particular environment variable why I kept this particular credential inside my environment variable because I don't want to share with my public I don't want to share with my audience okay so that see if you are creating any kind of secret file if you we are using any kind of secret credential in your project instead of keeping hard code what you can do you can keep inside environment variable and when you required it at that time you can read from the environment variable that should be the

most secured way okay to create a project that's why I'm showing this particular uh way like how we can kept our credential inside environment variable and how we can read it later on okay that is why actually I'm showing now if I just click on okay see it has been saved here okay it has been saved now I can easily load this particular envirment variable with the help of my operating system package okay I'll tell you how to load it later on I'll tell you how to load this particular uh environment variable with the help of operating system packes okay this is the

first approach okay you can follow those who are using actually command prompt if you're using anaga prompt guys if you're using anaga prompt or command promp to execute this particular project you can use this particular method and if you're using G bash guys so those who are using G bash those who are using G bash you can also create inside EMV file it is also fine well go you can create a EMV file inside that also you can mention okay there are so many way you can manage your like secret variable okay but I'm showing you this particular option because this is like most easiest

one you don't need to install any kinds of like package you can directly use your environment variable option inside windows and you can load it and those who are using gitbash guys for you there is another way you can follow so let me first of all clear my terminal see here I'm using my G bash so let me make a big screen so this is my G bash guys okay so those who are using G bash so what you can do you can execute one command so the command I already shared inside the readme file so let me show you I think

it's not there so let me just copy the command somewhere I kept this particular command so this is the command guys as of now I don't need the my AWS credential I'll just try to remove so see guys this is the command you have to execute so export mongodb I'll copy this particular command I will go back here and I'll take a new cell I'll try to mention now I'll copy the entire connection stream and here I'm going to mention it okay here I'm going to mention it so this is going to be your command guys export mongodb URL equal to you have to give the entire connection string now I'll try to copy the entire

command and here I'm going to execute inside my inside my uh V sorry inside my G bash now if I execute see this particular credential would be set as my environment variable okay this particular credential would be set as my environment variable so there is the two approach you can follow if you're using command prompt you can directly set inside your environment variable inside your Windows machine and if you're using any other operating system like Linux or Mac you can use this particular command export command and if you're also using gitbash to execute your project you can also use this particular method okay to

set your connection string I hope guys this part is clear all of you right now what I can do guys I can remove this particular connection string from here as of now because I'm not going to share with you so let me delete and I'm going to also delete this particular connection string so the command wise I have shared here guys so you can refer this particular command only you just need to replace your connection string here okay that is the idea now I'll save it now this particular file also I need

to save so here I can save it I just need to press contrl s and this particular file also I need to save here so now see this particular three variable I think you understand these are my constant see what is constant constant is the variable actually we don't uh see constant is a variable actually will be changing frequently let's say um the database name database name okay see here we are not going to use the same database name like I mean all the time so we may change the database name but if I'm hardcoding this particular value inside my code so what

I have to do whenever I need to change that particular database name I have to open up my code I have to open up my code inside my component inside my data inition there I need to change it manually so it is like very hardic task because let's say there is a like so many file and folder there is a very big code base you are having now finding that particular variable would be very much difficult for you so that's why instead of keeping that particular like variable as a hardcoded what I have to do I have to create a constant file inside that I can mention this particular variable and whenever I need

it whenever I need to change it I can open this particular constant file and here I can change it and everything will uh reflect inside my code so that's how I need I don't need to open up my code and there I don't need to change it manually so that is the automated way actually I'm following that's why we call it as a constant so whatever constant variable you need for your project and it should be changeable later on you can mention inside this particular constant and um you can change any time okay uh by opening this particular file I hope this part is clear guys what is constant let me know

so that's why you can see here I'm keeping this particular constant these are the constant variable I need to run my data injection pipeline okay great now uh you can see here I have updated my constant okay constant is updated now I also need to update my data in related constant because as of now I have updated some of the common constant you can see my mongodb URL my pipeline name US Visa artifact directory that means it will create artifact directory okay inside that it will save everything and whatever model actually want to save after training it would be the model.

pel that means we are going to save our model as a pel format okay that's why we kept this particular variable now I have to set some of the constant for the data in like data inion directory feature store file path okay these are the thing I have to like mentions for this I already prepared for my data inition related constant so these are my data in related constant guys you can see so these are my datation related constants so here what I have mentioned guys constant mean fix yeah you can consider as a fixed but later on if you

want to change it okay you can do it well it go okay now see guys data in collection name it's a Visa data okay it's a VA data you can see collection name is a Visa data then my data in directory name it's a data in directory that means inside a artifact folder it will create a folder called Data ination inside that it will store my data that means that's how it will create the folder first of all one folder would be created called artifact okay artifact inside that it will create a data ination folder inside that it will create a feature restore and in justed it would be more clear whenever I'll

execute this particular program and whenever I'll show you the output it would be more clear okay as of now just try to consider these are some variable I have taken and these are some folder actually I have to create okay during my data in and this is my Trend to speed ratio that means uh 80% data I want to keep for my training purpose and 20% data I want to keep for my testing purpose okay so these are my data in related constant I as of now okay great now the next thing I have to update uh called um entity okay I have to update my entity you can see in the

rme file I have mentioned I have to update my entity now let's open up my entity so here you can see this is my entity inside entity you are having two kinds of entity called artifact entity and config entity okay now let's try to understand what is config entity and what is artifact entity see what will happen so let's say this is my data in this is my data in data inje okay so to run the data in what I need I need some constant definitely I need some constant I think I show you I need some constant so I can

just yeah so I need constant okay see this constant is nothing but it's just a path actually you can see at the end it's just a folder that means it's a path because if I want to create something inside data ination folder I have to U Store inside feature store that means it's a path actually then it's a path I have to return as a path okay so to manipulate this this particular constant I need something called config entity so there is another things actually called config entity config entity config entity so what

config entity will do it will take the constant and it will make the path it will make the popup path okay this particular path will be provided in my data inion component okay and data indion would be executed and it will give you two things it will give you two things one is my uh train do CSV and this is like my test. CSV test.

CSV you can see here after running my data in what I'm getting train. csb and test. csb this is my output that means this is my artif yes or no guys this is my artifact right after in like executing the DAT I'm getting 10. csb and test. csb so this is my output okay now if I want to execute the next component what is the next component guys after data in I think you know which is nothing but my data validation data validation okay data validation now if I want to validate the data so what I have to do first of all I have to load my train train. CSV and test csb then I

have to check this particular data whether this particular data is in correct format or not if there is any data dat problem or not okay so that means what I'm doing I'm taking this particular tr. CSV as well as that test. CSV inside my next component which is coming after data ination which is nothing but my data validation that means you can see the output I'm getting from my first component this is my first component this is my second component that means the output I'm getting from my first component that is the input from the next component that is the input for the next component which is

nothing but my data validation that means this is my artifact this is my uh this is my artifact this is my artifact okay these are my artifact okay so this is called actually artifact entity that means we are getting some entity okay artifact entity that means we are getting some output and this particular output we are sending for the next component which is nothing but my data validation again data validation will execute it will give some output it will give some output and this particular output I have to send to my next

component which is nothing but my data transform okay data transformation okay data transformation again data transformation will give you some output okay I'll need to send this particular output to my next component which is nothing but my model trainer OKAY model trainer again model trer will give some output I'll send this particular output to the next component which is nothing but my model evaluation OKAY model evaluation okay after model evaluation again it will generate some output I will send this particular out output to my next component which is nothing but my model Pusher that means whenever I'm executing

the PIP pipeline it is taking some input it is generating some output this particular output is getting to the next component again it is giving some output again it is going as an input for the next component okay so this is the entire Pipeline and the output actually we are getting this is nothing but my artifact entity this is nothing but my artifact entity that means the output I'm getting that is called AR defect okay this is called AR defect okay I hope this part is clear right now okay now config entity is nothing but the constant we are manipulating with the help of config entity and artifact

entity is nothing but the output we are getting from the component which is the input for the next component this is called AR defect entity okay now let's try to write as a code okay how to write as a code so I open up my code editor so I will first of all Define the config entity so inside config entity what I can do I can uh Define the entire config entity just a minute um config entity so first of all I'm going to import some of the library here so first of all let's import operating system because with the help of

operating system we will be manipulating the path then I also need to import all of the constant here see all of the constant I'm having here okay so I have to import so how to import so here I can just do from US Visa so it is available inside my US Visa usor Visa inside that I'm having a folder called constant okay constant so here I can call this particular constant and I need to import okay I need to import everything from this particular constants for this I'm going to give this particular start icon okay I'm going to give this particular start icon okay that means I'm importing

everything okay I'm importing everything from this particular constant star means I want to import everything okay star means I want to import everything okay then what I have to do guys I have to uh again import another Library called Data class I'll tell you what is this data class as of now let's import data class and there is a method called Data class okay then I also need dat time Library so I can just write date uh time okay input uh date time okay date time so these are the package I need as of now so first of all I have to create a date

time that mean date time time stamp so it will create that it will take that current time stamp okay and it will save inside this particular variable called time stamp okay so here I'm going to create a class and this is not actually python class this is a data class you can see there is a decorator I have assigned called Data class so data class means it's a class only inside class actually without self actually I'm writing the variable see when we Define the class how we Define the class I think you remember if I am defining any class let's say class name is ABC okay

ABC this is my class inside that I just Define a Constructor okay I just Define a Constructor called diff underscore uncore okay underscore underscore init underscore uncore okay inside that I just mention the self okay then after that if I want to like Define us like class variable what I have to do guys I have to first of all take the self okay then I have to mention the variable name let's say variable name is test inside that I need to store some of the value let's say value is none okay as of now

none so that's how actually we usually Define class variable but here you can see instead of using self I can directly Define this particular class variable because this is a data class right now this is not a python class it's a data class okay with the help of this particular data class decorator you can Define the class variable without using any kinds of self method now let me delete this particular line now you can see what I'm doing the class name is training pipeline inside that this is the pipeline name so the pipeline name you can see the pipeline name I'm getting from the constant if you just press control and right click on top of

it it is coming from the it is coming from the constant itself okay now artifact directory so what is my artifact directory guys artifact directory you can see it is also coming from my uh constant so artifact and I'm just attaching with my time stamp that means it will create a time stamp folder inside that it will inest my data that means every time it will create a different different time stamp and that particular time stamp it will store my data this part would be more clear whenever I'll execute the code guys as of now just try to consider it will create a time stamp folder inside that it will inest the dat okay so that I

don't need to replace the data again and again okay because I need all of the experiment all of the artifact okay generated from my model now again this is a time stamp I'm returning here so this is a simple data class I have return now I need to write the data class for the data in but before that I need to first of all call this particular uh I just I just need to call this particular class so here I'm calling this particular class guys you can see so this is my training pipeline class I'm calling so this is the object and this is the variable and this is my um sorry sorry this is actually type

this is a class type this is my variable and this is my class name okay so here I'm calling it now inside that I need to Define my data in related constant you can see these are my data in related uh configuration so the class name is data inition and inside that I'm just creating a different different variable the variable you saw here my data inition directory feature restore file PA training file PA testing file path everything I have mentioned here okay one by one and this is the type type of this particular variable this is string type this is string type this is string

type this is the floating type this is the string type okay now inside that you can see with the help of o. part. join I'm defining all the path that means my training pipeline path first of all artifact directory and data in directory name okay then my data in directory data in uh feature directory name okay that's how I'm preparing all the path now file name is missing I think here so let me see guys what is file name here so this is the file name that means my data name okay my data name so this is the data name guys the data name is es visa. CSV I think

you saw that this is the data name okay this is the file name so this particular file name I need to save inside my ECT after downloading the data from my mongodb right so now I think everything is fine now there is two more variable is missing train file name and test file name so let's mention it here as well so train file name inside my constant so here is my training file name okay so train. CSV and test.

CSV test file name okay now let me say save and if I come here now see all the uh variable I have defined okay and this is simple python code guys with the help of w. part. join I'm just joining the path I'm not joining the path I'm just joining the path that's it okay that's it now you can simply understand you can print this particular uh things okay in a python uh python interpreter you can open it up and you can execute you can print it up that's how you can understand okay but I'm expecting guys you have this much of code level understanding okay here I'm not writing very fancy code I'm just only using w. part. join I'm just

joining the path that's it and from where this particular variable is coming you can see here you just press control and right click sorry left click you just press control and left click if you left click it will automatically redirect to you in that particular like variable okay that's how you can understand okay from where this particular variable is coming okay so this is nothing but my config config entity okay this is nothing but my config entity okay now see config entity is covered now I need to assign The artif Entity now I have to assign the artifact

entity now what is artifact entity you can see here so after doing the data in artifact will generate 10 csb and uh test. csb and tr. csb so these two file actually it will give me so let's define this particular entity as well so I'll open up this particular artifact entity so this is going to be my artifact entity again it's a data class uh the name of the data class is data in the artifact and this is my train file path and test file path and this is going to be a string string type path okay the string type variable that means it will return this particular two file and this particular

two file I need to receive in my next component which is nothing but my data validation okay I hope this part is clear right now okay so now you can see uh my uh config entities part is also clear that means we have updated my constant as well as the entity now the next thing what I have to do I have to make the connection I have to make the connection with my mongodb so to make the connection guys what I can do there is a connection folder I'm have sorry there is a configuration folder I'm having inside that I'm going to create a folder um sorry inside that I'm going to create a file called mongodb connection so

let's create a file here so inside connection configuration I'll create a file I'll just name it as mongodb so mongod DB mongoor DB uncore connection okay connection so why I'm creating this particular file separately I'll tell you first of all let me create yeah connection now inside that see the same code we have written the same code we have written in our jupyter notebook I have written in a I have written in a modular way that's it you can see here there is nothing new there is nothing new see I have created a let me select my environment

Visa now see pong be disappeared see so I'm importing first of all my exception login I think you know what is exception login because every time I I'm handling the exception as well as the logging also here I'm just doing the logging I think somewhere uh so logging I think I was doing somewh here see logging I'm Al also doing the logging operation so you can see here first of all I imported database name and mongodb URL from my constant you can see database name and this is My mongodb URL okay and from where mongodb URL would be getting it will get from the it will get from the uh envirment variable

itself okay so you can see here I'm loading now why I'm using certify because you remember I was getting the connection issue connection uh like I mean timeout issue somewh I think you were getting that connection timeout issue so to prevent this particular connection timeout issue what you can do you can use this certifi package okay certifi package okay now I have created a class called mongodb client inside that first of all this particular client would be none first of all I have to make the connection with my client see the same thing we have written inside this particular notebook uh for the connection see first

of all I was connecting with my mongodb URL creating the database then collection name the same thing I'm doing here first of all I'm making the connection but to make the connection first of all I need to load the mongodb URL and here you can see w.g EnV w.g EnV okay with the help of w.

g EnV I'm loading this particular mongodb key okay let me show you so I will open up my demo I can copy this particular code and here uh I can open up my demo. piy let me delete this particular code and let me m it here now here I need to import my operating system so import OS okay now weight. get EnV and what is my environment name in the constant it was mongodb URL I'll copy and inside that I need to mention it okay in a string I need to mention it okay now it will automatically load this particular mongodb URL from where from my environment variable okay I think you saw that inside environment table I already set it up here it it is already

there mongodb URL it is already there okay now let me show you so if I open up my terminal and if I execute this particular demo. Pi so python demo. Pi uh so there is nothing because I need to print it so print this particular mongodb uh mongodb URL let me save and again if I execute see this particular connection URL I'm getting this particular connection URL I'm getting automatically from my environment variable that's why we have kept my credential inside my environment varable and that's how we are reading this particular things okay with the help of get EnV and first of all I'm

taking mongod is none or not if none I'll raise the exception environment key is not set otherwise I'll make the connection you can see I'm making the connection the same way mongodb client the same connection I did inside my uh notebook as well okay you can see the same connection I'm doing that means the same code guys I've have written inside my jupyter notebook only I'm writing here as a functional way I'm in a modular way okay you can see after creating I'm just creating the data database and I'm just um returning this particular database name okay that's it and if some exception occurs okay what

I'm doing I'm just raising the exception okay my USB is exception that is the things actually I'm doing now see why I kept this particular Connection in a separate file in a separate file in a configuration because let's say whenever I want to do the data inition I want to do the data inition so whenever I have to make the connection with my mongodb what I have to do I have to write this mop code yes or no I have to write this masop code so if I'm writing this much of code here so it would be like I mean more messy okay so

there would be lots of line and to manage this particular code it would be very difficult for me it will look like more messy right so that's why what we kept we cre we created a different file and inside that I kept my connection string so whenever I need it whenever I need it whenever I need to make the connection with my mongod I will call this particular class only and it will automatically make the connection and create the database for me that is called actually modular coding instead of creating all the code inside one file you are creating a different different different file different different folder and there you keeping your code

and whenever you reord it you are calling that particular code from there only this is called modular coding and this is called object oriented programming okay that's why I told you oops concept was required and uh if you want to understand the modular coding you need the oops concept understanding okay that's how actually we write the production grade code in the company as well okay so this particular code would be very uh clean code and you can manage this particular code in easiest way okay going forward okay I hope this part is clear now let me save it and now I can

close this particular file okay now the next thing guys what I have to do if I cancel this one now if I open up this particular um I'll close this one this particular notebook see the next thing I was converting my data to a data frame that means whatever data actually I was getting from my DB I was converting as a data frame so how to convert as a data frame so for this I'm going to create another class in a different folder so here what I can do guys I can create a um I can create a just a minute H so inside US Visa I can create

a folder called Data access okay dataor access dataor access inside that I'm going to create a Constructor file because again it's a folder so I'll just write underscore uncore init uncore do PI okay this is going to be my Constructor folder now inside that I'm going to create a file called uh Visa data or I can just keep it as US Visa uncore data okay us visor dat.

Pi okay inside that I'm going to write the logic to convert my data from dictionary to data frame so again I prepared this particular code for you so let me show you again I'm sharing this particular code as a template guys you don't need to write from Scrat so if you're creating any kind of project you can use this particular code as it is okay no need to change anything okay it's automated code now you can see here guys I'm importing my client from the configuration the configuration I created so inside configuration mongodb connection you can see usva configuration mongodb connection and I was importing this particular class and

with the help of this particular class I was making the connection inside my data uh inside my uh here you can see here uh this this is the connection I was doing that's why I kept this particular class separately if I want to make the connection I can easily import it and I can use it okay now exception pandas and some of the optional Library have imported so here I've created one class so inside that I've created a Constructor so here I'm making the connection with my mongodb and if some exception occurs it will raise the exception and here the final method I have written export collection as data

frame that means you need to give the collection name database name okay it will return your data as a data frame object you can see first of all it will check the database whether database is none or not if uh it is none first of all it will create the connection name otherwise it will make the connection and it will get the data as a data okay it will remove the idid column and it will replace the Nan value if some Nan is there it will replace the Nan value and it will return the data it will return the data frame for you so whatever things I was doing manually here whatever things I was doing manually everything I've kept in a

everything I've kept in a here I think it is available Inside Yes Visa okay everything I have kept in a function in a class okay in a modular way okay so whenever I need it I will call this particular method and it will do my job automatically for me okay this is called modular coding instead of writing everything in a one one place I can separate out okay and whenever I require it I can use it from there only I hope this part is clear all of you let me know I'll share this particular code guys with all of you no need to worry so after that you can um like get that particular code in your system and you can execute from from your end okay that

would be more clear and guys um please try to explore this particular code it's not like that you are only watching this particular lecture it will help you a lot if you want to understand the code definitely you have to execute you have to execute in your system then you will able to uh understand what is happening here okay yeah so if you have any query guys you can ask me in between so it is ready data access I can close it uh as of now config entity I can also close demo I can also close mongodb connection also I can close okay that means I have created my data connection

data access that means export my data as a data Fame now finally I can work with my components okay now here I can mention after entity uh I can work with my components so components okay components so let me open up my components data inition inside components I'm having the data inition and here is the data inition code so it's like very simple code guys first of all let's import some Library so I'm importing pandas data frame T speed because I told you I have to do the test speed during data inion okay

that's why I need this and data ination config so definitely I need to import my configuration the configuration you have prepared because it is going to return all the path and I need to also import my data in artifact because it is going to be my return type that means for the next component what would be the return type so that's why I have imported this particular artifact entity as well okay then exception login and US Visa that means my export collection as a data frame okay this particular class I'm importing okay so these are the thing I

need for the data in now here I'm going to create a class called Data in so this is the class guys I've created called Data inition inside that it is taking data inition configuration the configuration I'm getting it from here and it is making the configuration object and if some exception occurs it will give the exception okay to all of you now inside that I'm going to create a method called uh export the data into feature store so this is the method guys I've created called export the data into feature store so this is the dock string I have written and from here your code

is starting inside Tri block first of all I'm logging the information getting the US visard data okay you can see this is the US visard data so this will give you the US visard data okay you can see export data as a collection I'm getting the data frame I'm checking the shape of it after that what I'm doing creating a directory inside that particular directory I'm saving my data as a CSV format okay and this is the feature restore file path I'm getting from my data in configuration and it is coming

from here only inside from from your configuration itself okay so that's how actually each and everything is connected with each other if you want to understand first of all try to open this particular code with your code editor press the and uh control and do the left click okay for uh like that you can understand from where this particular variable is coming okay now once my data has been exported inside my feature I have to do the 10 spit operation so this is what I have to do so again I have written another function here so this is the function guys called

split the data as a train train and test okay so you can see I'm logging the information and I'm doing the 10 to top operation after that I'm saving my training set as well as a testing set inside my training file path and testing file path okay the path actually I prepared here the path actually I prepared here inside my constant as well so this is my constant so here the path actually I have prepared okay I hope this part is clear now to execute this particular two function that means my export data sorry that means my export data into

feature is store and speed data as a 10 10 and test speed okay I have to write one final method so this is called initiate data inition initiate data inition you can see so here this is the return type data in artifact because I told you it will return 10 file path and test file path that's why this is the return type return type of a function in Python you can also mention the return type of a function okay now inside that you can see first of all I'm exporting the data that means I'm calling this particular method calling this particular method

exporting the data in a feature store and once it is done I'm logging the information then I'm doing the 10 spit operation you can see I'm calling this particular method okay inside this particular method that means if you execute this particular method this particular two method would be automatically executed that means one method is connected with each other okay one method is connected with each other okay this is called modular coding now once it is done I'm just defining my output that means my artifact that means as a artifact I will get my training directory as well as the testing

directory as well as the testing directory okay as an artifact you can see as an artifact my testing directory as well as my training directory for my next component and once it is done I'm returning my data in artifact and my exception is handling here now if you want to test this particular data in whether it's working fine or not what I can do I can open up my pipeline so inside US Visa I created one folder called pipeline now I need to update my pipeline after components fourth file I need to update which is nothing but my pipeline so pipe pipelines okay pipeline I have to

update I think it's a pipeline not pipelines let me save so now I'll open up this particular Pipeline and inside that I'm having two pipeline prediction Pipeline and training pipeline as of now prediction pipeline is not required I'll only update my training pipeline I'll open this particular training Pipeline and inside that first of all I need to import some of the libraries so let me import one by one so first of all I need to import my login exception and system package because I was handling that exception for this I need the system package then I also need done my data in

the data in component I created here so this is the class data in so if I want to import it so what I can do from US Visa okay us visa. component it is available inside component component folder and there is a file called Data in and here I need to import okay I need to import data inertion class the class actually I have created here okay then with that I also need to import guys my config entity as well as my artifact entity so let's import as well so from US Visa so entities available inside the entity folder you can see inside entity

folder my entity available so I can UT so from use Visa do entity do config entity okay so the first config entity I need to import which is nothing but my data in config okay then similar wise I also need to import from my artifact entity so us visa. entity and here I now need to import the artifact entity and inside that I have my uh data in artifact okay so these are the thing I need to import then once it is done I need to create a class and inside that particular class I need to initialize my data inion so so this is the class guys called training pipeline inside that I have create a Constructor and this is

going to be my data inition configuration okay this particular uh object I have initialize fine so now here I'll create a method called start data inition so inside that I'm going to create a method called start data in okay you can see so here what I'm doing guys I'm just calling this particular data inition I'm just calling this particular data ination object because this is going to be my final class and inside that I'm having this particular method called initiate data ination so this particular initiate data ination if I execute that means it will

execute your expert X export data as a feature store as well as the T speed that means I only need to call this particular method and everything would be executed okay so here what I am doing I'm just logging the information initializing my data inition and it will take the configuration that configuration I'm passing and from the object itself I'm calling the initiate data that means this particular function I'm calling okay sorry this particular function I'm calling and once this particular function is called that means your data in will start and it will inest the data inside the aric folder

and this would be clear guys let me first of all execute uh everything would be clear in your mind okay like what is happening here now once it is done uh I also need to create a final method called run pipeline okay because I need to run the pipeline so here this is going to be my method so see these are my common function going forward we'll be writing everything inside this particular fun function only that's why I told you if you understand the data in going forward all the component will be writing you can understand easily now see run pipeline one of method I've created uh this is

the dock string and inside that I'm initializing this particular start data in so this data in is calling here now once it is done I also need to uh fill out my exception block okay exception block I also need to fill out that means if I call this particular run pip planine my data transation will start now how to call the Run pip planine I'll open up my demo so demo. p let me open it up demo.

Pi so let me close everything now first of all I need to import uh so from US Visa us visa. pipeline okay so here I'm having something called training pipeline import so this is the class name I created called train pipeline okay now let me initialize the train pipeline in object so OBG is equal to train Pipeline and from the train pipeline is it OPG dot I'm going to run this particular run pipeline method okay so this is my run pipeline you can see if I execute the Run pipeline my entire pipeline will start now let me save and if I open up my terminal if I clear it see as of now there is nothing there is no artifact folder but if I execute this

particular demo. PI right now see my data would be ined from my database it will take some time make sure you set that particular environment key okay otherwise it w be working I already said that's why I'm executing so it's like very easy code guys uh only I have written in a modular way so sometimes you feel like okay um uh it's like a little bit complex but if you go through this particular code if you execute from your end definitely each and everything would be clear so make sure guys you execute this particular code from your end okay see currently my

will be ingested here so let's wait see my data in is done see my execution is successful there is no error there is no error that means it executed successfully now see artifact folder has created one artifact folder has created guys I think you can see artifact folder has created inside artifact one time stamp folder is created you can see this is the current time stamp the time stamp I have executed my program so you can see the current time stamp and dat this is the execution of my program inside

that it has created a folder called Data inition inside that feature stored and ingested that means the same thing we have written here artifact time stamp time stamp I been mentioned in this particular like flowchart but you can see feature restore and reged feature restore anded inside feature restore it has saved my usb. CSV us visa.

CSV inside inate it has saved my 10. csb and test. csb and this is my data guys okay you can see this is my data that means my data inition is successful my data in is successful from my data itself okay with the help of modular coding that means the code I have written as of now everything is running fine okay without any error okay so this is called artifact and every time it will create a different different time stamp let's say if I execute this particular code for the second time let's say second time I'm executing so what will happen let's

see so instead of replacing the previous file it will create a different time stamp inside that it will in the data that means I'm not losing my previous information this is called Data version controlling okay or your artifact version controlling I think people are asking sir how to do the data version controlling how to do the model version controlling that's how you can do the controlling okay with a different different time stem see execution is done now if I open up my artifact see this is my first execution this is my

second execution okay and again I got the same output I hope guys this part is clear and let me know guys the technique actually I'm following whether it's a uh I mean unique technique or not or you are enjoying this particular technique or not because uh this is the unique things okay I'm showing you uh if you have any better like idea you can follow because again end to implementation is completely Your Design pipeline this is the pipeline I usually follow okay and uh I saw like lots of Industry are

following this particular like pipeline okay this particular technique to write the code so how is this particular project guys let me know how is this particular code how is this particular project so let me commit the changes as well so you can see lots of comit I have to do but again I don't want to commit uh these are the artifact the artifact it has generated you can see because artifact I don't want to push in my GitHub so what I can do if I want to ignore it so I can open up this particular file called Dog ignore inside that you have to mention the artifact

so here I will mention the artifact folder so artifact I don't want to push it was artifacts or effect AR effect okay artif only inside that I want to ignore everything start now if I save it now see guys all the artifact has been removed now I can commit the changes so here I can write data inition data inition added commit and sing the changes okay so now if I go to my GitHub refresh the page so all the changes I think I have made here so data in added now let me uh give you this particular repo so that

you can get that code so this is all about our data injection part guys uh that's how we can inest that data and this is the entire explanation I have given on the notebook side as well as well as the modular coding how to write the modular coding and that's how we we need to write the modular coding and this is what ENT implementation happens okay so please try to see this particular code try to practice try to see like from where I'm calling the variable and all it's like very simple python code I'm writing guys only in just a class and Method okay I'm

just writing the code little bit hard for you valo so that's why I'm telling you first of all try to clone this particular repo open it up and try to understand this is the best way to debug the code okay otherwise by seeing the lecture I I know like still you will have some of the confusion okay unless and until you are not executing from your end so please try to execute definitely you can understand okay and you can write your own code that time okay yes now you can if I follow any kinds of let's say um like real world

implementation let's say if I follow K uh K GitHub let's let's let me show you the kasas actually implementation so inside kasas they're having different different apis let's say if I want to see the activation function so see guys uh so these are my Conant activation application so denet implementation uh only these are the things they have kept but where is the con uh yeah it is available inside layer okay so what's the board implementation Optimizer sh okay I think it is available inside SRC yeah SRC activation activation. Pi now you can see guys they

are also using uh function okay functional code they're also using class code okay classical code functional code they're also implementing in the same modular way they're also using decorator and all okay yeah yeah we'll see that uh lakman okay uh maybe I can um do the same thing okay yeah so I hope guys you uh got it like how we uh we have written this particular data in and how this data is working so tomorrow I'll just try to discuss this particular part called Data diff with evidently we'll see that in the model uh sorry in the

data validation part how to detect the data DFT if there is any data DFT and what is data DFT exactly if you don't know about data D I will tell you tomorrow the entire understand like detail and discussion and all I'll try to see like how to detect the data D with the help of one amazing ml tool C evidently okay so I hope guys uh you enjoyed this particular session and this is all from today uh let me know guys if you have enjoyed the session so give me a like and please like the session guys I can see very less like is there why guys because many people are watching but there is no like

okay there is no like and please try to subscribe to the channel guys please try to support the channel uh just give me some motivation so that I can keep different different unique content for you right so please try to support the Channel Try to subscribe try to um hit the like button and let's try to hit 12,000 subscriber as soon as possible tomorrow Sunday yeah tomorrow would be also CL I mean we'll be doing the live so you can uh get the update from this like live section try to do the project handling missing values and duplicates value and outliers we have already done okay in

our notebook experiment we have already done missing value duplicates value outliers okay you can check it out that particular session lakman even we'll be also doing in the modular coding whenever we will do the transformation part data transformation transformation part there will be doing so in this particular lecture I have already showed you how to handle it we have created the scalan pipeline for this okay you can check it out so all right guys I think uh I'm done with the session so let's connect tomorrow and try to uh see the evidently

experiment and if time allows I'll try to discuss the like uh model uh sorry data validation part as well tomorrow okay so with that guys thank you so much bye everyone take care uh bye everyone thanks bye

Show thinking

This session marked the beginning of building the project's **modular pipeline**, moving from experimental Jupyter Notebooks to production-grade Python code. The main focus was on creating the first and most crucial component of any ML system: **Data Ingestion**.

**📝 Today's Agenda: Building the First Component**

* **Goal:** To build the DataIngestion component from scratch.
* **Function:** This component is responsible for connecting to the **MongoDB database**, fetching the raw dataset, splitting it into training and testing sets, and saving them for the next stages of the pipeline.
* **Postponed Topic:** The presenter decided to cover data drift detection with the MLOps tool **Evidently AI** in the next session to allow for a more detailed focus on the data ingestion architecture today.

**⚙️ The Data Ingestion Workflow**

The presenter used a detailed flowchart to explain the step-by-step process of the data ingestion component.

1. **Connect to Database:** The pipeline initiates a connection to the MongoDB database using a secure connection string.
2. **Export Raw Data:** It fetches the entire "US Visa" dataset collection.
3. **Create Feature Store:** The raw data is converted to a CSV file and saved in a designated "feature store" directory. This preserves the original, untouched dataset for the current pipeline run.
4. **Train-Test Split:** The raw data is then split into training and testing sets (e.g., an 80/20 split).
5. **Save Ingested Data:** The newly created train.csv and test.csv files are saved into an "ingested" directory. These files will be the input for the next component in the pipeline (Data Validation).

**The "Artifacts" Directory**

A key concept introduced was the **artifacts** directory. This is the central location where all outputs from a pipeline run (like datasets, models, and evaluation reports) are stored. Each time the pipeline runs, it creates a new **timestamped sub-folder** within artifacts, ensuring that every experiment is versioned and reproducible.

**💻 Implementing the Modular Code**

The core of the session was writing the Python code for this workflow in a clean, modular, and reusable way.

* **Constants:** A dedicated file was created to hold all constant variables, such as database names, collection names, and file paths. This makes the configuration easy to manage and change without touching the core logic.
* **Environment Variables for Security:** The presenter demonstrated a crucial security practice: the **MongoDB connection string**, which contains credentials, was not hardcoded. Instead, it was stored as an **environment variable** on the system and loaded securely at runtime.
* **Entities (config\_entity & artifact\_entity):** Using Python's dataclass, two types of "entities" were created:
  + **Config Entity:** Defines the configuration for a component, such as all the file paths it will use.
  + **Artifact Entity:** Defines the *output* of a component. For Data Ingestion, the artifact is the file paths to the train.csv and test.csv files. This artifact is then passed as input to the next component.
* **Reusable Connection Classes:** The logic for connecting to MongoDB and for converting the fetched data into a pandas DataFrame was encapsulated into separate, reusable classes. This follows the DRY (Don't Repeat Yourself) principle.
* **The Main Component (data\_ingestion.py):** This file contains the main class that orchestrates the entire data ingestion workflow by using all the other modules (constants, entities, connection classes).
* **The Pipeline (training\_pipeline.py):** This is the master file that defines the sequence of operations. It initializes and runs the DataIngestion component. As more components are built (e.g., Data Validation, Model Training), they will be added to this pipeline file.

Finally, the presenter successfully ran the entire pipeline, showing how the artifacts folder and all the corresponding data files were automatically generated as designed.

**🚀 Next Steps**

The next session will focus on building the **Data Validation** component and will include the demonstration of detecting **data drift** using **Evidently AI**.

e e uh hello everyone uh yeah hi mes uh I think I'm audible guys let me know in the chat if you can hear me yes uh good evening everyone I hope I'm Audible and I'm visible guys to all of you let me

know so we'll wait one more minute guys then we can uh start the session yeah so let's wait one more minute then we can start the session uh with our uh agenda for today So today we're going to discuss like data D detection with the help of evidently the mlops tool and we'll try to complete the data validation component as well okay in this particular session today hi sir please go slow because onwards new topic for us today uh uh see actually uh the thing is like you are learning end to end implementation here so for this you need some basic

prerequisite as I already discussed in my first session so please try to uh complete that particular prerequisite first of all then it would be easy for you to understand okay otherwise things would be much uh complicated for you right again I'm going uh slow because it's a live session I'm taking after that recording is also available in my YouTube channel so if you're not getting something in the live session what you can do you can still what what's that particular recording and you can sort out to your issue okay so that's how actually you can complete you can revise that particular

video again okay this will give you the clear clearcut idea but again uh your prerequisite should be ready otherwise uh some difficulty will face right okay so now I think we can start a session guys so first of all let me write down the agenda like what uh we are going to cover in this particular session so I have opened my Blackboard just a minute I think my pain is not connected just a minute okay now I think it is fine so today is I think day six and today's discussion would be um data DFT demo with

evidently then the next thing we'll be covering the data validation part okay data validation part I'll tell you what uh what are the things you have to do in the data validation okay I'll show you that particular like diagram with the help of particular diagram you can understand what you have to do inside data validation but before starting with the data validation first of all let's complete this particular EV U data de detection with the help of evidently because this particular technique we're going to apply inside our data validation okay so what exactly

this validation validation means you want to validate your data let's say here I having my data so you can see this is my data so if I open up my data so it is having some of the features you can see it is having some of the features let's say it is having uh continent education of employee K ID has job experience or not so these are the features actually so what will happen actually let's say if I want to train this particular data so what I have to do if I want to train this particular data what I have to do I have to first of all take that particular training data then I'll will be then I will be

giving to my model okay but let's say there is a possibility your data may change okay like say after data in your data may change let's say one specific column is missing let's say one column is missing okay one column is missing in your data now if I give this particular data to my model will it work so definitely it is not going to work because one column is missing here okay one column is missing here all let's say you have trained your model you have trained your model with the help of training data okay with the help of training

data now during testing you got a completely new new data set okay that particular data set doesn't have any kinds of columns The Columns actually you can see here so that that time actually your model will fail your model will give you error like this particular entity is not available so that's why first of all before training that particular model I just need to validate the data whether my data is in correct form it or not that means what I have to do I have to check whether these are the columns are available or not inside my

data these are the features are available or not inside my data if it is available okay if it is available then I will return one statement I'll return one statement like validation status is true like my data data is in correct format so there is no issue with my data that's why this particular validation okay validation status it would be is equal to True otherwise what I have to do I have to make it as false and whenever it is false I'll break my code there I'll break a code there and I will raise the exception okay I'll raise exception I'll

tell the user your data is set is not in a correct format so please try to fix your data first of all then try to train the model this is the idea so that's why before training our data okay before training our data first of all we have to validate that particular data whether this particular data set is in the correct format or not I have to validate that particular schema okay so this is one level verification we'll be doing multi- level verification here I'll tell you what are the things we're going to do so uh but before that first of all

let's try to understand the data dift like uh what exactly this data D and why uh we face this particular data D problem inside our let's say machine learning okay first of all let's try to understand this particular data diff if you are completely new to this particular data DFT if you know if you don't know about data dift first of all let's try to uh understand the data dift then I will show you one practical demo okay how we can detect the data D so to explain the data D guys uh here what I can

do I can take actually two distribution let's say this is my training data I think I can take small size yeah so this is my training data okay training training uh data so training data distribution looks like that let's say this is my training data distribution okay this is my training data distribution all right now there is another data I'm having called test data okay now let's also plot the testing data distribution let's say it will look like that now guys please tell me whether this particular distribution is matching or not this training data distribution and testing data distribution is

matching or not yes or no is it matching I think no it's not messing because you can see training data having a completely new distribution and testing data is having uh completely another new distribution that means this particular distribution is not matching okay but what we have learned inside Statistics I think you remember this is a statistical concept so I think you know in statistics actually we are having something called population yes or no in statistics we are having something called population maybe you have learned this particular potion concept now if I take something

from the population if I take something from the population this would be considered as my sample okay let's say I have taken some of the sample from the population it should be considered as my sample okay now if you plot this particular sample if you plot this particular sample if you plot this particular sample let's say this is my distribution of the sample and I also plotted my population again this is the distribution of my population now most of the time you will see this particular distribution and this particular distribution will be

same why because this is a statistical concept they're telling if you're taking something from the population if you're are taking something from the population as a sample so that particular sample distribution will be same as per your population distribution because the same data you are picking from the population okay so that is the idea but here you can see U uh whether I have taken my data from my training example because what I have done let's say first of all I have done uh let's let's say first of all I had one data let's say I have one

main do CSV data main. CSV data what I did I did the T speed so after doing the T speed I got my training data I got my testing data then what I did I just plotted this particular data over the graph and I can see this is the training distribution and this is the testing distribution and I can see that this particular distribution and this particular distribution is not same it is different that time this is a problem of data D that means this particular training distribution and testing distribution is not matching whether you

have taken your data from the let's say population okay you have taken your data from the population but still your sample data is not matching with the population distribution okay so this is the problem of data drift data drift okay and this is like one of the very uh I mean um what I can say a very actually um I mean uh I mean very um very much complex problem inside machine learning so if you get data D problem so your accuracy would be like very low out there okay yeah uh how let's say I I will give you

another example yeah I'll give you another example actually let's say you working with computer vision okay let's say you are working with computer vision so there you have trained your model with the help of dog data set say this is my model what what you have done you have trained a data uh model with the help of dog data set and if I plot this particular dog dog data let's say this is the distribution of the dog is the distribution of the dog okay now whenever you are testing the model okay whenever you are testing the model that time you are giving the cat images to this particular model and

let's say C cat distribution look like that okay this is the cat distribution now tell me whenever your model got trained did your model saw these kinds of distribution or this kinds of distribution tell me this is my first distribution this is my second distribution your model saw this particular first distribution this kinds of second distribution it didn't see right so that's why whenever you are giving this particular cat image during testing you'll see your model will performing wrong okay there would be a

false output always okay there would be a false output always from the model right so this is actually again a data drift problem data drift problem okay data Ty problem because again your sample is not matching with your population distribution okay so this is a data diff problem I hope you got your answer right what is exactly this data D so what I have to achieve guys always I have to achieve whenever I'm having some population let's say when I'm I'm having some population let's say I'm having main do CSV data this my complete data

after doing the TR to spete what I will get I will get my training ex training set as well as my testing set okay and whenever plot this particular training set so this distribution and whenever I'll plot my testing set this distribution will be same okay it should same always okay that time I can consider it's not a it's not a data D problem okay data drift problem so we have to we have to achieve this particular output always okay we have to achieve this particular output if there is a data D problem again you have to

fix your data you have to tell that particular user okay or let's say you have to tell that particular developer uh there are some issue with your data with your sample data so please try to fix your data first then try to train that particular model okay this is the idea that's why actually U there are so many mlops tool over the market because see um if you don't have the mlfs tool what you have to do you have to manually write this particular code to detect the data D okay what you have to do you have

to manually use let's say cbond m plot Le you have to plot that distribution you have to get that particular value then again you have to write some manual logic then you can detect that particular data d whether my training distribution and testing distribution are same or not if this is a same there is no data D problem otherwise there is a data diff problem okay based on that you will make a decision whether your data validation Stace would be true or false if there is a data data D again the data validation status should be false okay uh and if there is not a data

d this particular status should be true this is the main idea what will happen if we create the model with data div data I told you now so let's say this is the example I show showed you right let's say you are training one model with the help of dog images but whenever you are giving the cat images that time actually will get a false output okay it won't be giving you the correct output it would be more robust on the dog side not a cat site actually okay I think you got your answer how to solve data what is the solution so for this again you have to

resample your data okay again you have to resample your data and you have to make sure that particular data is also matching with your population distribution otherwise again you have to collect the data again you have to increase the data site okay inside your database that is the idea so it will happen most of the time whenever you are having less amount of data okay whenever you are having less amount of data that time if you do 10 to split operation you will see that in the testing data you will have some of the data which is like very new this is not available inside your population

which is not available inside your train model okay that is the issue actually we'll be facing but whenever you are taking larger amount of data this particular problem you won't be facing okay so I'll show you one practical demo it would be much Clear like how we can visualize this particular uh training distribution as well as the testing distribution and how we can detect this particular data di okay for all the columns we are having here in this particular data set okay I hope this part is clear guys uh as a theoretical

high level understanding I hope you got like what is data D exactly let me know guys if it is clear so so give me a like thumbs up give me a any kinds of emoji in the chat so that I can get to know this part is clear to all of you let me know guys if it is clear to all of you the data di what exactly this data d so train and testing data should follow the same distribution yes aru you're right because this is the statistical concept okay in statistics I think uh if you learn the statistics I think you know whenever I pick something from the

population as a sample that particular sample distribution would be same as per your population distribution that is the idea just a minute all right so now uh how we can detect this particular data DFT with the help of mlop tool with the help of python now let's try to C1 demo so what I can do I can go back to my code editor and I think you remember we created one notebook folder inside that I'm going to create another notebook so let's create this particular notebook I'm going to name it as data okay dataor

driftor demo okay uh with evidently evident evidently do IP uh ynb file okay this is the BNB file that means Jupiter notebook file now let me create this particular file and inside that I'm going to select select my kernel the kernel I have created called Visa kernel okay now if I want to test it so I just write a print statement in this particular kernel it should give me the output that means this this is working fine okay this is working fine now see uh I think you remember we already install this particular evidently inside our environment in the requirement. txt I mention this

particular evidently you can see so evidently package we have already installed so the version was 0.2 uh 0.2.0 so this was theic version actually I was using now you can check the official website of the evidently so if I just search evidently AI so this is the official website uh you can see this is the official website and there is another thing actually will get called official GitHub so evidently evidently GitHub GitHub okay so it is also available and this is like very active resarch you can see the last comment was two days ago this is the official

evidently GitHub you can see okay they have already given the distribution and all they have already given the demo okay you can see they have already given the example demo how your data distribution will look like and everything they have given so here you can create a completely dashboard so this particular dashboard you can create interactive dashboard with the help of this particular dashboard you can understand okay about your data D so not only data diff with the help of that you can do lots of thing you can monitor your dashboard then you can um see I think they have written here different

different um you can generate the reports um then what you can do in the website I think they have mentioned goab uh evidently evaluate test and monitor so this particular three task you can perform evaluate test and monitor and the data D I think okay I think data D is somewhere they have mentioned uh see it is also integrated like ji so then jna also you can perform with the help of evidently AI so with the help of this particular evidently you can also evaluate your rag based application okay or your llm based application you can also monitor that particular LM based

application so going forward whenever we'll be discussing about the generative AI will be creating generative a project I'll be also showing you how we can introd integrate this particular evidently a for with the generative project as well okay but as of now we'll be doing the data diff one only okay data D we'll be doing so yeah so I was exploring this particular tool then I just created a jupyter notebook file guys for you so let me just uh show you the Cod snippit like how we are going to detect this particular data D

so I'm going to open up this particular file and here first of all what I have to do first of all I'll be importing some of the libraries here okay so let's import some of the libraries so here I can write uh import pandas aspd that means first of all I have to load my data okay I have to load my data but as a demo actually I'm not going to use my Visa data I'll use one uh very um common data set which is nothing but boost on house price prediction data because everyone familiar with this particular data so with the help of this particular data I will show you the demo then the same

demo I'll be applying inside my V US Visa data okay whenever we'll be doing the data dation they actually will'll be applying as a modular coding okay so first of all let's try to see the experiment how it can be done then I need another actually um function from evidently so first of all let me from evidently uh so inside that there is a method it will get called dashboard and here I need to import one class called dashboard okay this is the class I need to import then I also need something called uh do evidently okay evidently do

tabs uh input I need something called Data D tab okay data D tab so if you have the like data like numerical data but you can do you can use this particular method called Data diff tab with the help of that you can uh detect that particular data D okay in a numerical variable and for categorical variable you can use this particular method called Data data Target D tab okay data Target D tab this particular method you can use now I need to import some more additional package here so let me just write evidently do model OKAY model uh

profile uh input profile so with the help of this particular profile I'm going to create create the dashboard the uh HTML dashboard it will give you one HTML file that particular HTML file you have to render in your browser then you will be able to see that particular dashboard okay then I also need something called evidently uh do profile uh SE profile sections import data D profile section okay data D profile section this particular method ID now let me import all of them so it should execute properly guys you can see there is no error that means it working

fine so first of all let's load our Bon data so here I can write load our Boston data okay so here I can write Boston data so how to load the Boston data guys so for this what you can do you can Define one data URL so data \_ URL and uh there is a URL of the Boston data guys let me show you this particular URL if you don't know so there is a GitHub URL is available for the Boston data okay you can see that particular URL so this is the Boston Housing data and this is a CSV file this is a CSV file and it is hosted in the GitHub so I

can access this particular data with the help of pandas so I'll copy this particular link and I will go back to my code editor and here I'm just going to write uh data URL is equal to this particular URL okay then I'm going to use the fundas to read it so I'll just write pd. read uh pd.

read csb inside that I need to pass my data URL and this is going to give me my data object so I can store as a Boston okay boncore frame that means this is the data frame is equal to this my data frame okay now let me import this particular data now if you want to visualize this particular data simply you can do the head operation on top of this particular data so I can write uh Boston frame.

head and you can see this particular head of the data that means all of the column is it is visible okay whatever data actually you can see that particular server it is already available inside my local machine right now okay now I can also print the description of the data for this again you can write bonf frame.

info so it will give you the complete description of your data like how many columns you are having how how much size actually it took in the memory and what is the data type you can see these are the floating data type inder data type okay and there is any n value one not okay everything you will able to see and all the column names actually it will list down here all right so this is the actually simple statistics actually you will get if you apply this particular info method on top of your data frame okay now we'll be just doing the data D okay data D dashboard that means we'll plot the data D dashboard so let me do

it so first of all let's check the shape of my data so vone do shape here I can do so here I can see I'm having 506 rows I'm having 56 rows and I'm having 14 columns here you can see 56 rows and 14 columns all right now now first of all what I have to do I have to se excuse me first of all I have to separate out my training example and I have to separate out my testing example so what I told you you can see here this is my entire data that means this is my population now what I have to do I have to do the 10 test spit okay then after

doing the T speit what I can do I can use my uh sorry after doing the tent speit what I will get I will get my training example as well as my testing example so the same thing I have to perform that means this is my complete population boncore Fame this is my complete population it is having 56 records in this particular population now I'll do the 10 speed so how I can do it I can manually do the 10est spit I can just simply write Boston Fame and I want to take first 200 example for the training first 200 example for the training now if I execute see it will give me 200 example for the training

okay now if I want to take the testing so what I can do I can take last uh 200 that means last that means after 200 whatever date actually I'm having I can store as a testing data so here I can write uh 2 12 uh 200 like that okay now you can see uh after uh 200 till 306 actually rows I got the entire data that means this is going to be my training data here I can write this my going to be my training data okay this is my training data uh that means this is my reference data reference data reference data

okay and this is my test data test data okay that means my uh that means my actually sample data okay sample data actually I got it now that's how actually we can perform this particular T speed now what I have to do I have to create this particular dashboard so to create a dashboard I'll import this particular dashboard method I think you remember I imported here so dashboard I'm importing now inside that I'm going to pass my data okay I have to pass my data but before that I have to initialize one tabs here so these are the syntax from the evidently a okay this is not my syntax they're telling if

I want to if you want to detect the data diff you have to write the syntax in that way okay now inside that you have to pass this particular parameter called Data drift tab I think you remember we imported previously this particular function okay that's why we passing it here now it will give you one object so the object name should be boncore data \_ DFT okay netore Dore dashboard dashboard okay this is going to be the variable name I think it is fine underscore yeah everything is fine now let me execute let me execute see it's working fine now what I have to do I have to

calculate this particular uh data D that means I have to detect that particular data div okay for this what I can do I can call this particular object and there is a method called calculate okay inside that you have to pass your data like that so first of all I have to pass my training sample so this is the training sample and here I'm going to pass my training sample okay then I'm going to pass my testing sample so this is going to be my testing sample okay testing sample so once it is done now let me execute this particular code see my object has created now if

you want to show this particular dashboard so what you can do you can just write this particular codes in a pit um boost on data. D and there is a method called show Okay show you can use now it took it will take some time guys to render inside your vs code but if you don't want to wait what you can do you can also save it as a HTML file so for this what you can do you can again call this particular object and there is a method called save okay inside that you can mention the file so let's say this

is the file uh I can name this particular file Boston okay Boston uh underscore dataor drift okay data drift \_ report okay HTML so you have to save as a HTML not any other file format okay now if I execute you will see that inside my notebook folder one HTML file is created inside my notebook folder one HTML file is created now I'm going to open up this particular file I'll go back to the notebook and if I double click on this particular uh file okay you'll see it will open up One dashboard for me okay see it will open up One

dashboard for me so this is the complete dashboard I think guys this particular dashboard is visible to all of you let me know this is a complete amazing dashboard guys I have created okay with the help of evidently a uh thanks V for the Nvidia video thank you page apart from evidently which mlops tool we going to use in this particular playlist uh see we are we are starting with the like uh a simple project uh implementation so here only we'll be using uh this particular evidently and uh deployment wise I'll be using GitHub action and Docker and all

going forward whatever project will be we'll be using some more advanced tool like uh we can use let's say um we can use let's say genkins okay we we can use Circle C we can we'll be also exploring the kubernetes okay we'll be also exploring Bento ml Amazon sales maker okay different different things we'll be exploring pankas okay sorry I think is asking this particular question uh Na n what is the model training sir model training as of now we have been done model training we'll be doing after doing the data transformation stage okay

it just data validation St we are doing right now all right so now you can see this is the dashboard guys I got and in this particular dashboard I'm having all the features okay all the features you can see here all the features you can see in this particular data okay all the features are available okay and you can see the type of this particular features whether it's a number whether it's a category everything would be visible here all right now you can see that reference distribution and you can see the current distribution now what is

reference distribution guys I already told you reference distribution is nothing but your training sample okay your training sample that means the entire population okay that means the entire population okay and what is uh this particular Uh current distribution current distribution is nothing but your testing sample okay that means your test data now you can see both distribution it has plotted now you can see the reference distribution as well as the current distribution now you can see uh it is almost same distribution okay it is almost same distribution okay there is no chance small minor chance but

minor chance actually we don't detect okay you can see if this particular data D score is close to one that means there is no data diff detection if it is close to zero that means there is a data diff detection you can see that there is no detected okay there is no detected that means this particular distribution and this particular distrib are almost matching that's why it's not detected there is no data D in this particular column okay in the edge column and you can see the score of that particular data D it's almost 64% okay 64% actually

confident this particular distribution are same data validation means output of the test data no data validation means we are validating the data whether my data is in the correct format or not let's say whether my data is having all the features all the let's say uh numbers or not or it is having any kinds of data D or not if none what I have to do I have to make my status as true that means my data set is fine I can do the training otherwise I have to make as a false see in machine learning here you are not using the complete CSV data just try to

consider aru you are not using the complete US Visa CS data okay we don't take the complete csb data and we don't train the model what we do we do the 10 speit operation the experiment I showed you so there actually I was taking the data from my population only here okay the population only but here what I have done I have taken two kinds of data training and testing that's why this training would be considered as your population right now okay I hope this part is clear right so you have to consider training is your population that means your reference data and testing your

current data okay I won't say it's a population it's a reference data okay you have to say reference data inside evidently all right that's why it is telling you can see uh if I open up this particular dashboard it's a reference distribution it's a current distribution all right yes now for the next column you can see almost this particular distribution are same again not detected and this is the score and for the next column you can see again it is almost same there is no dete uh not detected okay again this is the uh isod okay now

you can also expand this particular things and you can also try to understand with different different graphs so you can also visualize this particular data distribution and you can see your uh current distribution as well as the reference distribution okay it's almost same okay almost same you can see the percent everything you can see and you can over it you can H your mouse you can see the interactive dashboard guys okay interactive dashboard okay that's how you can see for all the columns okay all the columns it is visible okay you can see that so this is one of the amazing tool they have developed uh and

if we don't have this particular tool what I have to do I have to write this particular code manually right so that's why I told you MLF is nothing but it's a tool actually at then it's a tool it will make your task automated okay most of the time yeah there is no no see there is detected also you can see there are so many data it is detected as a data D okay you can see now first three column there is no detected but after four column you can see there is a detection so you can see this particular distribution this particular distributions are different that's why

it's detected that means there is a data D and you can see this particular value is close to zero okay very go I think you understand right now again it's a detected this particular distribution are not matching again detected detected detected detected detected okay now you can go to the next page and here you can also see there is also detection okay that means only three column there is no detection rest of the columns are data D you can see data D detected for 78% of of the features like 11 out of 14 data set uh D is detected okay so it will

give you the entire like analytics okay it will give you the entire analytics and you can see that particular analytics as a visual understanding but I don't want to plot Des are the value I mean graph whenever I'll do the pipeline that means the modular coding I have to consider this particular number I have to consider this particular number and this particular data di uh detection uh status okay with the help of that I can consider whether this particular data set is having data D or not okay I'll tell you how to do it write this particular codee CIT as well okay so I

hope guys this part is clear all of you let me know so we'll follow the same technique to detect the data D inside our US Visa project as well this is the idea all right so now let's start with our uh next agenda which is nothing but data validation okay data validation because data D demo I already showed you that's how we have to detect the data diff now we can pick up any kinds of data you can use the evidently and you can plot this particular distribution and you can check the data d uh like detection okay now uh to uh explain this

particular uh uh this particular actually data validation first of all I'm going to show you one flowchart so let me open up this one flowchart so this is the flowchart is available inside flowchart folder even I will also share with you I'll will comit this particular changes in my gab you will get the changes here now guys you can see this is the flowchart I think I can make a bigger screen see this is the flowchart now in this particular flowchart again you can see um I have taken some of the constant okay I have taken some of the constant

and configuration so my data uh validation will run first of all what it will do guys it will take the train and test from where from the data inition artifact because I think you remember so yesterday I was running my data in and I was returning this particular Trend and test data as my artifact for the next entity right artifact from the next entity okay so that's why you can see data in this it will return train and test this particular train and test I have to receive inside my data validation I have to read the data then

I have to first of all validate the column I have to validate The Columns whether I'm having all the columns or not inside my data I'll first of all validate whether all the columns are having inside my data or not okay I'll just try to match K ID continent education employee has job experience I'll match all the column all the columns are available or not if it is available okay if it is available my status would be okay status would be you can see here uh train status test status status should be false columns are missing data frame uh I think here I

have written uh yeah you can see here so if it is a false so what will happen if this status is false it will give you one validation error if this particular stat status is FAL it will give you the validation error otherwise if it is true okay if it is true you can see if it is true that means one line I'm bringing here because it's a flowart you can design in any way so that's how actually I have designed if it is true then your data set would be detect uh data D would be detected that means first of all I

have to validate The Columns whether my number of columns are matching or not then I have to check if numerical column exist or not I think you remember in the data set we are having some numerical columns you can see we are having some numerical columns let's say this is the numerical columns this is a numerical columns this is a numerical columns these are the numerical column exist or not inside my data set it is if it is not EX again it would be the validation other otherwise it would be uh it will go inside my data data de detection phase okay then I have to check another

actually uh like layer which is nothing but is categorical column exist or not again you can see inside this particular data I'm also having some categorical data okay I'll also validate this particular categorical data exist or not if it is not exist validation eror otherwise it will go to the data diff detection part now if my three stage run successfully if this particular three stage run successfully that means validate number of columns is numerical column exist and is categorical column exist if all the status is true if all the status is true then it will go to my

data D detection phase and then I will start my data D detection that means the evidently uh like code I will execute that particular evently function I will be execute on top of the US Visa data and this particular data D will give you one report it will give you one report whether it is having data D or not if it is having data D if yes okay if it is having data if yes so it will again do the validation error otherwise what it will show you if it is a false it will give you the output it will give you

this particular report okay it will give you this particular report you can see I will get one report. EML file in that particular EML file I'll see that particular report okay whether it is a true or false okay then as an artifact actually it will return these are the path validation status validation uh train file path test file path okay so it will give you this particular uh actually startus that means this particular uh output for the next component which is nothing but my data transformation for for data transformation what I need I need

validation status whether this status is true or not let's say how things are working let's say this is my data validation component validation data validation data validation so as an artifact it will give you output validation status validation status okay start St if this particular status is true what I have to do I have to start my next component which is my data transformation data transformation data transformation I have rest started okay if it is false what I have to do I have to raise the I have to raise the error I have to raise the error or I

have to raise the exception okay exception to the user your data is not in a correct format this is the idea okay this is what actually we usually do inside our data validation okay data validation I hope this part is clear guys all of you right now let's try to implement this particular fler as my code so I'll go back to my code editor okay so here what I can do guys I can just uh uh yeah close this particular uh notebooks it's not required yeah so the same way I'll be updating my file guys I think yesterday I told you first of all I have to update my

constant then um entity then component then pipeline then my main file okay main file okay main file that means my end point because see if there is a data D so do you really think you have to perform the transformation because there is a issue with your data so you can't use this particular data now so that's why you have to check the data dep because after transformation your model tral will start now that's why it's just a p line it's a execution of flow just try to remember data injection validation transformation trainer

evaluation Pusher so if one of the St is failing okay it will give you the exception it won't be running the next uh I mean uh pipeline okay that time that's is the idea are sure right yeah so first of all I have to update my constant so let's open up this particular constant folder and there is a constant uh file I'm having so inside that I'm going to mention my data validation related constant okay so let's mention so what I can do um yeah I can mention some additional file as well so let me just mention

quickly sorry I just need to mention additional um additional variable okay let's say Target column is my case status I think you saw inside particular notebook Cas status is my target column current here I need because I need to calculate the age of the company okay that's why I'm taking this particular current here pre-processing object file name that means whenever you have to save this your uh like pre-processing object that means that particular pipeline object you have to save as a pre-processing DOL okay these are the very common step inside machine learning I think you know that okay that's why we

are setting the at the variable inside the constant now anything else I need to set let me just check we have set the mongodb URL TR file name test file name it is fine us visit is fine model. P we have also added and uh we have also added the US visa. CSV is available US Visa CSV it's also aable now we also need to add the schema file path URL okay so this is going to be my schema file paath URL so inside actually uh config we are having a file I think you see schema. EML inside that I have to mention the schema that means my data schema so let me show you how your

schema will look like so I already prepared this particular schema you can see this is the schema that means number of columns we are having in this particular data set you can see Cas ID continent education has experience required job training number of employees that means whatever things you can see in the data everything I have mentioned here everything I've mentioned here okay in this particular schema then I mentioned the numerical column like how many numerical column we are having so you can see we are having uh 1 2 3 three

numerical column so 1 2 3 then I have mentioned my categorical column how many categorical column we are having 1 2 3 4 5 6 7 8 9 10 you can see that 1 2 3 4 5 6 7 8 9 10 okay these are my categorical column and what are the columns I have to drop I have to drop the case ID and year of estabilish I think whenever I did the exploited data analysis I think you saw that this particular two columns is not re I can remove it that's why I written inside drop column section okay drop column section I have written now here I have written number of features

okay for the data transformation like what are the uh like transformation we have to apply on top of which are the column here I have mentioned now ordinal encoding column you can see these are the column I have to the ordinal encoding then one encoding these are the column I have to apply the one encoding and transform column the scaling column that means the power transformation I have to do on top of this particular uh column so everything I have done inside my Eda uh whatever things I have done

inside my Eda and inside my feature engineering so here I'm mentioning all of the column and this is called actually schema okay this is called actually schema okay I hope you got it fine this is called actually schema so that's how actually you have to prepare the entire schema of your data set okay you have to prepare your uh schema okay inside this particular file whether it can be any kinds of data first of try to see The Columns all the columns try to mention try to separate out the numerical column categorical column that column you have to drop number of uh uh numerical features okay

numerical features like you want to do the transformation uh whatever let's say you are performing in the transformation let's one encoding ordinal encoding um Power transformation you can mention in a uh section okay in a mention in a key value section so this is the ml file synex actually you can write this is a key this is the value and then uh file is a dictionary okay if you just read this particularl file it will give you dictionary kinds of output all right this is the idea so that's how actually we have to prepare the

schema fine so I think anything is required let me check schema I have ADD it uh okay I think there is nothing now I can mention my data validation related constant here is my data validation related constant guys okay you can see so I think you remember we need uh my data path so here I'm passing my data validation uh folder data diff report and report.

EML that means whatever things it will generate everything I'm keeping it here data d report uh then uh all the things I don't need you can delete it I only just need this particular three things like data validation that means in the artifact folder it will create a time stamp inside that it will create a data validation folder inside that it will generate this particular report.

EML okay inside this particular data report folder that is the idea whenever it execute the code it would be more clear as of now let's try to save this particular code fine so now what I can do uh I have updated my constant now I have to update my entity okay so let's update the entity so I will open up this particular entity folder right now so here is the config entity as well as my artifact entity both I have to update so again it's a it's just a like return return of path with the help of o.p. join you can return your path so let me just quickly show you first of all I have to update

my config entity so here is my config entity so again I'm returning the path with the help of operating system package so here I can write okay so this is the artifact entity have written sorry guys I have to save it now I can copy this particular entire code and inside uh config entity I have to mention okay again you can see I'm returning the path only my data validation directory path and data diff report file path okay now similar wise I have to update my artifact entity so again um you can see this is my artifact that means whatever thinks it will return as

an artifact okay after doing the data validation so it will return return validation status it should be the Bull type whether it's a true or false then masses okay whether it is having the data diff or not and the data diff report file path okay like that particular EML file path it will generate inside my artif this particular path would be given for the next component next component will load this particular path and we'll try to understand whether it is having some data problem or not that means the

validation Stratus is true or not if it is true it will uh run your data transformation otherwise it will raise the exception that is the idea now let me save I'll also save this particular file now I have to update the component so now let's go back to the component data validation and inside that let me update the particular component so data validation so again uh see it's a simple logic I have written first of all I'm importing all of the necessary Library I need then I defined a class called Data validation you can

see data validation it will take the data uh ination artifact as well as the data validation configuration inside that I'm writing the class variable okay then I've written one particular method you can see this particular method I have written validate number of columns validate number of columns it will validate all the columns so what for this what I'm am doing guys I'm first of all loading this particular data frame I'm reading all the columns and I'm also loading this particular I'm also loading this particular uh schema. ml file and

from here also I can checking all the number of columns are exist or not okay that is what actually it is happening inside that now once it is done I'm checking is column exist or not all the column exist or not you can see that with the help of this particular function uh I think I'm calling it somewhere just a minute yeah I'm calling it and this is the next function is column exist or not like all the column exist or not in that particular report file you can see in that particular schema. EML file I'm checking whether all the columns are exist or not again it's a simple python

logic guys here I'm just writing a f Loop I'm taking all the numerical column okay you can see I'm taking all the numerical column all the categorical column okay and I'm checking one by one all the columns are exist or not if exist what I have doing guys I'm just sending the status okay I'm just sending the status it's true otherwise it's a false then again you can see after writing this particular code CIP that means this particular things is done also this particular things is done see these two things actually I'm doing in

one function only in one function I'm doing is column exist or not here I'm checking my numerical column and categorical column both numerical column and categorical column both okay don't be confused guys if you see this particular function carefully first of all I'm checking the numerical column then I'm checking the categorical column okay then after that I'm starting with my data D detection and if you open that particular data notebook the same code I have written here okay I just replicated here as in modular coding you can see profile selection calculate then I'm saving as a

Json okay this particular things I think I haven't showed you how to save as a Json format so let me show you so let's say instead of saving inside a HTML I can also save as the Jon format it is also possible okay how to do it let's let me show you so for this what you can do H you can just simply write uh this particular code snip it first of all I'll uh initiate this particular profile okay profile is not defined so okay I think I have to execute this particular code again then uh I have to calculate then I have to save as a just on so here this is the

code now you can see this is the juston output I'm getting like all this code everything you can see here okay everything you can see here detected non detected is cod everything I'm getting as a just on format because here I'm writing the modular coding so I don't need to plot this particular graph so by seeing this particular number by seeing this particular activity I can do it okay whether there's a a data D or not you can see data D detected true data D detected true okay so with the help of this particular value I'll try to detect okay whether it is having some data D problem or not so that's why we have

written this particular function you can see here uh this particular function so you can see I'm taking the metric and data features checking the data detection if uh you can see the status actually I'm sending then inside initiate data validation I'm calling all the function y by for first of all validate column then is column exist or not then I'm doing the data D detection and finally the output I'm getting you can see data not detected detected so this is the error message actually I'm printing whether it's a true or false okay whether it's a true or false if it is true I'm sending as a true otherwise

I'm sending as a false because I need this particular statement for the next component okay just simply try to understand see what is happening here this is your data validation after validating the data it will first of all validate The Columns then it will do the data D detection if this particular statement is true that means there is a issue with your data if your column is not matching if there is a data D it will give you a uh like uh false okay false masses false means your data uh there is some issue with your data there are some issue with your data validation

that time you have to raise the exception and if it is true if it is true that means there is no data validation sorry there is no problem with your data okay okay that is what actually I'm doing in this particular codeit okay if you see it carefully I'm doing the same thing now let's test this particular code whether it's working or not I can save and I can now update my pipeline so to update the pipeline I will open this particular pipeline folder my training Pipeline and here I have to first of all import I have to import uh

here uh packages okay so let me import quickly first of all I have to import data validation then I need to import data validation config data validation config then I also need to import something called Data validation artifact okay then I also need to initialize my data validation config inside my training pipeline then I have to start the data validation so for this I'm going to write a uh function start data validation and again I'm calling this particular initiate data validation from my data validation

itself okay if I execute this particular method everything would be executed okay that is what actually I'm doing the same thing we did for the datation as well then now I have to call this particular method inside my run pipeline so here I'll be calling this particular method inside my run pipeline after data in your data validation will execute okay this is the flow of execution I'm handling here now I think everything is fine now I can open up my demo. pi and run pipeline is calling here now let me execute the code so I'll open up my

terminal I'll clear it now see guys this is my uh artifact as of now now if I execute this particular code again so python U demo. Pi now again it will inest the data uh okay so it should throw one error because I haven't set my environment variable let me check whether it is available or not I think yester you remember we set our environment variable so it should be there mongodb URL I think it is there let let's see whether it is able to load or not otherwise we'll set manually okay I think uh execution successfully that means it is able to take that particular credential from environment

and this is the execution guys and you can see the previous execution only data in was there now the last execution you can see data inition as well as the validation is also available data inition wise it has stored all of my data okay from my database and data validation I got one report you can see I got one report and this is the complete report guys I got you can see this is the complete data diff report it's like very big so it's not understandable like that okay so from here I only need to take the true false statement okay now if you want to take understand whether it is having data di

or not what you can do you can open up this particular log file so there is a log file this is my log file and if I go last so there is a data d uh data di see message is data diff not detected okay data diff not detected and validation status is true that means there is no issue with my data uh sorry I think I told you wrong if it is true that means there is no issue if it is false that that means there is a issue okay and you can see this particular message di not detected D not detected okay that means the data we are using this particular US Visa data there is no

issue with this particular data D problem okay I hope you got it that's how we can do the data D detection guys see only one particular column has detected as my data d uh which is nothing but this particular column I think I think you remember which column I'm talking about guys uh that particular case ID right I think you remember the case ID column which is having unique value right so this particular column has detected as my data D but again it's a majority calculation if one and two or three column is detecting as a data D it would it w't be considered if your majority columns are having actually not data D

it would be considered as a true okay otherwise it would be considered as a false I hope this part is clear guys fine so this is what actually we can do the data validation part now let me give you the source code so I can just push this particular code so here I can write data validation added will you use kubernetes in this project no in other project we'll be using kubernetes suras now please try to execute from your end guys uh once and you can let me know whether it is executing fine or not from your s uh and this part would be

much Clear okay that time if you're executing from your end see code is available everything I've shared only you just need to execute from your end okay yes so next class guys uh we'll be starting with the transformation data transformation in the transformation we'll be doing all the feature engineering task we have done in the notebook okay then after that we'll train the model we'll evaluate the model and we'll push the model to the3 buet then once my model Pusher and everything is completed I'll create the web application of my entire application okay then we'll try to deployment this

particular project over the cloud platform this is the entire idea so in case after production how we'll calculate the data D for the new data they will consider the previous training data as a reference friend uh no see in production Also let's say whenever you are hosting this particular application right so again if your new data is coming that means whenever you you are having this particular data uh whenever you are having this particular model in the let's say production server there also you need to train the model right and whenever you are training the model that means you are executing data inition

validation transformation everything that time you can see whether it is having data D or not if data D out there you have to fix it otherwise you can use that particular model okay and in future if your data source is changing again you have to return that particular model okay you have to return this particular entire project again okay that is what actually you have to do which test is used to calculate the data diff is score please explain those columns in evidently so there are some statistical column is used you can see

start test K KS and P value I am familiar with P value but I'm not familiar with the ks parameter maybe those who knows statistics better than me they can tell okay what is K Val KS parameter but I think you know the P value right P value statistics so this is actually P value calculation actually at the end sir uh if D detected uh resampling is manual or what is the process first of all do the do the resampling if not working try to increase the data provin okay that's the idea and try to take the data from the same s same population always not any other not from any other population I think I think I told you

now there's a dog cat example I have given let's say we are using dog images for the training no need to take any cat images for the testing so whenever training the model try to take the dog images as well as the cat images both images you have to take okay then you will be able to do the testing operation on top of it that's the idea I hope this part is here okay guys so now uh if you uh like this particular session guys so please try to like this particular video and try to subscribe to my channel as if you haven't subscribed yet uh because so

many people are watching but they are not getting subscribed to my channel okay let's hit uh 12 uh 12,000 subscriber guys as soon as possible okay so that I can bring uh lots of content for you lots of New U actually content for you in future right all right so yes guys this is all for uh today's actually session tomorrow U um actually I'll be doing that transformation part and the model training part and uh we'll try to see like how to do the transformation in a modular way and how to do the model training and all everything we'll try to see and once we complete this particular

project we'll do the deployment okay of this particular project and please try to see this the complete uh playlist so here actually I have added all the video and resources is also available let me share the resources again in the chat yeah so yes guys this is all from today's session uh so I think you like this particular session so thank you so much guys for joining this particular session bye everyone take care bye thank you

Show thinking

This session was dedicated to two critical topics in MLOps: understanding and detecting **Data Drift** and building the **Data Validation** component for the machine learning pipeline.

**📝 Today's Agenda: Data Drift and Validation**

* **Data Drift Demonstration:** A detailed explanation of what data drift is and a practical demo on how to detect it using the MLOps tool **Evidently AI**.
* **Data Validation Component:** Building the next stage of the modular pipeline, which acts as a quality gate to ensure the data is valid before model training.

**❓ What is Data Drift?**

Data drift is a common reason why machine learning models fail silently in production. It occurs when the statistical properties of the live data being fed to the model differ significantly from the data it was trained on.

* **The Core Problem:** A model is trained to recognize specific patterns. If the patterns in the new data change, the model's learned knowledge becomes outdated and its predictions will be inaccurate.
* **An Analogy:** Imagine you train a model to identify different types of dogs. It learns the features of dogs very well. If you then show it a picture of a cat, the model will likely fail because the distribution of "cat features" is completely different from the "dog features" it was trained on. This mismatch is data drift.

**🔍 Detecting Data Drift with Evidently AI**

The presenter demonstrated how to use **Evidently AI**, an open-source MLOps tool, to automatically detect data drift.

* **Process:** A Jupyter Notebook demo was conducted using the Boston Housing dataset.
  1. The data was split into a "reference" set (simulating the training data) and a "current" set (simulating new production data).
  2. Evidently AI was used to compare the two datasets.
  3. The tool generated an **interactive HTML report** that visualized the distribution of each feature in both datasets.
* **The Report:** The dashboard clearly highlights which features have drifted, provides a drift score, and shows the statistical tests (like the Kolmogorov-Smirnov test) used to make the determination. This automates a complex manual analysis process.

**✅ Building the Data Validation Component**

The concepts of data drift detection were then integrated into the project's modular **Data Validation** component. This component acts as a crucial quality check after data ingestion and before data transformation.

**Multi-Level Validation Checks**

The component was designed to perform several checks in sequence:

1. **Schema Validation:** It first checks if the incoming data (from the ingestion step) matches a predefined structure. A **schema.yaml** file was created to define all expected columns, their types (numerical/categorical), and columns to be dropped.
2. **Data Drift Detection:** If the schema is valid, the component programmatically runs the Evidently AI drift check on the training and testing datasets.
3. **Status Report:** Based on the results, the component generates an artifact that includes a validation\_status (either True or False).

**How it Works in the Pipeline**

* The DataValidation component receives the train and test file paths from the DataIngestion artifact.
* It performs all its checks.
* If any check fails (e.g., a column is missing or significant drift is detected), the validation\_status is set to False, and the pipeline will stop, raising an exception.
* If all checks pass, the status is True, and the pipeline proceeds to the next stage (Data Transformation).

The presenter successfully ran the updated pipeline, showing that the US Visa dataset passed all validation checks and that the drift report was correctly generated in the artifacts directory.

**🚀 Next Steps**

The next session will focus on building the **Data Transformation** and **Model Training** components of the pipeline.

e all uh yeah hello everyone good evening I hope I'm audible to all of you guys let me know uh good evening uh diar

G fine I think I'm audible so we can start with the session but I will wait one more minute so that everyone can join yeah and I hope guys uh your project is running fine uh there is no issue from your end you can let me know yeah hello P hello gab good evening guys uh sir daily upload one gni video please multimodel rag please playlist uh okay I'll try to do uh actually apart from uh YouTube I'm having small other work as

well so whenever I am available definitely I will um bring lots of content for you guys okay definitely there are so many things I'm planning to bring in this particular YouTube channel okay which is completely new and unique so definitely will be enjoying all right so now we can start the session guys and uh um I think uh you know like uh which playlist actually we're running currently so I created one playlist called mlos production ready machine learning project and inside that I am just taking uh lots of session regarding this particular project

implementation and we have completed till day six like we have completed till data Rift with evidently and data validation part so in this particular session uh we'll be learning about the data transformation and the model training part uh then uh I think tomorrow or after tomorrow we'll just try to see that evaluation part and the model Pusher to the AWS cloud like how we can push our model to the AWS cloud then we'll finally uh like create our front end application of this particular application then we'll try to do the deployment on the cloud platform with

that we'll end this particular project then after this I'll bring some more uh like end to end implementation with some more complex mlops tool um so that actually will be learning each and everything okay from this particular playlist only uh you are my uh you are my data science Guru I learned many implementation from you thank you gorov uh thanks for the good appreciation please make one video as DB video and Cassandra we'll try to uh create that particular video also right okay so now uh I think I can start the session uh but before that let me just write down

the agenda I'll be covering yeah I think my pen is running uh just a minute it's not running okay so today uh we'll be first of all talking about data transformation data transformation then we'll be talking about the model training part so first of all let's discuss about the data transformation inside data transformation what we have to do I think you remember we did uh already notebook experiment notebook experiment we already have done right so there I think you remember we did one notebook experiment on the feature engineering so

inside feature engineering guys what are the things I have done can anybody tell me if you referred that particular notebook I think remember so let me show you um there is a notebook I think you remember uh feature engineering notebook so we have done that uh so handle missing value duplicates value then uh we just converted this particular year uh like company year to the age then we uh did uh PR split then after that I just perform feature transformation because I was uh like uh I saw that this particular data is not a normally distributed some of the column that's

why part from this particular uh feature transformation and I saw like lots of outliers as well right so uh I applied something called power transformer and then I applied uh encoding and scaling technique as well so see this is the complete pipeline you can see first of all I did uh standard scalar one H trans uh Transformer that me one H encoding and ordinal encoding uh that means first of all I did one encoder ordinal encoder Transformer and standard scaler okay so these are the feature engineering task I perform in this particular data okay then after

that I got my uh like complete new data okay that means my clean data this particular clean data was uh passing to the uh smart because I saw that it is a imbalanced problem right there is some imbalanced problem in this particular data set that's why I was handling this particular imbalancing data okay with this particular smart package now after that I was performing D split then I was training this particular model so these are the task actually I have to perform inside uh feature engineering task that

means inside data transformation task okay so I'll do the same thing here I just written like python script so inside that uh my component okay the component actually I'm going to write now uh I'm going to just write right now I'll be writing as a modular coding okay so that is what actually I will show you but the thing I'll be referring from here only the same code I just like created as a function and class okay okay nothing change I have done but before that let me show you the diagram like how this particular diagram will look like so let me open the

flowchart so data transformation diagram so here is the data transformation diagram guys okay you can see the same thing I'm doing so again I'm taking some constant with that I'm creating the config okay with that I'm creating the config then uh I'm getting the data from my uh validation artifact itself because validation will give you the status as well the data data path I was reading the data then I was performing all the things like let's say drop column water encoding okay ordinal encoding transform column numerical column okay everything

I was doing then uh you can see that uh uh then we'll get the pre-process pre-processor object like whatever things you will be performing here whatever things you will be performing here so you also need to save this particular pipeline as a pre-processing like pel format so that you can also perform the same same thing on top of the test data because we have learned whatever things we usually perform on the train data the same thing we have to also perform on top of the test data okay that's why we have to create this particular pre-processor file okay then

we'll be also doing the target M mapping and all then uh we'll be doing this mod operation that means uh uh this U uh I mean imbalance data handling then after that you can see our uh data would be ready and this particular data I'm storing in the artif folder inside data transformation okay and as well as I'll also save this particular pre-processor P file because this is going to be my uh pre-processor Pipeline and with the help of that I will pre-process my test data as well and after um like performing all the operation I'll save my data as a train. npy test. NPD what is npy npy is

a format okay it's a file format it calls numpy format okay because it is going to be an array at the end if you see here um if you see here this is this is nothing but it's an array see after performing all the transformation it's an array now if you want to save this particular array in the format you can use this particular specific format called um this one npy npy is here it's a npy format I was asked in an interview today but you asked gab H so this is the flowchart we'll be following to implement this particular component so now let me go back to my code um but before that let me show you

the GitHub repo I'm committing my code so this is the GitHub repo so let me give you the link so from here you will get all of the code whatever things I have written so far right and make sure you set your uh environment variable guys that particular mongodb credential it should be available inside your environment variable okay otherwise you may uh get some issue that means he will U like face some issue whenever you will execute this particular project all right how to create a custom ml architecture okay fine so now uh

I'll open my code editor Let me close this at the tab H so the first thing I need to update my constant uh I think remember here I have already WR first of all I have to update my constant so let's update the constant so here uh this is the data transformation related constant I have already prepared see again I'm just creating a data transformation folder inside artifact inside artifact there would be a folder called Data transformation inside that there would be another folder called transform inside that I'm going to save my data so this particular data you can see this trend. npy and

test. npy so this data these two data should be available inside that transform folder then I'm creating another folder called transform object that means you can see one is pre-processing object we are having that means the entire pipeline object like whatever things we are performing let's say monot encoding ordinal encoding then uh we are also performing something called standard scalar whatever um pre-processing or let's whatever feature engineering task actually I'm doing with the help of this particular escalan pipeline I have to save that particular pipeline as a preprocessing p object so

that I can also perform the same thing on top of my test data as well okay that is what actually I'm saving this particular file inside this particular folder called transform object okay now you can see that this particular folder should be created now these are some actually uh very important uh constant related data transformation and this would be more clear whenever I will execute the program whenever it will create the folder inside the artifact you will automatically get the idea which folder will be created at the very first which folder will be created at the very last okay so yes this is what

actually we have to write inside constant now the next thing we have to update something called entity so again let's open up my entity which is nothing but my config entity and what is conf config entity config entity is nothing but it's a at the end it's just a path of my constant okay like the proper path of my constant so here inside this particular uh U entity I'll will assign my path so let me just show you data transformation okay now you can see here I'm just joining this particular path one by one and I'm just returning okay

I'm just returning in a data class uh I mean um variable okay so you can see data in directory transform file path then transform test file path and transer object file path that means whatever things I mentioned inside the constant I'm just returning the path one by one okay that is the task we we usually do inside config entity fine so now now the next thing what I have to update I have to update my artifact entity as well so let me open the artifact entity so inside artifact entity first of all I will update my

data transformation related artifact so it will just return your transform object file path that means these are the path it will return one by one your train data test data file path and as well as the pressing. pkl file path okay you can see transform object file path that means pre-processing file path transform train data file path and test data file path okay these are the things it will return after doing the transformation object that means what will happen if I open up my Blackboard let's say this is your data

transformation stage so this is your data transformation St after running it it will return uh three things actually as an artifa the first thing it will that particular pel file then it will return the train data and it will return the test data for the model training OKAY model training so it will uh receive this particular uh object and it will train the model then it will give you the your train model so you will get your train model OKAY Trend model you will get here this is the complete idea so here I have written this particular data transform artifact as well now one thing uh we

have to do inside uh I mean uh model training I'll tell you whenever I'll train the model I also need to calculate the accuracy score so this is a classification problem statement I think you know that in classification what are the evaluation matrics we are having we are having something called F1 score we have having something called Precision recall okay so these are the like uh I mean Evolution metrics I have to consider okay in this particular project so this thing we'll be writing whenever we'll do that model training but as of

now it's not required then uh after that I'll have to update my component so let's open up my component so here is my component called Data in. bu and from here I'm going to import some of the library at the very first so these are the libraries are required now one thing I don't have which is nothing but my target value mapping okay so I have to create this particular Target value mapping inside entity so inside entity see what is Target value I think you remember so if I open this particular in feature engineering so here one thing I was

doing I think you remember let me show you see here we are doing the target value mapping because Target value is nothing but it just a if you see the target okay if you see the target of my data so here is the target it's a categorical data it's a denied or certified that means yes or no statement right but it is written in a category but I have to do what I have to convert this particular statement as a numerical representation so that is what actually I was doing here I was doing the target mapping so wherever it is having the denied it will make it as one otherwise it will make it as a zero

okay so this particular things I have to do that means Target valum mapping instead of writing manually this thing okay instead of writing manually this thing inside our code what we will be doing we'll be just creating a file here inside entity I'm going to create a file and I'm going to name it as St met okay estimator do PI so inside this particular estimator do PI I have to map my target value okay so how to map the target value so let me import some Library first so these are the libraries are required let's say I'm using system package pandas pipeline okay uh then

exception and login so here I'm going to write a simple class okay and I'm going to name it as Target value mapping now you can see here certified would be zero and denied would be one when so reverse mapping here I'm doing so whatever things actually we are getting we're just mapping this particular information okay that means what will happen if you give your uh like data complete data whenever it is having the certified it will make it as zero whenever it is having the denied it will make it as a one that is the thing actually I'm doing inside this particular class okay that's

it I'll save this particular file and I will go back to my data transformation and let me check if anything is missing or not so everything is fine I think you can see so we have imported all of the library whatever Library we need and this particular Library I think you will already familiar with we already used inside our jupyter notebook some of the libraries are new let's say we are importing this save object save naai ARR read ml file drop columns okay so these are the library uh function are available inside your main utils and what is main utils I already told you

what is the use of utility inside a project okay so these are the like uh frequent uh written function I just kept inside the main utility whenever I required it I'll just call it from here okay I'll just call it from here so that is what actually I'm doing fine then I'm importing my target value mapping from my estimator itself so what uh what is the use of this particular pckl file uh I think you know that see whenever you do the you do the let's say standard scaling monitor encoding okay ordinal encoding on the

test data on the train data so the same uh I mean the pre-processing technique you have to also per on the test data yes or no because test data would be the same data now as per your training data so you also need to clean that particular data you also need to analyze that particular data then you will give to the model okay that is why this particular P file is required so this particular P file contains the entire pipeline we have done inside the training okay yeah yes you are right so that is what

actually you have to do now uh I just need to uh create this particular US Visa uh um data transformation class so see it's a very simple class uh it might be like uh little bit bigger codewise it might be little bit bigger because we have added so many actually logging you can see so many logging we have added that's why this this code seems to be like a little bit bigger but it's not actually that much of bigger actually you can see here so this is my data transformation class inside that I'm just writing class

variable and uh here you can see I've written one start static method what is static method static method means the method we can write without using any kinds of self self statement you can see everywhere I have to write the self statement but inside static method self is not there so I can access this particular function anywhere right then I have created another function called Data Transformer object okay what is data Transformer object so this particular function will make everything that means that that particular entire pipeline it will run so let me show you so this particular entire pipeline it will run

see the entire pipeline it is running if you check it here the entire pipeline it is running mon encoding a coding Transformer standard scalar that means whatever things we have written down inside our notebook everything we are doing inside this particular Pi file as a python isit as a python class and function okay that is the only idea now I'm logging everywhere like uh after each and every like line I'm logging the information so that I can also get to know okay so that's how actually it's working okay it's a best practice

section always now once my uh this particular function is completed I just return another function called initiate data transformation because this fun particular function I have to call otherwise I'm I mean answer until I'm not calling this particular function your code won't be executed so I have to call that particular function why I'm calling I'm calling inside initiate data transformation inside that you can see I'm calling this particular function called Data transformation object okay I think here sorry get data transformation object yeah here actually I'm calling then I'm calling the processor object

I'm loading the train data and test data I'm just dropping that uh this particular column like Target column and I'm taking the input variable and Target variable then I'm just calculating the company Edge then uh I'm just dropping the schema like what the drop schema you need see here I have written how many columns you have to drop these are the two columns you have to drop so everything I have written in that particular file okay then I'm just dropping then I think after that what I was doing I'm just uh saving my data okay I'm just saving uh sorry I was just

performing the target value mapping then after target bow mapping you can see uh I'm doing for the test data as well then I was saving my data see in a uh in know NPI format you can see np. array now once it is done I was saving my data in the NPI format see save data to the Nay array and this particular function I have written inside utils okay here is the function it's a very simple implementation only you just need to go through this particular code BS you will get each and every idea uh so here I've written in a model bar way I'm

wrting in a class and function that's why it looks little bit bit but but uh it's like very easy to understand at the end right yes then if exception is going on I'm just raising the exception so I think my component is ready now I have to execute this particular component uh in a pipeline so let me open up this particular pipeline training Pipeline and inside that I'm going to mention um just a minute rning pipeline inside that I'm going to first of all UT my data transformation then I need my data transformation

config then I need something called Data transformation arate this is a data transformation artifact so here I have to mention the artifact and here I have to mention the config now everything is fine you can see everything is fine then I will initialize this particular data transformation inside this particular class variable then I'm going to create this particular method called start data transformation start data transformation okay so now it will call this particular data transformation from the component itself and it will call this particular method called initiate data transformation which I have written

here so somewhere I have written initiate dat see this one it would be called now let me save I'll go back and here I'll mention this particular uh method inside run pipeline so let me just mention it is available inside run pipeline so it will take the data in artifact and data validation artifact why it will take the data in and data validation artifact because it should load data okay it should load the data and it should see that status validation status whether this particular status is true and false okay that's why I loaded

this particular two object now if everything is fine I'll save this particular file and let me execute the program so for this I'm going to execute this particular demo. pi and here I'm going to open up my terminal and let me clear it I'll just write python python M by sorry it's a python demo.

my GitHub token expired so you have to create a permanent token for this how to create uh see I'll tell you how to create it first of all let uh just um complete this particular execution see execution is going on and uh it has created I think some AR let me check yeah so data validation okay so it's executing still I think let's wait see for GitHub token uh what you can do just go to your GitHub go to the settings so here is the settings you will get uh I'll go to the settings and left hand side you can see uh developer settings go to actually personal access token go to token classic and here you

will get something uh generate a new token then token classic then you have to give the password okay let me give the password so here you can give a name of your token and just expiration you should keep no expiration then after that select this particular repo and create this particular generate token and here you will get this particular token copy and try to use okay so I think who is asking the right you will get your token like that so till data ination project is simple after data ination project is getting uh complex definitely see uh

there are some complexity level of a project so it's an end to end implementation we are doing and uh for this actually I'm expecting you are familiar with uh like Python programming language because everything we are writing as a class and Method at the end see everything we writing as a class at metal okay that's why I told you so if you are not familiar with python so it would be a little bit difficult for you right but it's it not actually complex uh so see I think execution is done it's not complex see I have given

all the flowchart and everything just try to understand and just execute this particular code from your end and try to understand see how we can debug a code let me tell you how you can debug a code and you can understand let's say this is my endpoint demo.

pi and I can see here I have imported one particular class which is nothing but my training pipeline so here I've initialized the training pipeline then I'm calling this particular run pipeline so here my execution is happening so from here my execution is happening what I will do I press control and I will just simply left click on top of this particular run pipeline then I will see here is my my run pipeline okay from here actually it is executing my code now here I'll just try to understand okay inside run pipeline what is happening first of all it is running the data in okay now let me see why this particular data in is coming again press control and left click on top of it now see data inion is

redirecting where it is redirecting to this uh here start data in it is available inside my you can see it is available uh inside my yeah component uh component I think yeah you can see start data inion it is available inside my component the component we have implemented see data inion okay component uh no no it's not in a component it's in a pipeline I think so sorry let me go back start data yeah it's in the pipeline you can see it's in the pipeline training pipeline you can see pipeline training pipeline okay it is inside training pipeline now just try to see inside this particular data inition

what you are executing you're executing this particular class called Data inition now just try to see where it is available you can see I'm calling this particular method I'll just go inside that and now it is coming from the data injetion component okay now inside that just try to see what I'm doing the same thing whatever things we have performed inside our notebook whatever things we have performed inside our notebook okay the same thing we are doing here the same thing we are doing here in a modular way okay there I was writing as a python simple script and here I'm writing in a class and Method

see here I'm writing in a class class and Method so that's why I'm telling you op concept is required unless and until you are not familiar with OP concept it would be little bit difficult for you okay and I can't do anything that's why I told you for the prerequisite you should know the advanced python here okay because see Basics thing every word teach you but the complex thing nobody will teach you right so that's why I'm bringing this particular complex thing to you like how like actual uh endn

implementation happens and what level of code you have to write because we all know about coding but we are not utilizing that particular coding skill in our project okay that is why I'm bringing these kinds of project so that you can understand okay we have learned the class and Method inside uh uh we are uh we learning class and um like method inside our python but why to use this particular class and Method okay we have uh let's say learned static method we have learned let's say in inheritance concept but where we have

to use it in a real world application so there you have to use it okay that's why I'm using using this particular class concept otherwise I could have written this particular code in a simple python script way okay got it yeah thank you thank you everyone so I know that initially it may u i mean a little bit complex and it will uh yeah yeah so it I mean initially it would be a little bit difficult for you but uh if you put some dedication okay if you just put some dedication if you just uh put put little bit of time to understand what is happening inside this particular code definitely after

learning one implementation you can use this particular simplate for your entire project implementation you'll be doing okay because it's not possible just write the code line by line in front of a class because it's just a one and a half hour of session okay one and a half hour of session so I can't write the entire project in front of you because it's a very big project if I want to write the code from scratch so maybe I I could have conduct like one month of session right 30 days of like project implementation session so but

again it's not I mean suitable right so that's why I I'm just explaining the things in a high level way I'm giving you this particular flowchart and everything and I'm showing this particular code and how we have written okay now what should be your task your task should be just debug this particular code the way I showed you but first of all figure out the end point and try to debug try to debug that s okay try to debug that so this is called debugging fine that's how flow happens see I've already written all the things like uh which file you have to refer at the very

first constant entity component pipeline main file that is what we need now to understand one implementation because here we are not doing any kinds of simple python project like uh I mean any kinds of um simple scripting project we are not doing Cassandra DB Cassandra DB if requires I'll tell you but learning Kasandra is not a big deal if you can connect with mongod you can also connect with Cassandra as well just need to know see don't ask these kinds of question let's say how I connected with my MB

remember see see that how I connected with my mongodb I showed you the mongodb script right so here is my mongodb script uh let's say if I want to connect with my mongodb here is the script I was using p package right I was using p package I was creating the client I was creating the DV I was creating the collection and there was I was inserting my data now try to see how to connect with the cander DB for this you can simply search on Google how to use Cassandra okay Cassandra with python you will see the

code snipp it how to connect it okay you will see the entire code snipp it how how to connect it python implementation I in here see the entire connection code it is available entire connection code it is available okay just try to read now it it will take only five minutes to read right it's not a very big actually I mean uh uh uh blog just try to spend five minutes try to read and try to copy the code as it is do the experiment on your notebook on your notebook try to see whether it is able to make the connection or not if it is able to make the connection

congratulations you can use that particular C DB no need to learn in depth only just try to see how to connect how to insert the data how to retrieve the data this particular three task you to them that's it okay yeah because see at the end what you have to be you have to be a self explor I mean say a self- Explorer uh if I'm showing something uh you have to improve that particular things okay by yourself and try to tag me as well on LinkedIn or any kinds of social media uh you have done this particular work okay

by referring my video I will happy to see that particular things okay from your side okay so definitely I'm expecting guys uh you should give some of the like dedication from your end then your actual learning will happen instead of relying on the mentor always okay yeah fine so as a guide I can give you this particular I mean uh knowledge and everything but to be a developer you have to write okay you have to write always not the copy paste so try to open this particular code try to line one by one in your system okay this will make you like good

knowledge about the coding fine yes so I think now uh data transformation has executed now if I show you my execution data inition done transformation is also done validation is done now if I show you the transformation see this is my data in a NPI format and here is the pel format that means completely uh this particular three stage is running so my data ination validation transformation without any error yeah thank you Narendra fine so till data transformation we have written everything till datation we have written

everything and again see it's not necessary to understand my flowchart always because I have created the flowart in my understanding with by my understanding okay how I understand this particular project but for you how you are understanding this particular project that's how you can create this particular flowchart and again to create the flowchart you can use this particular website called ml.

com okay so this is the website you can use to create a flowchart okay you can use this particular website so just create an account let's say I will login with my account and you can uh create an account inside that it is a like very simple interface with the help of that you can create this kinds of flowchart okay this kinds of flowchart fine so now let's see um so now let's do the model training part so for model training again I'll will follow the same approach so let me first of all close these at the tab one by one I don't need it h so the first thing I have to update the constru again because constant

always we have to update at the very first because it's a variable declaration we are doing right always so again U I have prepared the model trainer related constant let me show you model trainer constant uh here is the model trainer constant see model trainer constant again it will create a folder inside artifact called Model trainer inside that it will create another folder called train model inside that it will say by model called model.

pel so this particular model I'm going to train right now this particular model I'm going to train right now with my data right but what should be the name this should be the name model. p now you can give any kinds of name it's up to you now training expected score okay training expected score means how much score you are expecting from your model okay let's say uh you are training one model and you are expecting some accuracy let's say after calculating my evaluation Matrix we're expecting some accuracy if my accuracy is getting that much of accuracy I'll will accept this

particular mother otherwise I won't be executing this particular model okay so this thing is required whenever we do the model push operation in the production so there first of all I will check the train model and the production model uh let's say how things will H happen let's say this is my train model the model actually I have train right now this is my train model the TR model I have train right now and this is my production model the model is available inside my production server so after training I will bring this particular

model I'll also bring this particular model and I will calculate the evaluation Matrix okay then I will check which model is performing better if this particular model is performing better I'll push this particular model to the production okay if this model is not performing well not performing well I own be pushing this particular model to the production because already my production model is working fine I don't need to push this particular model but if accuracy is fine if it is selected then I will push this particular model that's why this

particular score is required okay so as of now I just decrease this particular score because it just a uh like I mean demo I'm showing you okay but whenever you are creating your own model okay whenever you are creating your own model try to increase this particular expected score at least uh more than 80% more than 80% of accuracy you will expect from your model okay otherwise you won't be accepting this particular model as of now let's keep it as Z 0.

6 okay just to show you f okay so now uh this particular path my model. yl file path what is p p model means from where it will get production model it would be available inside S3 bucket okay uh it would be clear whenever we'll do the model pushup part that then it would be cleared uh whenever we'll be doing the production model okay setup I will be using Amazon Amazon S3 bucket okay to keep my production model there so here you can see I have my config and inside that I'm having my model. EML but inside model.

EML I don't have anything okay see what I did inside my notebook remember what I did inside my notebook uh in the model trainer notebook I just uh I have taken so many model you can see I have taken so many model all the classification model then I was training one by one all the model I was training one by one all the model then I was checking the accuracy score I was selecting the model based on the accuracy let's I selected three model then what I was doing again I was doing the hyper parameter tuning operation with the help of randoms RCV okay so

this thing instead of writing here okay instead of writing this thing here I can use one package called neurom MF okay neuro MF so this is one like machine learning practice you can see with the help of neur MF you can do do this particular automation okay you can do this particular automation see when I was uh working in Inon okay Inon like development team we had implemented this particular NE MF package okay so see this is from my NE organization only and we had implemented this particular package even I personally uh I mean

hosted one package which is nothing but image Seeker see I'm more interested in deep learning so most of the work you will get around deep learning okay if you see my work see this is a package actually I have developed okay entirely I have developed and this is my package so image Seeker is a package for the image classification so let's say if you don't know about coding if you don't know like how to do the image classification with the help of tensor flow okay kasas what you can do you can use IM SEC card so see if you want to do the classification you have to write tons of code okay in Kaz or tensorflow

whatever framework you're using but if you're using my package only just need to install IM secer open it you will get an interface give your data okay train it after training simply you can do the prediction see this is the prediction UI just upload your image you will get the prediction okay so this is the package actually I have developed long back I think it's around two years back I have developed so let me show you the source code this is the homepage and here is the source code of the IM sec I have developed see this is the source code it is available for public you can also

contribute in this particular package okay see this is the entire package I have develop okay so that's some uh some work actually we have done uh for the Community as well so that they can use see NE we have developed because we saw that whenever people are doing these kinds of hyper parameter tuning so they are having lots of issues okay they're having lots of issues like they have to set the like parameters okay they have to like Define the model everything they have to do from scratch okay but can you automate this particular process definitely we can automate this particular process

okay that's why we have created this particular NE MF package so NE package what it will do it's an open source Library designed to avoid writing duplicates code you can use this particular model psychic Lear without writing any code model training can be controlled by the configuration file itself okay only you just need to prepare one configuration file that's how Okay the config file I have given inside the config EML this is the model. EML so this particular config file you have to manage how to manage this config

file let me show you it's very easy if you see the documentation of in MF you will see that that's how we can Define this particular model. emlc let me explain okay what I have done so here first of all the first key I have done which particular hyper parameter method you want to use I want to use GD sear CB okay GD s CB you can see I want to use escalan okay scalan library for this and you have to give the CB parameter I think you know inside you having some parameter like CV value varos value okay then model selection which model you want to use I want to

use K sber model I want to use random first classifier okay these are the model only because from my experiment I saw that uh from my experiment I saw that random forest and Kish neighbor is working better okay this particular two model is working better I don't need the ex boost classifier because it's having low accuracy so I'll only do the hyper parameter tuning with the help of this particular two model you can also use ex boost for this you need to additionally add the ex boost here but I'm not using

ex boost I can see that canish neighbor and random for is working better so I'll do the hyper parameter tuning between this particular model so that's why I have set it so my all the parameter of my carious neighbor all the parameter of my random Forest classifier the same parameter I have mentioned here you can see here the same parameter I mentioned here also okay here also the same parameter I just copy pasted okay GD search CB is a hyper parameter tuning like I mean technique okay you can search on escal and you will get it then

this particular configuration file is ready no need to do anything only this particular configuration file you have to write and you have to install one package called neur MF I think it is available inside requirement. TTC neur MF we are installing okay now everything would be done automatically now let me show you so now I have to update my config entity first of all let me update the config uh here is my config entity for the model training I'll copy and here I'm going to mention it see this is my model trainer config again I'm just returning the path One By

One The Path I have mentioned inside my constant that that's it then I have to also mention my artifact entity so let me open the artifact entity and here let me mention it so artifact entity Dev I have to mention two things my model trainer artifact as well as the classification metric artif because I'm going to create a classification right uh I want to evaluate my model right how to evaluate for the classification model F1 is score precision and recall okay this particular three metric I'll calculating I'll save it and this is my mod artific that means after training the model it will return the model and

it will return the classification metric as well okay for the next component okay the next component will be doing so once it is ready then I have to update my component so let's go back to my component called Model trainer inside that I'm going to write my model trainer so let me just write it quickly see again the same model trainer I'm doing nothing else I'm just training the model doing the hyper parameter tuning training the model now one thing it is missing called s estimator US Visa model now here inside that I have to mention my Visa

model so estimator do PI here is my Visa model see I'm just loading my pre-processing object that means pipeline object the pre-processing object I have saved here inside my transform folder because with the help of that I'll pre-process my data now okay then I was loading my train model object as well okay then after that what I'm doing see this is what actually I'm using for what this is what I'm using for the prediction that means the evaluation purpose I will do the prediction on top of my test data okay

that's why I'm creating this particular function instead of writing here I'm keeping inside a separate file so that whenever my model will get train I'll test my data on top of this particular train model okay that's why I'm keeping this particular uh I mean class in a separate separate actually file called US Visa model okay now you can see that I have already imported everything now everything have imported by by one now here is the class I have written called Model trainer inside that I'm defining the class variable get model object and report so here you can see this is the

neur MF work you can see so you have to write this m of code actually for the neur MF just Define the model Factory from NE MF see from nef model Factory inside that just give your configuration file the configuration file we have created this configuration file path you can see I'm giving this particular configuration file path okay configuration file path now I'm getting my data train train data y train data X TR data y t data okay then I'm there is a method inside model Factory called G based model see the example is already given here if I show you the code wise if I show

you here model Factory then after that you have to get the best model you have to pass your data and you will get the best model OKAY the same same thing we are doing so get the best model give your data and the model you get actually you have to save this particular model as your best model okay then after that I was calculating the acces code and saving the ACC Cod inside my classification metric okay inside the artifact now I uh once it is done I was just initiating this particular model trainer again I'm loading my data okay

then I'm getting my best model from my new MF I'm loading my preprocessing object I'm just uh doing the prediction of of my test data then I was saving my model and I'm returning this particular artifact that's it very simple code and with the help of NE MF I'm automating everything okay the same uh code actually have written inside feature engineering for the hyper parameter and all okay the tons of code we have written no need to write prepare this particular config file uh put uh put the data inside NE everything would be done automatically okay so this one Library

you can use okay going forward whatever ml project you are doing and if you want to do the hyper parameter tuning you can use this particular new M package for this fine so it's done then uh let me uh mention in a pipeline now so I'll open up my pipeline training Pipeline and let me import my model trainer training pipeline I will import my model trainer then I'll will import my model trainer config then I'll import my model trainer artifact then I will initialize my model trainer config and then I'll write my model trainer function start my model

trainer after data transform by model TR start let me save then after that I will run the pipeline so inside that I will call my model so model trainer object I have to call it after transformation I'll call my model trainer here so I hope everything is fine let me check one more time m my model trainer uh imported everything there is nothing missing okay fine so now let me execute the code one more time and let's see whether it is able to train the model on so I'll clear and I will execute my demo. again okay let's see so this is going to be my

new execution right now so let's wait for some times and see uh the model training part so in between if you have any query guys you can ask me in the chat okay I think injection is happening injection is done I think it will start the see my time stamp has created it is

doing the feature store that means data in is happening right now it's of training the model every time can't to use the model dople no unless and until you are not training the model how you'll get it now how we get that model. P file now I'll get the model.

P file see my execution is done and why it is so fast because see for your system it might be slow for your system it might be slow it will take some time because it will do the hyper parameter tuning but for me it is super fast because I'm having good configuration machine actually I'm using core9 processor that's why it's like super fast for me but if executing it will take some time now here you will see the model the train model model trainer see model.

pel is now available okay so that means my trainer is working fine as well now here is the model trainer flowchart the way I explained everything is mentioned in the flowchart as well see same thing you can see later on so now let me push the code in my GitHub so model trainer added done so now I think if you go back compare I will show you okay in the next class Nara will show you whenever we'll do the model Pusher part that time I'll show you how to compare the model model with my production model as of now just try to consider you you have trained your model the data you have

prepared so far you have trained one model okay and this is my first model right now and this model would be pushed to my production because in the production at the first time there would be no model now because we are training the model for the first time so whatever model we're training for the first time this would be available directly in my production then whenever it will the second time execution that mean second time retraining okay that time it will compare with the production model whether the current train model is good

than my production model or not if good we'll push it otherwise we won't be pushing it that is the idea I hope it is clear right now okay so successfully we have trained the model and everything is working fine and all the code is available inside my GitHub please try to clone this particular repo and execute in your system one more time okay yeah so yes uh this is uh all from this particular session today tomorrow we'll try to see the model evaluation and the pushup part we'll push the model to this

three bucket and we'll try to create a user application for this particular application so we'll be using first API okay to create the user application so here you can see I have mentioned the first API I think somewhere yeah first API will'll be using you can also use flask Django um then stream lit but I will be using first API because first API is a production ready uh framework you can use okay instead of using any other one so yes uh this is all guys uh and please try to subscribe to my channel if you haven't subscribed yet um

and please try to like like this particular video and share it with your friends and family so let's hit 12 12,000 subscriber as soon as possible guys it's like very close uh so please support my channel so I'll bring lots of content for you fine so yes I think now we can end the session if you have any qu you can ask me otherwise I think we can drop off the session let me know guys if you have any query otherwise I think we can end the session which your um you will get to know very soon Valu okay yeah thank you uh sir awesome yeah thank

you okay fine so uh bye everyone take care so please join in the next class and next class update you will get from the live session itself so live section please try to check uh whenever I will schedule any kinds of session it will come here yeah so with that guys thank you so much for joining the session bye everyone take care bye good night

Show thinking

This session focused on implementing the core machine learning steps of the project: **Data Transformation** and **Model Training**. The logic previously developed in Jupyter Notebooks was converted into robust, modular components for the automated pipeline.

**⚙️ Data Transformation Component**

This component is responsible for all the feature engineering tasks required to prepare the raw data for a machine learning model.

**Workflow**

1. **Input:** It takes the validated train.csv and test.csv files from the Data Validation component's artifact.
2. **Preprocessing Pipeline:** The heart of this component is a **scikit-learn Pipeline** that chains together several transformation steps:
   * **Categorical Encoding:** Applies One-Hot Encoding and Ordinal Encoding to the respective categorical columns.
   * **Handling Skewness:** Uses a Power Transformer to make the numerical data more normally distributed.
   * **Scaling:** Applies StandardScaler to all numerical features.
3. **Saving the Preprocessor:** This entire scikit-learn Pipeline object is saved as a pickle file (preprocessor.pkl). This is critical because the exact same transformations must be applied to new data during inference.
4. **Handling Class Imbalance:** The **SMOTE** algorithm is applied to the training data to create a balanced dataset, preventing model bias.
5. **Output (Artifact):** The component outputs the paths to:
   * The saved preprocessor object (preprocessor.pkl).
   * The transformed training data (as train.npy).
   * The transformed testing data (as test.npy).

**🧠 Model Training Component**

This component takes the transformed data and trains the machine learning model. A key highlight of this section was the introduction of a tool to automate model selection.

**Automating Model Selection with neuro-mf**

Instead of manually coding the complex logic for comparing multiple models and tuning their hyperparameters, the presenter used **neuro-mf**, an open-source library designed to simplify this process.

* **How it Works:** You define the models you want to test (e.g., k-Nearest Neighbors, Random Forest) and their hyperparameter search spaces in a **YAML configuration file**. neuro-mf reads this file, runs the search (e.g., Grid Search), and automatically returns the best-performing model. This makes the training code cleaner and more configurable.

**Workflow**

1. **Input:** It takes the transformed train.npy and test.npy arrays from the Data Transformation component's artifact.
2. **Automated Training:** It uses the neuro-mf library to find the best model based on the specifications in the model.yaml file.
3. **Saving the Model:** The best-trained model is saved as a pickle file (model.pkl).
4. **Output (Artifact):** The component outputs the path to the trained model.pkl and its performance metrics (like F1-score and precision) for the next stage.

The presenter successfully ran the entire pipeline, which now includes Data Ingestion, Validation, Transformation, and Training. The execution resulted in the creation of the preprocessor object and the final trained model, all stored neatly in the artifacts directory.

**🚀 Next Steps**

The upcoming sessions will focus on the final stages of the pipeline:

* **Model Evaluation:** Comparing the newly trained model against any existing model in production.
* **Model Pusher:** Pushing the best model to a cloud storage location like AWS S3.
* **Prediction Pipeline & Deployment:** Building a FastAPI application and deploying the entire project.

e e yeah guys uh hello everyone am I audible to all of you just let me know in the chat if I'm Audible just uh give me a quick confirmation if I'm audible to all of

you I and uh it's been long guys uh again I started uh taking the live extremely sorry for that because um I was having some personal issue and uh because of because of my busy schedule actually I was not able to conduct the session so extremely sorry for that so uh from today actually I'll try to again start the community uh this uh live session and all and I have lots of plan guys uh yeah so let's complete this implementation and uh again I have lots of plan on for you in this channel actually I'll be taking actually different different content different different live session so that it will

help you a lot yeah hi hi s hi Aris hi gopal yeah thank you guys thank you yeah sorry guys uh I already told you I was uh having some personal issue and uh I had some busy schedule that's why I was not able to continue but no need to worry again I'll try to continue okay H so I mean currently machine learning project implementation is going on I have also plann uh I will also bring up deep planning implementation uh computer vision related project implementation then NLP then generate okay all kinds of domain actually we'll be discussing the

end to end project implementation yeah hi Aris yes okay so guys if everything is fine uh you can give me a quick uh confirmation I can start the session let me know if everything is fine can I start and please uh share this live with your friends and family so that they can also get to know again uh we are contining with the live session yeah everything is fine anupam yeah so please uh hit the like guys all of you please hit the like and uh if you haven't subscribed to my channel so please try to do the Subscribe okay let's hit the 14 14k as

soon as possible yeah thank you thank you AR and Happy Teachers Day all of you I would like to wish you Happy Teachers Day all of you okay so I think now I can start with the session guys yeah so thanks thanks for joining the session and please uh try to subscribe and hit the Bell icon also so that if I'm uh actually bringing new new content you will get the notification okay immediately fine so as of now I think you remember we have completed till I think uh model trainer if I'm not wrong if I go to the live section so yes so I just uh

completed till data transformation and model training part okay so in this video actually we'll be discussing about uh this one uh our model evaluation model pushup as well as the prediction pipeline we'll be implementing yeah hi hi everyone Hi sik hi sandip uh hi BRX Danny yeah so I'll I'll give you the quick revision whatever things we have covered h fine so let me open up my Blackboard and let me write down uh everything here okay so first of all let's see the quick revision uh so far actually what are the

things we have covered um so I think we have uh already seen the introduction of the project introduction of project and setup okay then I had also discussed about the database okay database like how we can use mongodb database how we can connect okay and how we can get the let's say uh connection string and all everything have I've just discussed okay then I also discussed about uh I think Eda that means exploor data analysis then feature Engineering also okay feature engineering I also covered then I I think I started with modular coding uh I showed you the data injetion

part okay data inje from the mongodb atlas okay then after completing data ination I already discussed about data validation validation and inside data validation I discuss about how we can verify okay uh how you can verify schema okay schema of your data and what is schema anyone remember what is schema yeah Happy Teachers Day thank you thank you DPI guys uh do you remember what is schema actually schema of our data remember uh in the data validation um I think uh session I showed you how we can validate our data

anyone remember guys schema schema means I just prepared all of my column okay Columns of my data as well as the their data types okay data types So based on that actually I was validating my data whether my training data okay as well as the test data is having the same schema or not let's say whenever you'll be doing the prediction okay prediction or retraining operation so there is a possibility your data might be changed okay your data might be changed that means let's say some of the column might uh would be

let's say missing inside your data that time your model will uh sorry your let's say code will uh I mean break okay in the production so it's not actually uh efficient way we can uh let's say deploy our project to that cloud platform so always we need to verify our data first of all let's say if you're doing retaining operation if you're doing let's say any kinds of testing operation first of all try to verify the data whether it is having the same schema or not okay then I think another things I discussed called Data drift prediction

okay the data drift with the help of evidently have evidently evidently evidently AI okay then uh I had discussed data [Music] transformation data transformation and inside data transformation what do we saw guys uh how we can uh actually process our data that means we created the entire pipeline right like whatever things we used to perform inside our feature engineering let's say the missing value handling then standardization then uh let's say outlier handling whatever let's say things we perform inside our feature engineering everything we did

inside data transformation pipeline then finally I think I showed you the model training part model trainer inside model trainer what I did I trained actually different different model model then I perform something called hyper parameter tuning okay hyper parameter hyper parameter tuning okay so we used one uh Library I think remember called NE MF neuro MF okay with the help of that what we did we selected our best model okay so I think till here we just completed this project okay till model trainer that means we are successfully able to

train our model okay I think you remember and this is the uh GitHub link of our project now you can see the last comment I added model trainer so if I zoom this screen see model trainer I added now let me share the link with you as well so I'm sharing the link in the chat so now let me know if you are ready with the setup and everything like uh did you open this project in your system or you are only watching because uh I always uh say like uh if you are doing any kinds of let's say live session and if you are doing any kinds of projects it's better to open up your system as well and try to code with

me okay so that immediately you can ask me the question if you're having any kinds of Doubt or let's see if you're facing any kinds of error fine so please try to open up your system guys don't join from the mobile phone try to join from your system everyone and try to open your project with the help of V visual code Studio see I already opened and even I also activated my environment I think you remember we created one environment called Visa okay I also activated the environment that means everything is set everything is set and

this was my end point demo. Pi because demo. Pi I was executing my entire training pipeline okay I think remember and if you want to see the pipeline since I training pipeline I created all of the pipeline that means first of all data in will execute then data validation then data transformation then uh this one model trainer OKAY model trainer will execute now the next thing we'll be doing in this video uh sorry in this session actually we'll be doing something called uh agenda for today first of all we'll try to see how we can perform model

evaluation model evaluation okay then after model Evolution we'll see the model Pusher model Pusher to production and here we'll be using something called uh Amazon S3 okay S3 for the model saving that means we'll push our model to the production um we'll be using something called Amazon S3 okay S3 bucket to store our model then once my model Pusher is completed I will show you the prediction pipeline then we'll be creating the user app that means we'll be creating the front end okay of our application and here we'll be using something called Fast

API okay and tomorrow if tomorrow I free so what I will do I'll show you the deployment okay deployment we'll be using something called cicd okay cicd continuous integration and continuous uh delivery okay then I think this project would be completed now after this project I have a plan actually I'll try to uh show you one uh deep learning project implementation and to and to and deep learning production grid project then second I will show you end to end uh computer vision project then I will show you n

NLP project then I'll also show you N2 generative AI okay project uh after this project one more MLS project you will explain yes Vu I already listed down one from De deep learning one from computer vision one from NLP one from Genera VII I think uh this will cover everything okay this will cover everything right apart from that if you want to I mean learn anything you can let me know I think apart from that we don't have any domain now nowadays so inside artificial intelligence what we have we have machine learning deep learning computer

vision NLP and jni yes or no fine what are the key components of the stand out resume for an entry level data science of a position uh no need to AR aash uh actually what I'm planning tomorrow I will try to show you the deployment and after that I will keep one session for the carrier guidance and how I I'll I'll tell you that time how we can prepare your resume okay and how we can mention the projects how we can mention your experience everything I'll discuss in that particular let's say session fine so that you will have the complete project implementation after

that you will have one more session so that you can prepare your resume and then you can prepare for the interview that will help you a lot which MLF tools you are going to use cicd so here we'll be following actually AWS like cicd with the help of GitHub action but uh whatever project actually I have listed down here so there I will also show you how we can use the circle CI so we'll be learning about Circle CI Circle CI then Jenkins okay we'll be covering them as well fine all kinds of C CSD tools will be exploring how do you see the job

market evolving for the data science and ml POS Professionals in the next few years see uh just try to remember uh today I mean this world is all about artificial intelligence okay so I believe that in the job market actually ml DL then NLP CV okay jna so definitely this job would be available okay but people are more focusing on jna nowadays CU um we are having like very powerful model like different different large language model and all right but if you just think if you want to learn jni if you want to

learn the jni so what are the skill you should have first of all let's say you are very new to this field you are very new to this gener uh let's say artificial intellig Ence or let's say data science so let's say if you want to learn so first of all which technology will you'll be learning directly will jump in the Gen or you have to learn ml DL NLP CV these are the thing because if you have these are the knowledge definitely it would be easy for you to learn the jni okay and whenever you are learning these are the technology definitely you will also get some of the job Ro in the market because

it's not like that jni came in the market ml job has been destroyed no it's not like that still ml jobs are there because let's say we are having different different problem statement so if one of the problem statement can be solved with the help of machine learning then why people will be using gni for that because managing large language model is not easy it's like very costly so as a let's say company owner you have to also take care about the cost so if you can solve with the machine learning approach then why you'll be using large language model okay so I I feel like U

still this job would be there okay in future so no need to worry about like artificial intelligence whether uh it would be there or not definitely it would be there okay but you have to gain the skill okay you have to be a good developer it's not like that you are only learning as a high level no not like that if you're a good developer so you will be getting thousands of job okay yeah Docker also will try to cover Docker will be using to containerize uh enre uh enter our application and if you need separate Docker session I can also

contact okay fine guys so I think now all the agenda is clear now let's try to see the model evaluation uh how we can perform the model evaluation okay okay so see whenever I'm talking about deployment sorry whenever I'm talking about model evaluation so what do you think about model Evolution guys let me know let's say I'm talking about model evaluation tell me what we usually perform inside model evaluation any idea so I think remember we trained one model so let's say after doing model training model training what we got

guys we got one model let's say we got one uh this model I think you remember inside uh time stamp model trainer we got one model. pel okay Piel file we got so let me just write we got something called model do pickle okay we got this one now what we have to do we have to evaluate this model yes or no so it will go to the next stage which is nothing but model evaluation yeah so dendra is telling we check the matrix like R2 then MSC rmsc okay fine but first of all try to see what kinds of statement we are solving here if you have attended for the let's I mean from the very first session I

think you know we are solving one classification problem statement here now can you please name some of the classification metrics inside machine learning like we are having uh accuracy score we are having uh I think uh confusion metrics we are having AOC curve right then F1 is score recall is score right so these are the metrics we are having okay so in the model evaluation what what we usually perform we just try to see the model actually we got for our production that means we'll just try to test this model on top of my testing data so let's say here I'm having my

what I'm having my test data I'm having my test data test. CSV okay so this data I'm having and from where I'm getting this test do csb I'm getting from the data inition I think you remember just a minute okay so we are getting this test. CSV from where from our data in pipeline right because whenever you are investing the data I think you remember whenever you are injesting the data you are getting inside feature restore sorry inside not feature inside interestate folder training. csb as well as the testing. csb so training. csb we have

already utilized okay during training of our model now we have to take the test. csb and we have to evaluate the model okay so that is the main thing uh so this this is the main thing actually we have to perform here fine so now here what I do I'll try to take the test data and I will evaluate my model OKAY the model actually are getting model.

pel and here I'll just try to do I'll just try to check okay I I'll just try to check one uh let's say accuracy score let's say I will first of all set one um let's say threshold value okay I'll just try to take one threshold value threshold uh threshold score I can talk about why I'm taking is called I I'll tell you let's say as of now uh I'm considering let's say 0 uh 0 point let's say 6 that means 60% also you can increase and decrease it's up to you you have to decide based on your requirement let's say you need a very good model okay you need a very good model

that time you can keep this threshold score uh let's say close to 100% that means 90% 80% but if you see any kinds of let's say model machine learning model if you train you will get around 80% accuracy 85% accuracy so I'll suest try to keep this special value around let's say 0.6 or 0.

7 that means at least 70% that means what you will do you will evaluate this model on top of the test data okay that means you will be calculating what you will be calculating accuracy score okay accuracy score you'll be calculating let's say uh uh this one uh you can also calculate F1 score you can also calculate let's say this one uh Precision recall Precision recall AOC curve anything you can calculate here but let's as of now try to consider we are only cons let's say calculating acuracy code okay we are checking the model accuracy okay now let's see if my model accuracy if my model accuracy that

means after uh let's say feeding my test data if my model accuracy is 0.7 that means 70% okay if I'm getting 70% accuracy okay 70% per accuracy what I have to do okay what I have to do so here I'll just simply do I'll just return one statement here so there would be a parameter I'll be taking called is model accepted okay is model accepted is equal to I'll return true okay I'll return return true why I will return return true because in the next stage what I have to perform guys I think remember next St I have to perform

something called Model pushup model pushup and where I will push the model in the S3 bucket I have to push the model okay see if my model is accepted that means if my accuracy is let's say 0.7% uh that means 70% accuracy I'm getting okay after let's say feeding my test data that time I'll will give the permission yes you have to push the model right now this parameter would be true otherwise this particular parameter would be false okay this parameter would be false false means uh we didn't actually got

70% accuracy let's say you got 50% accuracy that time you are not considering this model okay you will only consider whenever your model accuracy is 70% that time your actually model will get push where to the S3 bucket S3 S3 bucket it will push the model okay but see this is the simple idea I have given at the very very initial phasee okay at the very initial initial phase now see we'll be updating this uh actually pipeline slightly so what we'll try to do see instead of only relying on our current model the model we have trained we also need to rely on the

production model now can anybody tell me initially at the very first time will be there any model in my S3 bucket tell me is there any model you can see in the S3 bucket at the very first let's say you are training this model for the first time so is there any model in my S3 bucket yes or no let me know S3 is my let's say production production actually model you can also call it as production model production model will be there any model in this three initially I think no right there won't be any kinds of model okay there won't be any kinds of model in this three

because we are training for the first time so at the first time what I have to do at the first time whenever I will run my model evaluation model evaluation okay on top of my test data on top of my test data so it will um what it will return tell me is model accepted this particular statement whether it would be true or false at the very first time let's say whatever model you are getting no need to worry let's say you are getting 50% let's say you are getting 60% accuracy or let's say you getting 70% accuracy so it doesn't matter so at the very first time what you have to do you

have to return this statement as true because in my production server there is no model so at the very first time what I have to do I have to first of all push the model push my train model in my production server so what will happen so thisel model Evolution will return model accepted is true then my model Pusher will execute model Pusher will execute and what it will do whatever model actually you train that means your train model uh your train model that means this one uh let me write train model your model dopel so this should be pushed to where pushed

to S3 bucket okay that means to the production server okay and there you will have one model right now okay now let's see you are training for the second time now let's see your training for the second time second time now what do we have to do so let's say we are performing model evaluation model evaluation and we got one model after model training which model model.

pel and we are having our test data okay now let's say my uh threshold is let's say 0.6 that mean 60% I'm expecting if my accuracy score is 60% that time what I will return is model accepted is equal to True otherwise it would be false okay this is the idea fine but see what you have to do you not only only you not only uh check this particular model this uh your train model but also you will be also checking the production model that means this model evaluation okay this model let's say this is my S3 S3

bucket and it's say S3 bucket I'm having my production model production model okay so first of all it will fetch this production model it will fetch this production model okay that means it will uh load this production model from the production uh uh from the S3 buet okay then how many model you will have guys how many model you will have how many model you will have you have your train model the first model your train model and second model your production model production model okay so what it will do it will apply this test data on top of both

model as train model as well as the production model now let's say if uh your train model is giving uh let's say 60% accuracy okay and production model is giving let's say 55% accuracy now tell me which model is good my current train model or my production model previous production model tell me which model is good current current train model or production model which I just fix from my three bucket very good difference between precision and recall actually both both are actually classification matrics uh

and if you want to understand them you can uh simply search on Google actually you'll see one equation okay like true positive false positive two negative false negative okay this kind of Stu I think you learn in your machine learning this thing so guys let me know which model is better your current train model or production model you just download it please reply so that we can have interactive s session guys Al together definitely my train model because here we are having okay yes here we are having 60% TC now tell me whenever my

model Pusher will execute let's say next stage is my model Pusher so which so which model actually it should push guys which which model it should push to the S3 bucket again that means in my production server train model right that means this model will get pushed right now why because you can see here we are getting good accuracy that means my previous production model whatever I had it will replace with my current model OKAY current model and in case let's say you are not getting this Mass accuracy let's say you are getting uh train model

55% and uh uh production model let's say you are getting 60% now tell me uh if I'm getting this kinds of let's say accuracy should I execute my model Pusher that time should I execute yes or no if you're getting this uh I mean this kinds of condition no so I don't need to push my model Pusher right now because I can see already my production model is working better so why I'll be pushing my I mean current model because I can see my production model is good if my production model is not good if my current train model is good that time I should have replace otherwise I won't be replacing okay so there would be a if

condition if my uh production model is not good that time model Pusher will execute otherwise model Pusher won't be executed difference between train model and production model both both are same right yeah see both model are same see initially what you did initially you didn't have any model you just directly push the model to this three bucket that means same model.

pickle right then second run whenever again you are training again you are evaluating what you have to do first of all you have to check the production model whether this model is better than your current model or not if not you will push that current model okay otherwise you won't be pushing the current model got it I hope it is clear thank you sir thank you so much along with that we yeah thank you thank you sash Shashank fine now tell me guys this this uh concept is clear yes or no this model evaluation concept is clear clear yes or no how model evaluation and model Pusher

will work let me know S3 bucket is a service from AWS okay it's a it's a storage service very good so there you can store any goft file so let me know if this concept is clear model evaluation and model Pusher how it will work in my project I mean how we are going to implement because see the same way I explained now will be also implementing with the code okay let me know if it is clear or if you're having any doubt I hope it is clear guys how model evaluation and model Pusher will execute right okay great then so now I think we can

uh we can start with our practical then right so let's open up our code guys so I think you remember the first thing what we have to update uh I already mentioned in the workflow as well can you same process on the Azure also see a is just a it's a cloud platform whatever deployment I will be doing in the AWS you can also do the same thing in the Azure also but only some of the services will be like different okay otherwise my concept is same at the end so in Azure also you are having a storage service I think Azure blob right if I'm

not wrong Azure blob is there it is similar kinds of um I mean S3 bucket as your blob storage right so in that case actually you have to use a your blob right instead of using S3 bucket and if you're if you're using Google Google Cloud that time you have to use gcp bucket all the cloud are having same Services only but name is different got it so no need to worry about cloud guys just try to learn only one cloud and uh same concept you can apply but see if you're switching to another Cloud only you just need to know about

the cloud like how to use the Azure blob service if I want to connect with the Azure blob service what I need to do okay these kinds of things you to learn only is S3 speed more than trainer model in become then the production model no speed actually speed I mean both will have the same speed now the speed actually we're not considering we're considering the let's say performance of the model okay that is the thing so first of all let's try to update our con guys I think you remember the first thing we usually update the constant okay but before that I think uh

what you need to do guys remember this inside me file I I already written you have to export your mongodb uh I mean this one connection string if I want to execute this code yes or no inside environment variable either you can set inside environment variable if I open my environment variable see I already set it environment variable if I show you see mongodb URL I already set so previously I showed you so if you miss missed that session guys please try to check my previous recording okay previous let's say live session there I already showed you how to set this environment variable and here I'm using

G bash and if you're running from G bash what you have to do you have to export it and this is the command okay inside that you need to pass your mongodb uh connection string and how to get the connection string I also explained this part fine so I already uh see I already exported the connection string if I show you uh see I already exported my connection string this is my connection string I already exported let me export again see I already exported now I'll just try to clear so please try to export your

connection string guys all of you so see this is my mongod Atlas so here this is my project inside that I created one uh this one database and here is my connection okay and here is my data and how to get the connection string if I go back click on uh Connect Now divers select your python version 3.

6 or letterer and this is the connection string just try to copy and try to add it okay so I already explained I'm not going to explain again in S3 okay I already uh answered fine so now first of all add my uh this one uh my constant okay I first of all I'll be adding the constant so let's the add the constant variable so see the first constant variable I be I'll be updating my AWS credential I think I can add it here after this target column I can add it here AWS access key ID AWS secret access key ID and a uh region name now tell me why I need to add my AWS access key ID secret key ID and region

name because I already showed you now here we'll be using AWS service that means we'll be using S3 bucket okay and now if I want to let's say authenticate with this three bucket I need these are the credential and how to get the credential I'll tell you okay no need to worry I'll tell you how you can get the credential as well so all of you are having AWS account here let me know if you don't have the account no need to worry later on you you can also try whenever you will have the account you can see the I

mean um session as of now later on you can try fine okay then we'll be adding some more constant here uh here is the constant and this is related our model evaluation here I can write model evaluation related constant now I think fine see here I'm giving the threshold score I think you remember guys what is threshold score I explained in my Blackboard so as of now I'm taking 0.

2 because here I'm showing you the demo I know that my model won't be working better sometimes so that's why I just kept very little actually uh threshold score 0.2 but whenever you are training your actual model that time what you have to do you have to increase this particular number just try to keep 0.7 okay that means 70% 70% score you'll be expecting from the model but as of now I'll keep 0.

2 only just I just I just wanted to show you how it will push the model to the S3 bucket okay that's why I kept the size and another uh constant I have taken my model bucket name that means My3 bucket name with the help of this uh name actually I'll be creating my bucket and model Pusher S3 key you have to take model resist okay you can take any name I have taken model register okay it's up to you all right so the first thing what we'll be doing guys first of all we'll be set uping our AWS okay AWS let's say server that means we'll be set uping our three

okay three connection for this open up your AWS account all of you see I already logged in with my account this is my account now here make sure you are inside uh this reason Us East one if you're using my code see I already mention I'm inside Us East one okay that means not Virginia but if you're inside any other region let's say Asia Pacific Mumbai you have to give AP South one okay similar wise you can see all the region name so I'm inside Us East one not Virginia I'll be taking us East one okay this is the name even I also mention inside my constant

okay now the second thing you have to search for is three service okay three service you can see the three bucket it's a storage in Cloud okay now I'll open this three okay I'm getting one error S3 service P currently unavailable you can monitor okay guys just a minute can anyone check can anyone check guys whether your S3 is working or not or it is having in my system only okay now see it is working I think okay now it is working fine now the first thing we'll be creating one bucket so let's create a bucket uh first of all give the name of

the bucket so I'll give this name let's say here is the name US Visa model 20124 I'll give the bucket name here then what you have to do you can uncheck mark this option block all the public access I acknowledge then just keep it everything as default and create create the bucket okay so guys you can see my bucket is created okay US Visa model 2024 my bucket has created and inside bucket I have nothing right now okay I think so far everything is clear now the next thing we have to uh actually manage our access key and

secret access key okay okay so for this what I will do again I'll go to the AWS main page and here I will search for I am that means identity access management just click here and click on the user we'll be creating a different user here so click on create user give the name let's say I'll give Visa user you can give any name click on next attached policy directly so here I will give the S3 S3 access only okay because I want to use S3 service that is why it's loading just a minute it's

loading let me create again Visa user okay now I got it now here search for S3 okay S3 now you can see Amazon S3 full access just try to give the full access here okay and apart from that I don't need any other Service as of now so what I will do I'll just click on the uh next now click on the create user now I have created the user now click on the user and just go below you will see one option called security credential now go below you will see access Keys okay now try to create the access key select the common line interface I understand the above

recommendation click on the next okay now I'll create the access key and guys don't share this access key with anyone okay if you're sharing they will also able to access your S3 service okay and you will get charged that time okay no need to share with anyone I'm showing you because after the session I will try to delete my user okay so I'll try to download it first of all if I want to download see there is a button called download CSV file just try to download and now what I have to do I have to save it as environment variable

okay now if you want to save it as environment variable what you have to do guys if you're using G bash if you're using G bash you have to run export command you have to run export command so let me show you the command inside read me I can mention so this two command you have to execute so export AWS access key ID here you have to pass the access key ID and aw secret access key ID here you have to pass the secret access key ID okay now let me show you how we can add so here is my access key I'll copy the access

key and here I'm trying to mention so I I can make a different copy and here I can mention this my access key now I'll copy this command and I will execute from my G bash let me copy again now just try to execute now similar wise I will copy for my secret access key this is my secret access key I'll copy and here I'll be mentioning it and I'll execute okay now those are using environment variable so open up your environment variable uh click on the environment variable go to the system variable and create a new one create a new one here you have to give the variable name that

means uh first of all you have to give the a access key ID okay and here inside that you have to pass the value what is the value access keid this is the value I'll copy inside that I'll just try to paste it and click on okay now see this is set now similar wise you have to do it for the uh this one a uh secret key okay secret key this one also you need to do the same thing as of now I'll just try to remove because I'll be using uh this one my uh uh like G bash so I'll be removing this one so I'll select and

I'll try to delete it as of now fine then I'm closing it now let me remove my keys from here otherwise people will use my one I know that yeah so now let me know guys is it clear how to set the access key and secret access key I hope it is clear see it will cost like very less amount so here we are not going to let's say keep our model running once everything is done we'll try to REM remove all the resources okay raguram so no need to worry it will take very less amount of money only one to two rupees I think fine okay now one thing I need to also set this AWS region so if I show

you uh this one AWS region also region let me just check h no reason I I don't need to say because reason I have added directly now esist one so sorry guys I don't need to set the region region is already set okay fine so my setup is completed now let's write the code see the first thing what I have to do I have to uh create one file here so I think remember inside connection we created our mongodb connection now we'll be creating another file and this would be for the AWS connection so I'll click here I'll create a file and I'll just name it as AWS

uncore uh connection do PI so this uh file will help me to connect with my AWS okay now I already written the code let me show you so here we'll be connecting with the S3 bucket see if you search on internet how to connect with S3 so you'll get these kinds of codes inip it okay so here I'm using the same Cod code as it is only I written as a class and Method okay this is this is what I have done only okay just a minute see guys so here I've have written a class called S3 client inside that I've taken a Constructor and here is the connection okay so whenever you will call this class it will

automatically do the connection with my S3 and for this I'm using boto 3 I think remember inside requirement. txt we install something called boto 3 see so boto 3 is a library python library with the help of that you can connect with your AWS service okay here you can see resources wise I'm using S3 bucket and if I want to connect with the three bucket I have to pass my access key ID security access key ID and region name and from where I'm getting I'm getting from my environment variable you can see get environment variable and here is the key name okay I'm passing the key name

so it load from my environment okay that time because directly I can't give if I'm giving directly people will see my credential and they will use my credential okay I don't want that that's why I'm passing as a environment variable okay I hope it is clear now now if you call this class this uh object would be initialize that means you will get your uh connection string okay that means your S3 connection string that time S3 client you will get fine I hope it's is clear now hm all right then we'll be creating one folder here inside this US Visa uh I'll

be creating one folder I'll name it as uh Cloud stories cloud unor storage so this inside this folder I'll be creating one Constructor file first of all because uh it it is actually local package I'll create this underscore unit uncore dop okay then I'm also going to create one file here and I'll name it as AWS stoes AWS stoes do file so inside this file I'll be uh writing all the AWS functionality like how it it will let's say uh get the model from the S3 how it will push the model from my local to S3 okay all the let's say operation related code I'll be

writing here so let me show you so I already prepared one class for you so this is the class guys I've written again all the functionality you don't need we'll be using some of the functionality okay from here let me show you see uh s see guys uh uh yeah thank you so guys if you feel uh this session is really helpful and if you like my actually effort so what you can do guys you can um uh do the little bit contribution so let me show you what you can do so you can uh buy me a coffee guys everyone if you really like my session so I'm sharing my uh my

profile so here's the profile guys so here you can buy me a coffee for me if you feel like this session is really help helps you a lot so you can buy me a coffee okay so this will give me more motivation even uh you can see lots of people has already did okay if you really like and if you want to appreciate me so you can uh um you can buy me a coffee for me here so again I have shared the link in the chat okay fine so now I'll open it up great okay so

now okay fine so now what I'll do guys uh I'll just try to uh okay explain this code see this is the a storage so here I created a class called simple storage service and here you can see I've imported this configuration S3 client okay so this S3 client we have written now that means this is the connection connection object we'll be getting okay this is the connection I'll be getting so this connection will be using to connect with my S3 bucket and after connecting guys what we can perform we can read the object in this three we can get the bucket name we can

get a file from this three we can load any kinds of model we can create a folder there we can upload a file okay we can uh let's upload a data frame as a CSV that means whatever things you can perform okay manually in the3 bucket so let me open my three again service S3 see whatever things you can create okay with the help of user interface same thing you can also perform with the help of python code right now okay see all the function I have written see all the function you don't need okay all the function you don't need we'll be only

using this these are the function let me show you we'll be using load object upload a file that means if I want to let's say push my model I'll be using this function upload file because I'll be uploading one model and if I want to load any model I'll be using this function load model okay because I told you now first of all it will get the production model from my three and then it will compare with my current model as well as the production model okay for this I need these are the functionality got it so always keep this particular

code as a template guys you can use this code in every project okay this code will not change the way I have written you can use this class as it is in your every project as well fine so this is called actually template code here I created as a template it's not a scripting code you can see it's a complete class I have written whenever you are using S3 service you can use this code as it is fine so I think this one is also done now um H now another thing we be uh creating inside my entity so let's open the

entity inside entity we'll be creating another file and we'll just name it as S3 uncore uh s estimator estimator do PI so what is S3 uh S3 estimator see I told you now what will happen first of all it will uh get the model from my production and you will check whether this model is working fine or not okay whether this model is working fine or not and if you want to do it so that's why I've created a separate class for this you can see usva estimator and here I'm uh importing this simple storage service from where from

my AWS storage okay because inside that I'm having all the functionality that's if I want to load model I can load model also now see here I'm loading the model only so from this one I'm loading the model see load the model that means whatever model I have in my production server first of all I will load the model then what I will do I'll save the model okay then I will do the prediction on top of it that means I will load uh I will uh load my test data on top of that I will do the prediction okay how you are managing your day you have huge knowledge how you started at

uh common cats it's a good question uh see it will uh automatically come so whenever you will be more consistent he will be doing continuous learning he will be doing let's say continuous development then definitely uh you can uh gain these kinds of let's say skill and knowledge fine it's not a one day one day actually effort okay every day we have to do the same thing and you have to be let's say motivated with always your work let's say uh if you're learning something try to complete it no need to give any Gap try

to be consistent try to develop something okay build different different things then you will automatically see that uh you can also manage all of everything okay all together okay fine so now uh I think huh so I think you got it so why we created this S3 estimator because here will be getting the model from my production server and here we'll be doing the prediction on top of it okay we'll just try to check whether this model is fine or not uh we'll be comparing our current model as well as my production model and if one of the model is working fine that means my current model is working fine

I'll push it otherwise I be pushing it okay this is the simple idea that means whatever things I explained here now I'm just implementing with the code that's it all right now let me save now I think everything is fine Nows connection I can uh close it I can also close it I I also don't need one I can close it I can also close it okay that means everything is fine okay so now the next thing what we'll do guys yeah I think my AWS connection is ready uh now we can start writing our component okay but before writing the component I think you remember we used to update uh two file actually entity

that means my config entity as well as the artifact entity so both file both file I have to update so let's update this file as well so here I will update first of all my config entity with my model evaluation OKAY model evaluation config so it will basically return all the path okay one by one let's say this is my changeable threshold and from where I'm getting the threshold I'm getting from my constant itself okay then bucket name this is the bucket name okay then let's say three model keyth this the three model keypath okay that means this model

would be updated model. pel okay I hope it is clear now similar wise I'll update my artifact entity so let's update my artifact entity as well so artifact wise this is my model evaluation artifact that means after running the pipeline what it will return it will return these are the thing is model accepted or not change model accuracy S3 uh model path as well as the train model path now let me just show you see open my piic pen see what is e model accepted this is the param sorry just a minute h is model accepted this is the parameter I think remember if my current

model current model is better so it would be true okay otherwise it would be false okay change accuracy means let's say you calculated your let's say current model current model accuracy current accuracy you also calculated the production accuracy production model accuracy okay let's say here you got 60% here you got let's say 55% now changed accuracy means let's say you will be substract current uh substraction of production okay then what what you will get 60+ 60 minus 55 so here you will get sorry it should be 55 here you will get five that means

changed accuracy is 5% only okay this is what I mean I just wanted to see how much actually accuracy drop is there that is why I'm calculating this I hope it is clear now guys thank you Aris thank you it's really motivate me a lot thank you it is clear guys what is changed accuracy that means here I'm just calculating my current accuracy as well as the production accuracy then I'm doing the substraction I want to see the changed accuracy that means how much accuracy dropped okay between my current model as well as the production model that is what I just wanted to calculate here I hope it is clear now let me clear

my screen and S3 model path that means uh see what is S3 model S3 model path means let's say whenever it will push the model to my S3 it will give me the S3 model path that means this is the I mean my model. Pi now that name actually it return and trade TR model path means my this model uh inside artifact inside model trainer this is the path okay so both path actually it will return because I need for my model Pusher because if I want to push the model which model I'll push I'll push my train model now and how we get the path here I'm I'll be rning the path inside this folder I'm

having the model OKAY simple got it so this is how I have prepared my config entity as well as the artifact entity now let me close them now finally I can create my component so let's open the component right now inside component folder I'm having uh model evaluation okay now I already prepared the model evaluation let me show you so here is the model Evolution guys so here I'm combining each and everything all together and I'm creating the final class or model evaluation see I've imported all the library necessary

I need model evalution config trainer artifact inje artifact model evaluation artifact why I need the injection artifact can anybody tell me because I think you know that inside my model uh sorry data in I have my ingested data that means t.

CSV yes or no and if I want to get the path from where I can get I can get from my data in the artifact okay because it's a return I if I open the flow chart I think you remember if I open the data in see after investing the data it will return two path my Trend csb path as well as the test csb path okay okay I I think you already know previously I already explained so that is why I've imported data in this artifact then some matrixes I imported like F1 is code then you can see USB exception login okay these are the thing I have imported one by one and again Target value mapping whatever things see we perform inside our train

data now inside train data we perform some feature engineering task let's say we dropped some of the column we did the target value mapping the same thing we also need to do inside my test data because test data is also the same it's a raw data now that's why before giving to the model I also need to perform the same thing and if you missed out that session guys please try to check because in my YouTube channel I already covered I think in this session in the data transformation um this particular session I already talked about okay how we can do the different different feature engineering okay yes and one class I'm

creating called evaluation model response that means it will give me train model F1 is score best model F1 is score is model accepted difference between that means I want to capture the F1 score okay here okay F1 score that means here we are not considering the accuracy score instead of that we will be considering the F1 score okay F1 score Matrix we'll be using for my final uh let's say uh I mean decision okay this is the main idea and is model accepted this is the main parameter if this parameter is true that means we'll

be pushing the model otherwise we won't be pushing the model now this is the final class model evaluation here I'm initializing everything first of all I'm getting my best model from where from my S3 bucket you can see US Visa estimator okay this is the class we written inside us estimator I think you remember inside that we are loading the model saving it as well as the we predicting it got it then we are evaluating the model which model my current model my train model then here we are comparing the accuracy you can see here we are comparing the

accuracy somewhere I dided the comparation here if best model is not none that means if my best best model that means best model is this one if my uh let's say production model is empty let's say you are running for the first time if it is empty that time what will happen this will give me true that means I don't need to execute my model Pusher so d uh sorry uh I have to run my model pushup that means my is model accepted parameter would be true okay that means I will directly push my current model otherwise what I'll do I'll just try to check the accuracy you can see if my best model is scor okay if it is more

than that what I'm doing I'm just storing everything okay and returning the results that's it fine now let's go to the initiate model trainer so here you can see I'm uh calling my model Evolution artifact inside that I'm passing everything and I'm calling my this one evalate model okay this function evalate model that means whatever think I explained here it will first of all compare the model if my current model is better this EAS accept model would be true otherwise it would be false okay and will return me all the matrixes that is the simple idea and it will return me the model path as

well fine now let's test it whether it's working fine or not for this I'll open up my uh this one uh train Pipeline and first thing I have to import my um model evil version Let Me import just a minute uh model evaluation okay then I have to import my model uh model Evolution config then I have to import my model evaluation artifact okay then I have to initialize it here my model evaluation config inside my Constructor then I'll be creating a function called start model evaluation

after model trainer the same concept previously whatever I explained the same concept only the function name is changing and here I'm calling my main function which is nothing but initiate model evaluation and it is coming from my component this is the initiate model trainer uh initiate model Evolution that means if you call it all the execution would be happening right so that is what actually we're doing let me open my uh just a minute onig I don't need training pipeline okay now I have to call it call it inside my run pipeline but how I will be calling inside run

pipeline guys uh as of now I'll be directly calling model Pusher would be a little bit different I'll tell you so after model trainer I'll call my model evaluation fine I hope it is clear so far okay now similar wise I will also add my model Pusher because I also need to add the model Pusher because model Pusher is dependent on my model evaluation if my evaluation resp response is true it will start the pushing otherwise it won't be start the pushing so let's add the model Pusher as well it's very simple again I'll open up

my uh constant and I will add some parameter here let me see whether I have to add the constant parameter or not oh yeah so model Pusher I don't need any constant so I'll directly add the entity so I'll open the artifact as well as the config so I'll add the model Pusher config so here it will only return the model bucket name as well as the model file name that means my uh model.

pel okay that file name I think remember model. pel as well as the this one model bucket name what is bucket name bucket name this is the bucket US Visa Model 24 that means inside this bucket it will push the model okay this bucket it will push the model that is the idea now initially there is empty there is nothing now let me save now I can close it now I also need to set the artifact model push artif it will return to path actually my bucket name as well as the S3 model file path okay you can also add you can also remove it's up to you but everywhere I'm adding my uh this one artifact that's

why I'm adding it okay otherwise you can also um delete it okay it's up to you if you want to see the the bucket name if you want to see the SD model path that time you can see if you don't want to see you can delete it okay it's up to you again it's your design pipeline yes s definitely we'll be doing no need to worry okay all the project we'll see as an end to end now everything is set now I will add the component so model Pusher now let's add the model Pusher component so simply it will uh push the

model that means upload the file to the three bucket see here I'm using the same simple stories and from here I will be using something called uh if I show you model Pusher initiate save model okay this function I'm using that means it will upload this file my model trainer the this Pi file to the production that means to my three bucket that is what actually it is doing now you can see that is what actually it is doing simple code only I just written as a modular way that's why you can see so many line

but but if you go through it you will able to understand okay what I have written because uh I mean uh from the very scratch we are writing this kinds of code only okay and I already told you what is the structure of my code if you missed out the previous session it would be a little bit difficult for you and if you complete them it would be easy for you fine so it will push the model now I'll call it inside my pipeline so training pipeline first of all I have to import them first of all I will import my model Pusher then I'll import my model Pusher

config sorry model Pusher config why it's not coming okay sorry I have to give the comma here now fine now I have my model Pusher artifact now I have to initialize it here model Pusher config okay so now what I will do I'll write a function here called Model Pusher start model Pusher after evaluation okay so it will call my initiate model Pusher that means this function and my model Pusher will start now I have to install inside my run pipeline now see uh one small logic you have to write here let me show you

inside um run pipeline first of all you will check uh you will just check one if condition see that's how first of all you will check if my model evation artifact this parameter is model accepted if it is not not true okay if it is not true you will raise one uh let's say message model not accepted that means my current model is not working good because it's not accepted you can see it's false otherwise it will execute the model Pusher in else condition simple code you can see start model Pusher will execute fine this is the small ail

condition you have to apply now everything is fine now let let's test it whether it is working or not now let me show you initially in the S3 bucket I don't have anything it's completely empty if I refresh see it's completely empty now see if I execute my code right now I already set all of my environment variable now I'll simply execute my demo.

Pi so python demo. Pi I hope it should work okay let's test guys okay I'll execute see first of all it will uh inest the data from my mongod DV then it will do the validation then transformation model trainer evaluation pusher still running let's wait e

in between if you have any question guys you can ask me I'll try to answer and let me show you uh the flowchart as well see whatever things I have explained I also mention as a flowchart see the same thing config get best model S3 model OKAY test data then we'll be just checking our best model as as well as the trend model okay then we'll comparing based on that we'll try to decide whether I have to push or I have to keep it okay this is the flow chart you can see and this is the model pushup flow chart that means if my uh this one uh this um uh this statement is

true that means my model is accepted I'll push the model you can see initiate the model Pusher save the model and we'll just push the model to the um S3 bucket okay this is the flowchart I have created but again if you feel like this is a complex flowchart what you can do you can also create your own flowchart okay I already showed you one website with the help of that you can also create your own flowchart let me close it right now now let me see still it is running okay let's wait I think my internet connection is little bit slow that's why it's taking time e

any question guys you can ask me okay now see my training has started uh I think this is the last yeah so now you can see this is the last run it is running now see training is going on that means it is doing the hyper parameter tuning right now cross validation we are using neurom MF Library okay execution is done let me check guys I got my model now if I go to my S3 bucket if I go to my S3 bucket right now m what is my 3 S3 this is my S3 I'll go to the S3 okay here is my S3 now if I refresh now can you see the model guys all of you can you see the

model yes or no see beautifully it has pushed my model to my production server okay model. pel and this is my final model this is my production model okay so whenever I'll will be creating my prediction pipeline this model I'll try to use for the prediction so so far whatever things we have done to achieve this model actually let me know so far everything is clear yes yes or no I hope it is clear okay now let's uh see how we can create the prediction pipeline now see everything is working fine now as of now

we have completed our training pipeline because uh we are able to train now we need also prediction pipeline because user will pass the data and we also need to do the prediction now for prediction pipeline first of all you have to update this one let me show you you have to update actually this uh constant inside constant actually you have to mention two parameter apphost and apport because here we'll be implementing web application now with the help of P API so this is my apphost and apport number 8080 you can give any kinds of number

it's up to you now let me save and I will also add my config let me see so here it will give me the model model file path and model bucket name okay because uh we'll be using that S3 model so that's why I need these are the uh file uh I I need deser the name okay this is the idea now I also need the artifact entity arct entity I don't need because here we're doing the prediction only so simply I'll go to the here prediction Pipeline and here I already created the prediction pipeline see this is my prediction pipeline so here I'm importing my all of the required package USB is a prediction config okay

USB is estimator and here is the estimator guys you will get so here you can predict it that means it will load the this model okay and you can predict with the help of this particular function predict function okay that is what actually we are doing now here we'll be taking the input uh first of all from the user fine so here first of all we'll be taking the input from the user and tell me how many input user will give continent education of employee then has job experience required job training number of employee region of uh let's

say employment privilege W unit of w then fulltime position company is so these are the input actually I will take from the user and based on that I'll just tell them whether they are able to uh let's say uh accept it for the Visa my model will try to predict okay because this is my independent let's say uh I mean features based on that it will give me the label okay now see here we're restoring everything then here we are getting the see here what we are doing actually we are getting the Visa data as a dictionary I can see here we

are converting everything as a dictionary why because I want to convert them to the data frame and if I want to convert anything to the data frame the first thing what I have to do I have to convert them to the dictionary and you can easily convert dictionary to the data frame okay that is what actually we are doing Simple function guys if you check it you will understand okay what I'm doing I'm just calling one function inside another one that is the idea now this is my us a classifier it will do the prediction now see here we

doing the prediction I taking my model and model is doing the prediction and here is the result and result I'm just returning that's it that means I'm loading my S3 bucket model I'm predicting and returning the results that's it now with the help of this function I'll be creating my user application for this uh what I will do I'll create a file here called app. Pi see app.

Pi is already created now let me delete this line is not required and two more uh for folder I need here guys if you're using fast API you need something called this one template let me create template because inside that you'll be keeping the HTML related code templates I think let me check whether it's a templates yeah templates and inside templates you will be creating another file call HTML file US Visa do HTML now inside that I'll will be mentioning my HTML code now again no need to worry about the HTML code guys because you can uh you can take any kinds of HTML code okay here so it's not

your uh I mean responsibility you can take any kinds of HTML code you can take help from the front end developer they will create this particular form for you but what I've have done I've just created a basic form here that is the idea okay okay fine so now this is my HTML then I also need another actually folder called Static if you have little bit knowledge about actually U this one flask I think you know why we are using this file static inside that we'll be creating another file called CSS folder CSS

inside that I'll be keeping my CSS related code so inside CSS I'll be creating a file I'll just name it as style dot CSS okay now inside that I'll mention the CSS code that's it as of now this is my Basics website I've have created okay basic let's say web application I created now let me write the app. pyod see so here we are using fast API so we have imported all the fast API related package then I'm importing my apphost and apport from my constant then I'm importing US Visa data and classifier then I'm importing training Pipeline and this is the classifier that means this is my prediction pip plan

okay this is the classifier it will give you the prediction that time okay this is the idea now here you have to initialize the first API like that so first of all initialize the first API get the app object with the help of app object just try to create the template like that so whenever you'll execute the first thing what you will do it will look for the template folder inside template folder whatever files you are having that HTML okay this HTML would be rendered okay this is your landing page now here I'll be taking the data from the user and here I'm taking the data

from the form you can see so from form itself I'm getting all the data so user will pass the data and I will collect the data all of the data I'll collect okay then this is my main route that means this is my landing page it will render my us visa.

HTML page and this is the training route that means if you give SL train it will start the training and if you give the SL predict it will start the prediction that time okay two route three route I kept one is the default route one is the training route and other one is the prediction route okay and finally I'm running my application that's it now let me show you if I run python app.

py now see my application will run now you have to allow the access Now open up your Google Chrome and search for local host port number 8080 now see guys this is your user interface okay this is your user interface this is a simple form I have created this is the simple form I have created guys okay nothing else so that user can pass any kinds of data now let's say I'll select the first data let's say continent is equal to I'll give Asia education let's say uh doctorate job experience yes required job training no region let's say I'll give northest privil voice let's say I will give select between 700 to

7,000 I can give let's say um 700 I can give as of now contact trainer let's say m month part-time yes age of the company let's say 11 now if I predict okay if I predict this data now let's see whether this Visa would be approved or not okay I have to give this number okay number of employee let's say let me see the count okay 15,000 I can give now see first of all it will get the model from my S3 bucket okay see it will download the model from the S3 bucket then it will do the prediction so let's wait

it will create a folder here called uh predicted model inside that actually it will uh download the model and it will do the prediction so my internet connection is little bit slow guys that's why it's taking time e let me know if you have any question guys

okay so it is downloading the model still I think to it is taking that much of time e e I don't know what happened uh I think uh my internet is not

reachable anyone just try and let me know how can fresher get into mlops field first of all try to learn um like some mlops technology uh let's say whatever you need try to use it in your project implementation okay that's how whenever let's see you will be getting familiar with mlops you can understand let's say mlops architecture and all that time you can apply okay that's how we can enter into the MLS position mlops is nothing but a tool only okay you have to use them as a tool that's it I don't know I think my server got hang some of the website actually I'm

not able to open today with my internet I think that's why I don't know but it will work I already tested okay I already tested it will work it will give you the prediction maybe it is not able to um download this model because of my internet connection it is blocking I think maybe let me stop and try again so can I can execute it again ab.

P running it I can refresh e yeah so because of the internet actually it got hang somewhere but you can try I know it will work tomorrow I'll show you tomorrow I'll show you actually it will work I have to restart it then I think it will work fine so that's how actually you can uh implement the entire projects and this is what actually uh you have to follow okay whenever you are creating any kind of pipeline okay now tomorrow we'll just try to deploy this project to the AWS Cloud we'll be following cicd approach okay to deploy this projects

this is the idea so let me know if you have any query guys let me know any query if you have any query you can let me know otherwise uh okay let me push the code as well so prediction and app addit I'm done now if I go to the GitHub now see everything is up to date now you can clone and you can try okay in your system guys all of you okay I'm trying to stop it but it's not it got hang actually my terminal got hang okay so tomorrow I'll show show you okay it will work I already tested I don't know some internet issue from my

side today so yes guys this is all about uh today's session I hope you liked it and if you like my session guys uh if you want to appreciate me so what you can do you can buy me a coffee so here's the link I'm giving all right so yes thanks for joining everyone and please like the session and try to share with your friends and family and if you haven't subscribed try to subscribe to my channel Let's uh connect tomorrow uh I think tomorrow is uh Saturday sorry tomorrow is Friday I think okay uh Friday uh we'll see if I'm available Friday definitely I will conduct the

session otherwise I will reschedule on the Saturday or Sunday fine so with that guys thank you so much uh for joining the session bye everyone good night take care bye thank you

Show thinking

This comprehensive session focused on building the final components of the ML pipeline: **Model Evaluation** and **Model Pusher**. It also covered creating the **Prediction Pipeline** and a **FastAPI web application** to serve the model to end-users.

**📝 Today's Agenda: Finalizing the Pipeline & Building the App**

* **Model Evaluation:** Creating a component to check if the newly trained model is better than the one currently in production.
* **Model Pusher:** Building a component to "push" an accepted model to a production storage location.
* **Prediction Pipeline:** Designing a separate pipeline specifically for making real-time predictions.
* **FastAPI Application:** Developing a user-facing web application with a form to interact with the model.

**✅ Model Evaluation: The Quality Gate**

This component acts as a critical quality check to decide whether a new model is worthy of being deployed.

**Workflow**

The evaluation logic is designed to be robust, especially for retraining scenarios:

1. **Check for a Production Model:** The component first checks an **AWS S3 bucket** to see if a production model already exists. S3 is a cloud storage service from AWS used here to store the production-ready model.
2. **First-Time Run:** If no model is found in S3 (i.e., the first time the pipeline is run), the newly trained model is automatically considered **"accepted."**
3. **Retraining Run:** If a production model *does* exist in S3, the pipeline performs a "challenger vs. champion" comparison:
   * It downloads the existing "champion" model from S3.
   * It evaluates *both* the new "challenger" model and the old "champion" model on the same test dataset.
   * The new model is only **"accepted"** if its performance (e.g., F1-score) is better than the existing model's performance.
4. **Output (Artifact):** The component's main output is a boolean flag, is\_model\_accepted, which will be True or False.

**🚀 Model Pusher: Deploying to Production Storage**

The Model Pusher component is straightforward and is controlled by the output of the Model Evaluation step.

* **Trigger:** This component only runs if is\_model\_accepted is True.
* **Action:** It takes the newly trained model.pkl file from the local artifacts directory and uploads it to the designated **AWS S3 bucket**, replacing the old model. This action effectively promotes the new model to production.

**🤖 Prediction Pipeline & FastAPI Application**

With the training pipeline complete, the focus shifted to creating a way for users to interact with the model.

* **Prediction Pipeline:** A separate pipeline was created specifically for inference. It takes raw user input (e.g., from a web form), loads the production model and preprocessor from S3, transforms the input data, and returns a prediction.
* **FastAPI Web App:** A web application was built using **FastAPI**, a modern Python web framework.
  + **app.py:** The main file containing the server logic.
  + **User Interface:** A simple HTML form was created for users to enter data for a visa application.
  + **API Endpoints:** Three routes were defined:
    1. /: The main page that displays the prediction form.
    2. /train: An endpoint to trigger a full run of the training pipeline.
    3. /predict: An endpoint that takes the user's form data, passes it through the **prediction pipeline**, and displays the result (e.g., "Visa Approved" or "Visa Denied").

The presenter ran the full pipeline, which successfully trained, evaluated, and pushed the model to the S3 bucket. He then launched the FastAPI application to demonstrate the user interface.

**🔜 Next Steps**

The final session will focus on deploying the entire FastAPI application to the cloud using a **CI/CD (Continuous Integration/Continuous Deployment)** approach.

 e e uh hello everyone am I audible to all of you can I please confirm in the

chat good evening everyone hello everyone to all of you I think I'm aible guys let me know okay fine so now let's start with the session guys so uh we are continuing with our mlos production ready machine learning project series and if you haven't check out my previous session so guys here you can check this is the complete playlist so from day one we have completed till day eight that means we have learned how we can do the model evaluation model Pusher as well as the prediction pipeline that means we uh

completely created our application now uh in this video actually we'll be trying to do the deployment okay uh we'll be doing awcd deployment here fine so make sure you have the AWS account with you so here we'll be using AWS and uh in my next project implementation I will also show you how we can deploy in any other Cloud as well okay so all the cloud will be exploring one by one okay so now let's start with the session guys uh let me open up my uh Blackboard and let me tell you about uh this cicd okay uh I'll just uh give you

the brief introduction of the cicd like what exactly cicd and how we'll be deing our project okay so can you please tell me what is the full form of cicd guys cicd anyone yeah hi Aris I'm doing great what about you see the full form of cicd is continuous integation and continuous delivery or you can also call deployment okay H just a minute

H yeah continuous integration and continuous deployment yeah this is the last lecture for this project because uh here we'll be doing the final deployment and I will keep uh some of the session like uh I think one session I will keep U for the let's say resume preparation as well as the um like job preparation okay how we can attend the interview and how we can discuss the projects and all okay this thing we'll be discussing about there even we'll be also learning about uh how we can uh let's say create our resume okay if I if we are a fresher if we are a experienced person that time

how we can prepare the resume we'll be explored exploring this part as well now see here uh what happens let's say as a developer let's say you are a developer as a developer so where you write your code I mean in which environment you write your code I think you know inside development environment Dev environment right yes or no that means you use your local machine okay local machine to implement the entire projects then as of now what we are doing we are pushing this code where guys to the GitHub okay GitHub we are pushing all of

our source code fine so now see so far we are able to uh develop our projects and commit to GitHub okay now what we have to do we have to create a complete cicd pipeline so that whenever as a developer I'm committing my code to the GitHub so it should automatically get updated okay okay otherwise what will happen let's say you have to manually update your application okay if you're doing manual update so what you will do let's say you are updating your code to the AWS okay AWS Cloud that means you deploying to

the AWS Cloud so manually what you have to do you have to let's say this is your code okay this is your code manually you have to first of all do the setup okay do the setup of the environment then you have to create the environment okay everything you have to do manually then you have to manually upload this code to the AWS okay manually upload this codee to the AWS then your application would be running but let's say after 6 month after 6 month you want to again add some new features in your application so what you have to do again you have to stop the application okay stop the application it

is running on the cloud then you'll be updating your code again then again you will start the machine okay again you will start the machine then again your application will be keep on running but whenever you did this stop operation now whenever you did the stop operation that times your audience okay audience won't be able to audience won be able to access your application okay so this is an issue actually this is an issue so how we can solve this issue because if you see any of the software any of the software let's say YouTube Facebook whatever software if you see whatever they're

updating let's say day today all the softwares is getting updated but it's not like that their server is down it's never like that okay server is running but instead will see the update okay how because they following cicd approach now in cicd approach what will happen let's say whenever uh you are pushing the code to the GitHub so GitHub will automatically connect to the AWS Cloud okay AWS cloud and whatever change actually you'll be making here whatever change you'll be making in your code it will automatically get updated here okay

updated here that me this application is running this application is running I don't need to stop this application it is running on top of that I'm updating my code and it is taking the new features and it is showing in front of the user that's why I don't need to stop my machine here if I'm following this cicd approach okay cicd approach continuous integration and continuous delivery or deployment now I think you got the meaning okay what is continuous integration and continuous delivery or deployment okay and this is the manual approach that means here you are uh uploading your code manually and

whenever you want to add some new features so first first of all you have to stop the application then you will be updating the code then again you will restart that particular application okay but this is not a uh actually efficient way you can manage your deployment okay so that's why we always follow this cicd approach I hope it is clear now fine okay now there are some cicd tool you have to use see if I want to automate this task let's say whenever I'm updating my code in my GitHub so there should be some tool there should be some tool so this tool will make the task automated let's say this tool will

automatically connect with my let's say AWS it will automatically let's say pull my code from my GitHub it will automatically push my code to my let's say cloud so these kinds of tool actually we need so this is called actually cicd tool so there are some cicd tools are available the first one very popular GitHub action and this is like mostly used nowadays why because it is having already set up it is having all the setup integrated okay you don't need to do manual setup it is already pre-built with the GitHub then there is another one called

genkins okay if you want to use genkins you have to manually set up everything manually set up Jenkin server okay then there is another one called Circle CI Circle CI so this one also you have to manually set up manually set up okay and this one one is automated Auto one fine that's why this one is more popular because here as a developer you don't need to set up it it is already integrated with GitHub only you just need to write one EML file okay EML file and this file will take care everything okay and genkins we have to create a

separate server Circle C also we have to use one Circle C website okay there actually we can create the entire pipeline okay so we'll be also exploring this two tool but uh in this particular deployment we'll be learning about GitHub action okay then one by one we'll be exploring each and everything okay this is my agenda got it so here the deployment will be doing called AWS GitHub action GitHub action cicd okay deployment deployment okay this is the idea that means if you upload your code to the GitHub GitHub will automatically

update to the AWS Cloud okay and user will get the endpoint endpoint this is the complete idea now you can ask me what are the tools actually I'm going to use see I'm going to use some of the services from the AWS inside AWS you are having lots of services but I don't need all the services the first thing I need uh which is nothing but uh AWS that means Amazon Amazon ec2 okay ec2 and what is ec2 ec2 is a virtual machine it's it's a virtual machine it's like a computer okay so here you will get CPU GPU Ram okay Rome everything you will get because at the end I have to run my application in a machine okay and

in the cloud I can take a machine and this is called actually ec2 okay E2 service from the Amazon now the next thing I need something called Amazon ECR that means elastic container regist so here we'll be storing our Docker image okay Docker image that means uh with the help of Docker we'll be dockerizing our entire source code and we'll be building an image and the this image actually I'll be storing in the ECR elastic container register so this is the service we'll be using so here you can save any kinds of Docker image

okay in this service fine then we'll be using GitHub action to create our self hosted draw self hosted Runner means uh we'll be connecting with our uh let's say AWS and whatever changes actually will be updating in the GitHub it will automatically push to the AWS Cloud okay so these are the services we'll be using one by month this is the idea okay so now uh let's go to our AWS first of all because your application is ready so this is the application guys this is the uh GitHub link I'm sharing again so here is the GitHub and if you

see uh in the rme file I already updated the deployment uh let's see St okay so these are the step actually will be following first of all we have to loging with our AWS console create IM am user uh so we here we'll be uh giving two policy AWS ec2 container and uh Amazon e to full access that means ECR Service as well as the ec2 service okay we have to give now we'll be creating a ECR repository then create a E2 machine open ec2 and install doart there then we'll be configurating ec2 as a self hosted Runner then we'll

be setting some GitHub Secrets okay then finally we'll be doing the deployment and here is the process guys so deployment is description first of all we'll be building the docker image of our entire source code push your Docker image to the ECR launch your is2 machine pull your image uh from ECR in ec2 launch your Docker image in the ec2 okay then finally we'll be executing our application there so that user will get the endpoint okay so for this we have to do some Port mapping operation fine I hope it is clear okay so you can refresh and you can see the

deployment step because whatever things I'm going to do right now I'll be following this re file only okay so it's better if I can open a new one because this one I'll be referring for my deployment and this one I'll be referring for my documentation okay that is why now everyone just try to login up with your uh AWS console yeah definitely there would be a charge okay nothing is free in the cloud but if you're uh let's say opening this account for the first time if you're adding your card you will get $300 credit okay and you can use those credit

actually free credit it is called got it what is Runner responsible to ingate charge uh changes can you explain a little bit more about it Runner responsible for the intrig changes I didn't get your question e

yeah S3 bucket we created uh we stored our model I think you remember okay sorry so now let's uh huh so so I'm inside my AWS console guys now the first thing we have to create IM am user so let me click on IM am okay now let me click here I think it is loading

just a minute I don't know why it is not opening just a minute it's loading okay now I think it's working okay now let me click on the I am

okay so here I'll click on the user now let me create a user here now let me give the name let's say Visa I can give Visa you can give any name Visa user I can give now let me add some policy now policy wise I already mentioned here first of all you have to add this policy for the ECR sear here and select it and click on next but before clicking on next I'll just add another one now if I click on next you will see both policy have added now let me create the user now inside Visa user uh just a minute

guys e e okay fine so now uh click on security credential and uh create access key

command line utility I understand and next the same way we created the user previously I'm following the same way now this is the accs secret key now let's download the CSV file and just keep it with okay later later on actually I need it now let me get back now the second thing what I have to do guys here I have added the second thing uh so the next thing I have to uh create a ECR repository to save my doad image there so now let's do it so here just search for ECR it's loading just a minute

so ECR so elastic container is St I'll open it up and create a new repository and you can give the name so let's give the name so I'll give let's say Visa then uh everything I'll just keep it default let me check now let me create the repository now just copy the URI and keep it somewh let's say I'll keep here as of now this is my URI I'll just try to replace because later on you need it okay that's why I just save it here fine okay now we have successfully created the uh this one uh ECR repository now I need to create a E2 machine okay so let's go

back then I'll will search for E2 so I'll click on launch instance so here you can give the name let's say Visa machine I can give anything then select uh your operating system I'll will be using ubon to here now I'll keep it everything as default now you need to select the instance type like uh what would be the machine configuration so at least just

try to take uh 8 GB memory okay 8GB Ram so T2 large be taking and no no need to worry after deployment I'll try to remove the instance okay now you can create a key value here let's say I can give any name so it will create a p file later on if you want to let's say connect this machine with the uh third party let's say tool like mobile stream then puty so you can use this U Pam pile that time fine now I'll just select this two option allow HTTP and https now everything is fine now see configuration stories at least take

let's say 30 GB and just keep everything as default now laun instance now see status is pending so keep on refreshing once it is running then uh we'll we'll be making the setup here fine now see it is running now let's click on the instance ID now let's connect the instance so I'll connect so it will give me one uh like uh blank terminal there I have to make all the setup okay now let me clear now let me Zoom a little bit now I think it is visible now

let's uh make the set up so see to do the setup I written all the command you need to install so let's copy one by one and let me just execute there so first of all you have to do the Pudo update that means your package manager you need to do the update and all okay now next command now let's give the permission yes and guys if you don't know I start started one generative actually playlist

so this is the second channel so here you can follow actually generate VI see uh if you go to the playlist so here is the generate VI complete generate VI um so here we'll be covering each and everything you need to master about generi so you can also subscribe this channel okay so here also you will get uh the content related generate python okay even uh intern project implementation everything you will get so make sure guys you also hit the Subscribe in this channel as well so I'm sharing the link in the

chat okay I think it is done now uh I'll just simply install the dock ad e now let me execute this command as well now if you want to test whether your Docker is running or not you can execute this command Docker hypen hypen version now see this is the version that means Docker has installed successfully fine now the next thing what I have to

do I have to configure my2 as my self hosted runner for this just go to your GitHub make sure you are using your GitHub okay not my GitHub you published your code in your GitHub then just go to the settings and here you'll get one option col actions go to the runner uh new self hosted Runner let's take INX machine and copy the command one by one and execute it here that means I'm making the connection with my gab right now okay so now the last two command I have to

run now see it is connected with my GitHub now it is asking for the runner group to add the name enter the another name Runner group I don't need to give any name hit enter now it is asking enter the r of the N I'll give self typen posted okay self hypen hosted uh so you need to give the same name don't give any other name otherwise it w be working fine now it is asking to the label I don't need to give any label just simply press enter now see uh folder again press enter now see it's done now I'll execute the last

command now see it will connect with my GitHub uh see listening for the jobs that means it is connected to my GitHub right now that means whatever changes I'll be updating in my GitHub okay everything it will automatically pull from there but at the last what I have to do I have to set the GitHub secret as well so let's add the secret so I'll come back here I'll go to the secrets and variable click on the actions now here you can select uh repository secret new repository secret now give the key name so this is my key name a this access key

ID so here let me add and why will get the access key ID guys I think you remember we downloaded one CSV file let me open it up so this one so this is my uh access key I think it was access key right let me check I access key ID I'll copy and here I'll paste it then the next one I have to add the secret key secret access key and this is the secret access key then the next thing AWS default region so default region is you can see I'm inside North Virginia that means Us East

one so here I can give us East one so let me just write Us typen East typen one okay now let me add and ECR repo but before that let me uh add one file here called GitHub action okay so let me just add I'll create a folder here called GitHub in inside that I'll be creating another folder workf flows and said that I'm going to create a file called AWS do yaml and I'm going to add all the yl command

okay so these are the EML will help to do the automatic cicd okay so here I have written some of the command even you will also get this EML file from the Google itself if you just search AWS cicd with GitHub action you will get this kinds of EML okay this is the idea now I'll be creating another file called Docker file Docker file and inside that I think Docker file it is already present uh yes so inside that I'm going to write my Docker command so this is the docker command first of all it will uh like create a 3.8 python then inside that it will

create app folder and it will copy all the source code then it will install my requirements then it will run my app. by okay this is a simple command I have written now let me update everything now let me see the secret right now ax secret key uh AWS access key ID access key ID AWS access key ID AWS Secret key then a default Region C ACR repo default region so two more secret I have to add ECR repo and mongodb URL okay I think you remember mongodb URL we had uh so from mongod actually we um like um investing our data okay so for this we have to also add it mm ECR repo so fine so now let me open

it here so here is my ECR repo so what I can do I can create another new secret and I can add ECR repo and this is myo then my mongod as well so I already have my let me just show you this is My mongodb URL

fine now I think everything is set now we can go back and we can comit our changes right now um just a minute okay it is giving one error allow personal access token to create the workflow without workflows scope okay so

I think just a minute guys do GitHub workflows A.L it is fine I think but why it's not working um

just a minute R poster remote trying to pull the first to ingate your changes e I create a new personal access token

it was working but I don't know why is giving me the no I don't need to pull I think they have updated their um policy I don't know just a minute because previously also I updated this uh project okay I know that fix SS list to with additional J login Lo okay sorry guys um I think I got the issue just a

minute what I can do I can go to the settings here you will have the developer settings personal access token token classic then uh generate new token classic let me give my password so repo workflow no expiration now let me create now let me

try again it's not working just a minute [Music] um e just a minute guys h lots of thing they have actually

updated because I haven't checked so far now so okay it's not working so what I can do just a minute settings deploy keys e e

please try to comit all see I'm trying to comit but it's not taking just a minute maybe I need to add this key somewhere e e just a minute I just need

to copy e okay so um it has given me an

error e yeah you will uh get it sort of I'll announce it definitely internet connection is also

slow e just giving me error disconnected maybe my internet just a minute let me download the Z file e okay now I think download it now let me just un and get that dogit

file why is the dotg f okay if you download actually so you won't be able to get the dogit file so you just need to somehow clone it Iden file I activate I activated but it's not showing there e e

yeah downloading actually won't be getting okay it's giving me actually issue because uh my internet connection is slow that's I'm not able to clone it uh see it is almost done so only just need to push the changes so you will see automatically your Runner will run and it will deploy everything to the cloud okay see connection is very slow so what I can do maybe I can uh conduct a um like separate class I think tomorrow or after tomorrow and we can sort out this issue okay because I uh this deployment uh I mean um some month

back actually that time actually it was working fine but I don't know I think they have updated something I have to check okay fine guys so I'll just try to explore and I'll try to fix it then I'll show you this deployment okay again fine so so far actually whatever setup we have done you also need to do the same setup only I just need to add this uh I mean uh workflow okay somehow I'll just try to explore and I'll let you know okay how to do it now fine so now let's end the session uh because I have some other work to do

otherwise I could have solved it okay let's connect uh in the next session and I'll try to I'll show you the solve okay like how we can solve it yeah thank you everyone

Show thinking

This was the final session of the "MLOps Production Ready Machine Learning Project" series, with the primary goal of deploying the completed FastAPI application to the cloud using a **CI/CD pipeline**.

**📝 Today's Agenda: CI/CD Deployment**

* **CI/CD Concepts:** An introduction to what Continuous Integration and Continuous Delivery/Deployment (CI/CD) means and why it's essential for modern software development.
* **Deployment Architecture:** Outlining the plan to use **GitHub Actions** to automatically deploy the application to **Amazon Web Services (AWS)**.
* **Hands-on Setup:** A step-by-step walkthrough of configuring all the necessary services on AWS and GitHub.

**🤔 What is CI/CD?**

CI/CD is a methodology that automates the software development and deployment lifecycle.

* **The Problem with Manual Deployment:** Manually deploying an application often requires stopping the live server, uploading the new code, and restarting it. This process is slow, error-prone, and causes **downtime**, making the application unavailable to users.
* **The CI/CD Solution:** With a CI/CD pipeline, a developer simply pushes their code changes to a GitHub repository. This push automatically triggers a series of actions that build, test, and deploy the new version of the application to the live server **without any downtime**.

The tool chosen for this project was **GitHub Actions** because it's seamlessly integrated into GitHub and doesn't require setting up a separate server like other tools such as Jenkins or CircleCI.

**☁️ The Deployment Architecture**

The deployment plan involved using GitHub Actions to orchestrate a workflow between the code repository and several AWS services.

**Key Services Used**

* **Amazon EC2 (Elastic Compute Cloud):** A virtual machine in the cloud that acts as the server where the application will run.
* **Amazon ECR (Elastic Container Registry):** A private registry to store the application's **Docker** image. A Docker image is a self-contained package that includes everything the application needs to run.

**The Automated Workflow**

1. A developer pushes new code to the GitHub repository.
2. **GitHub Actions** is automatically triggered.
3. The workflow builds a **Docker image** of the application.
4. The new Docker image is pushed to **Amazon ECR** for storage.
5. GitHub Actions securely connects to the **Amazon EC2** instance.
6. On the EC2 instance, the new Docker image is pulled from ECR and run, making the updated application live.

**⚙️ Step-by-Step Setup**

The presenter walked through the entire configuration process:

1. **AWS IAM User:** An AWS user was created with the necessary permissions to access EC2 and ECR.
2. **ECR Repository:** A new repository was created in Amazon ECR to host the Docker image.
3. **EC2 Instance:** A new Ubuntu-based EC2 virtual machine was launched and configured with Docker installed.
4. **GitHub Self-Hosted Runner:** The EC2 instance was configured as a "self-hosted runner" for GitHub Actions. This allows GitHub to securely execute commands directly on the AWS server.
5. **GitHub Secrets:** All sensitive information—such as AWS credentials, the ECR repository name, and the MongoDB URL—was stored as encrypted **secrets** in the GitHub repository settings.
6. **Workflow & Docker Files:** A Dockerfile was created to containerize the application, and a GitHub Actions workflow file (.github/workflows/aws.yaml) was written to define all the deployment steps.

**🚧 The Final Hurdle**

During the session, when attempting to push the final workflow files to GitHub, the presenter encountered an unexpected authentication issue, likely due to a recent change in GitHub's security policies. Due to the difficulty of debugging the issue live, he concluded the session and announced that he would resolve it and cover the final push in a follow-up video.

for e yeah hello guys uh good evening everyone am I audible to all of you let me know in the chat yeah yeah hi everyone Hi H S sidat hi

suant good evening all of you I will wait uh one more minute then I will start the session let everyone uh I mean everyone join yeah all it yeah hi anupam so today uh first of all I will uh show you the last deployment part because I think yesterday I was having some issue so I just rectified I actually got the problem what was the problem because see I switched uh to my I mean new system okay recently because previously I was using my old system there actually I had everything uh I mean set up okay in that system but uh

in this new system actually I mean I haven't done the setup okay yet that's why uh there was a I mean settings actually I have to enable in the GitHub uh but I forgot to add that GitHub okay GitHub setting that's why it's not able to commit the code in my GitHub okay that was the issue yeah yeah I will resolve us issue okay then I will uh start with our today's agenda like resume discussion and carer guidance so please uh like the video guys everyone hit the like and share this video with your friends and family and if you haven't subscribe to my channel just uh do the Subscribe guys uh

let's hit the 14k uh 14k subscriber and if you don't know this is the playlist guys we are continuing uh this is called actually ml production machine learning project okay and today is day n and today is the last session uh of this series okay apart from that if you have any query if you want to connect with me you can connect to my LinkedIn so I'm sharing my link so LinkedIn also you can connect me you can follow me here okay fine okay so let's start with the session so so see I'm not going to discuss the architecture diagram because yesterday I already clarified each and everything what are the service we are

going to use and all so I'm going to quickly actually set up everything and I'm going to show you the deployment okay so first of all what I will do I'll create IM user because I deleted everything yesterday okay whatever I created let's name it visa user attach policy so I already added the policy here I think remember so this two policy have to add I don't know why it takes time actually

just a minute yeah so let me show you the issue actually yesterday I was having see uh if you need to go to your settings okay let me show you see this is the settings your GitHub account settings so there you have to enable one option so just go to the developer settings and here we'll get one option called personal access token token classic so I created my token already this is my token let me give the password so you have to enable this option workflows okay because here I'm

using GitHub action and GitHub action needs this workflows folder I think remember so you have to enable it okay otherwise it w't be working that's why yesterday it was throwing me the error okay I think you got it now let me get back to my repository okay so now I got it now let me set the uh permission then I need another one you see2 apart from that I also need S3 I think remember because if you want to do the model training definitely you need the S3 because in S3 actually your model will get updated so let me also give the S3 Services now I'll click on

next now see three permission I have given okay now simply create the user now I'll click on my user security credential access Keys command line uh then I'll generate it let me open it up so this is my Visa user fine so now next thing I have to create my uh ECR repository so let me create this repository quickly so you can see elastic container St then I'm going to create a new repo I'll name it as let's say Visa

repo uh everything keep it default and create it now just copy the URI and update here then the next thing I have to launch my ec2 machine so let's launch E2 instance I launch instance give the name let's say Visa machine anything you can give UB to I'll select and I'll at least take 8GB memory okay this one key value PA I think yesterday I created one I'll check mark this two option and configuration stor at least take

30gb everything keep it default and launch instance okay now it is running I'll go inside then I'll connect the machine then you have to set up uh these are the thing okay you have to execute desert the command one by one okay so let me clear now let me

p e e yeah hi Anand hi okay done now let me set up the

docker e okay so completed now let me check so I'll just write dock hpen hpen version see my dock is running now the next thing I have to create my self hosted Runner so let's create so what I will do I'll just try to make a

copy I'll go to the settings and here you'll get one option called actions go to Runners uh yesterday I created One S posted I'll just try to delete it I'll create a new one one okay now new sell firsted Runner Linux operating system and let's execute all the command one by one for I'll skip it I'll press enter Then enter the name of the runner

I'll give self hyen posted then everything you just need to escape okay enter enter enter now let me execute the final command okay so it is connected to my GitHub and listening for the jobs now I'll go back to my GitHub and I have to update my secret variable so yesterday I already updated but I already removed that uh like secret file what I will do I again update my value first of all AWS access key ID let's update value I'm going to update AWS XS K this is the accs K copy and here I'll will update

then uh AWS default region Reon it will remain same Us East only because you can see I'm inside North Virginia fine then aw secret access key let's update then one more thing you have to update ECR repo okay yesterday I have given the full link but you only need to give the name uh because I in the yaml file I already handled the URL you can see it will automatically get the URL okay so here I have written uh step login ECR output registery okay it will automatically get the URL so you just only need to give

the name so Visa repo okay now let's update okay I think everything is up to date mongodb the same URL I don't need to update it again because yesterday already updated fine now I'll go back and let me push the changes right now so here I'm going to push it right now so let's say CI CD updated now see uh if I go to my GitHub and refresh see my action has started now if I click on action now see this is running first of us continuous integration is running then continuous deployment so inside continuous integration it will do all the setups it will build and push the docker image

okay to the Amazon ECF everything would be happening automatically it's just a one time setup guys okay if you do one time setup so next time everything would be automated here so let's wait it will take some time and I think you got it uh why we are getting the error yesterday because I have to enable that option in the setting okay so if you're getting the same issue we just need to enable that option e

now see almost done it is uh set uping everything and code is updated in my GitHub guys you can get the code from there and also try to start this repo guys start this repo you can Fork this repo okay now see it is pushing the image okay now if I uh if you go to the ECR uh let me show you if I go to the ECR right now container history so inside that you will see the docker image as of now I think it hasn't updated H you have to wait once pushing is done you will see the docker image there okay see it's done now continuous deployment has started now if I go to my ECR and refresh you will see this is the

image okay my image has been published now inside continuous deployment what is what will happen it will uh pull pull that image and it will run in my ec2 machine see it is pulling after pulling it will run okay see completed now if I go back to my C2 and let's go to the instance now I can close this tab it's not required now I'll go to my instance ID now copy the public IP okay so this one and if I paste it here now see if I

execute it won't be working why because I was running this application in a different port I think remember if I open my app. Pi port number I was using 8080 so I have to do the port mapping here see it's not reaching out that server so let's do the port mapping to to apart from the port mapping go to the security then uh security groups here you will see one option call edit inbound rules click here and add rules custom TCP port number you have to give 8080 and just select 0000 0 okay this one and save the

rules now if I go to the instance again now copy the public IP and again just paste it here and now give the port number clone port number 8080 8080 now see if I execute my application should be running see guys all of you my application is running see guys my application is running now if you hit this URL you will see my application but I'm not going to share uh if you hit so it will charge me so I'll delete it okay so you can do the deployment like that fine guys so we just fix the issue whatever issue actually we are having now see uh if you want to terminate

everything uh see if deployment is done so what you can do you can terminate everything so let me show you how to terminate see if I now uh close this window close this tab also still my application would be running okay see it is still running got it now let me terminate everything to terminate go to the instance first of all select the instance and click on the instant St and terminate okay terminate and delete the instance so it will uh stop and it will delete okay after some times it will take some time then it will delete

everything then you also need to delete your ECR repo so go to the elastic container history then select it and delete it it is still contains some image delete okay I think my machine has been stopped now let me just try okay I think my E2 is still deleting that's why it's taking time after some time actually you can delete now let me delete the user I am user select it and delete it so you have to give the name Visa user and you you can also delete your bucket okay so bucket wise I already deleted yesterday so everything is deleted

fine so now get back all right so guys our deployment part is done I think you saw how we can perform the deployment so in other video actually we'll also see how we can use any different like tool I'm not getting easy to instance connect option while connecting to the instance I'm not sure your account is active or not you just need to add your card okay make sure and you got that $300 credit D okay then definitely we'll get the option because we we are getting the option now okay so now uh let's start

with our main agenda uh I'll be discussing about actually resume discussion and carer guardiance see many people has this question actually like how to like mention projects in my resume and how to prepare the resume in a proper format okay so everything we'll try to clarify in this video so this resume and carrier all right see whenever you are preparing the resume sorry so you have to keep some of the point in your mind so the first thing you should have one uh proper summary about yourself

okay proper summary like uh what is your current let's say job role how many years of experience you are having okay uh what are the skill sets you are having you can just write one small summary on top of it I'll show you some of the demo of the resume that I'm this part would be more clear as of now just try to remember this is one thing you have to always mention in your resume proper summary about yourself okay at the very first of the resume then you have to mention your uh um employment history employ employment history or you can also consider job

experience okay job experiences this should be the second priority okay always try to remember guys and whenever you are mentioning this employment history and job experience so make sure you mention the latest job okay latest job at the very fast and older job lost okay this thing you have to maintain then the third thing you have to mention the project okay projects at least uh three project just a minute three project should be there okay I'll tell you how to choose the Project based on that you have to choose

okay what kinds of project you have to mention whether it is from machine learning deep learning NLP computer visual jni okay based on the job description here let me note Down based on job description okay J fine then you have to mention the skills okay skills you have then you can mention your education okay education again whenever you are mentioning the education make sure your current education will come at the very first then oldest fine then uh if you have any research and publication research and

publication then you can mention and uh if you are having any awards okay you can also mention and if you have any certificate ific at also you can also mention certificates okay certificates and those who are actually interested uh teaching so what you can do if you have any online lectures you can also mention online lectures okay online lectures what happens see whenever you represent these are the thing now uh actually whoever seeing your resume they will get impressed by seeing your resume like you

know these are the thing you have these are the skills yes or no got it so that's why uh if you have any skill set like uh online lectures uh let's see you can take let's see if you have done any open source contribution also you can mention it if you have any certification you can also mention it if you have any awards you can also mention it okay if you have any research and publication you can also mention it so these are like extra okay extra I mean benefit you will get okay extra benefit you will get unless until

you are not showing your let's say achievements so how people will get to know you know these are the thing got it and one thing uh you have to mention um after actually uh see for this actually we don't create any kinds of section but here you can consider proper summary uh you can also mention your online profile online profiles okay like your GitHub LinkedIn then if you have any YouTube channel you can also mention them okay one by one so these are the point actually you have to remember whenever you are preparing any

kinds of resum it is for fresard it is for experienced person okay any any of the people you can follow these are the point if you are preparing your resume and for freshers you have to complete your resume in just one page okay always try to remember if you a fresher if you just passed out let's say from your college what you can do uh create a one page resume instead of taking multiple pages okay but if you have experiences if you have multi projects okay if you have research publication Awards then you can take two P resume if you have let's say more than like 8 years 15

years 16 years experience that time you can take three pages resume okay but I will try to suggest try to complete um one to two pages okay always one to two pages this is preferable actually why because those who will be checking your resume so they don't have that much of time to go through all the information Okay so that's why you have to keep your resume short no need to write unnecessary information only just try to write necessary information okay whatever actually let's say it is required for that JD okay that is what

you have to remember now let me show you one resume template see this resume actually I prepared long back whenever I was uh I was actually student that time actually I prepared this resume okay uh I I I haven't updated this resume so far okay but this is the older version of the resume but again this is a good format actually okay this is a good format whatever things I mention in my board everything I have added here that's why I'm I'm going to show you the example with this resume guys and no need to worry I'm having lots of resume

template if you want I can also share so let me share guys so this is one repository I created called data science resume template and here you will get all kinds of resume freshers experience if you have four four 3 years experience 15 years experience 12 year experience six years experience okay all kinds of resumes are available guys all kinds of resume templates are available you can download anything and you can update your resume now let me share the link so this is the link you can forkit you can start it so that uh you will get

a you will get actually material okay you can uh refer it later on now let me show this resume see this is the resume uh I had prepared see the first thing I told you mention your profile uh just a minute why it's not working okay now it's working fine but before that try to mention your name as well as your current designation if you if you are already working you can mention your current designation okay if you are not working what you can give you can give data scientist Enthusiast okay or whatever let you can write here you can if you see the

template I have shared with you so there you will see okay whatever things you have to write here now mention your profile profile means your u i mean one small summary about yourself so here I have written one summary you can see data scientist and lecturer with 2.

6 years of working experience in the field of machine learning deep learning microcontrollers electronic system okay hands-on experience in classification regression clustering computer vision naal language processing and transform learning model to solve challenging business problem see in this short summary I have mentioned everything I know in the short summary I have mentioned everything I know that's how you also need to prepare One summary about yourself so that if reading if anyone reading your profile if anyone reading your summary they can easily get to know okay so you are a data scientist you are having 2.6 years of experience and you can uh work with machine

learning deep learning microcontrollers electronic system even you can also work with classification regression computer vision natural language processing and so on these are the technology okay so that's why you always need to take care this profile part guys all of you this is very much important because I have seen lots of resume they won't be mentioning this part so this is the first impression of your resume always try to remember if anyone let's say u i mean wants to review your resume if they're getting this kinds of let's say summary definitely after reading actually they

will get interest to look through your resume like uh I mean what are the things you have done so far got it then you can also mention your details let's say details wise you can mention your phone number email address as well as your uh address the current address you are staying and all you can mention it because let's say if your resume gets short rested so what they will do they will contact okay they will contact you just a minute fine then you have to mention your online profile as I already told you if you have any online profile let's say

LinkedIn GitHub YouTube If you are also having let's say research related online profile let's say Google Scholar research gate you can also mention it see I already have these are the profile I already mentioned if I show you my Google Scholar this is my Google Scholar this is my research G okay and this is my YouTube channel I think you know GitHub LinkedIn so LinkedIn GitHub it is very much important guys if you don't have these are the social media is completely fine but you have to create LinkedIn as well as the GitHub okay both

profile you have to mention guys make sure because this is very much important if you going for any tech company this thing you have to mention and if you are having any research get profile you can also mention if you're having any let's say Google scorer profile you can also mention see I already have some publication so that's why I mention these are the thing got it fine okay okay now next thing employment history that means your job experience so here I have added my job experience you can see so that time actually I was a junior data scientist

at Inon intelligence private limited and this was my joining date okay then uh this is my responsibility there in that company like managing and developing various research project in the field of machine learning deep learning computer vision NLP monitoring students and job Professionals in the data science career transition content creation based on the latest resarch paper and Technology collaborating with the developer team to build some products so these are the activity I used to perform whenever I was a junior data scientist at Inon okay so that's how uh I mean whatever let's

say job responsibility you are having currently in your company you can mention it here okay in a bullet point and no need to mention unnecessary information guys again just try to write in a neat and clean way okay this is super important because if you're writing I mean too much information in that let's say bullet point again people won't be able to interest to look through your resume got it if you see my resume it's like very neat and clean very neat and clean okay that's why you also need to make a neat and clean resume

always fine now let me remove now see next thing I have added my second job experience so I was working as a j machine learning engineer as a Japan Bangladesh Robotics and advanced technology Research Center this was my joining this and this was your ending date uh and again you have to mention your job responsibility there okay what you used to perform that's how if you have let's say multiple job experience at least try to add three okay three recent job experience but if you have multiple sorry if you have if you have only one or two you can mention all of

them but if you have more than three more than three that time you can only mention three because if you're adding let's say um I mean so many job experience here then your let's say space will be occupied and you won't be able to write any other thing there again your resume will be bigger that time got it so again it's super important you have to remember whenever you are preparing your resum always try to mention the currect current actually job respons I mean current job Ro not the older one try to mention current two to three job Ro it is completely

fine okay then you have to mention your skills okay see skills wise I have mention at the left hand side why because I don't want to take any space here I just wanted to let's say free up my space that's why left hand side I created one boundary and here actually I mention all of the skills I'm having okay so this was my actually previous skill old old skill actually I told you this is like very old resume I I haven't updated so far that's why you can see so that's how you can mention all of the skill you are having let's say python web scrapping machine learning deep learning naal language processing okay if you know any hardware let's

Rasberry P RDO then operating system to Windows okay then mlops uh then DBC ml FL whatever things you know just try to mention because let's say if this resume is checking by the automatic resume parser Auto resume parser I think you know currently in the company they're using Auto resume parser let's say software so they will up upload your resume and this software will check whether your resume is suitable for the job roll or not so this kinds of tag will help you that time this kinds of tag will help you let's say in the job

description they're looking for someone who knows mlops so if you mention mlops in the skill set definitely this will match and there would be like highest probability let's say 80% probability it is messing with your resume this job description that time your resume would be short listed okay otherwise it would be rejected so you also need to take care this part guys see resume preparation is not a easy task guys okay you just need to research you just need to research everything like how you can prepare how you can write everything in just one

place okay in a neat and clean way so that it would be suitable for the human as well it would be suitable for the automatic software as well let me know guys no response guys please uh give some response in the chat whether you are enjoying the session or not if you want this kinds of session you can let me know fine now the next thing you have to mention about your projects project is super important guys now you can ask me how to mention my projects okay how to mention my projects in my resume see that's how you have to mention the project let's say I had implemented one

projects whenever I was working in the Inon attendance system using facial recognization okay so just try to give the title first of all okay title of your project then give the date like in which date you started implementing and which date you finish that particular project okay yes you will also get everything in the GitHub I will share with you some okay first of all try to check now see I have also added the GitHub I have also added the GitHub link of that project let's say if someone is interested to check my project they can click here and

they will redirect to the GitHub page and they can get to know but see whenever you are uploading your project to the GitHub whenever you are committing make sure you maintain the readme file properly see the way we have maintained the readme file now here everything we have mentioned like how to set up my project what is the workflow of my project okay then what are the things you have to set up as an environment variable okay everything I have set up everything I have mentioned here so that if you're referring my if you're refer ing my let's say GitHub repository so

you will be able to set up this project in your system so always try to mention sorry always try to let's say maintain this R me file all of you this is super important no need to only commit the code in the GitHub after committing the GitHub you have to maintain the readme file in a proper way so so that if anyone let's say referring your repository they will be a able to uh let's say install this project in their system as well that thing you have to always take care fine so that's why uh here you can see I have added the

GitHub link now here you have to mention uh I mean what is the problem statement you are solving and what kinds of Technology you are using here see I have written in a smarter way automated the end process of the attendance of the employee in the company using deep learning uh deep neural network computer vision and mongodb database see whatever problem statement I solve I also mention as well as the technology I'm using I also mentioned here okay in just one line now reducing 4 hours of manual attend in working uh time per day that means what kinds of problem statement you are solving what is the benefit you

are having okay after implementing this project everything I have mentioned in just one line so this is like very smarter approach guys always try to remember uh I mean this technique whatever project you are implementing first of all try to give the problem statement then whatever technology you are using and also try to give the solution like what kinds of solution you are proposing and how it is helping the people okay here you can see I have added one very powerful point 4 hours of manual attendance working time per day that means we are reducing four hours let's say load okay four hours actually

workload we are lo I mean reducing after implementing this attendance system got it then created a desktop application using uh kter and python for employee so that they can they could interact easily okay then we used mtcnn for the detecting the ph and Arc pH for extracting the face embedding and recognizing uh can reduce uh the workload by 50% okay that's how you have to give some bullet point around three to four bullet point you have to give and inside that you have to mention each and everything of your project like

whatever technology whatever problem statement whatever architecture okay everything you can mention here I hope it is clear guys all of you now this is the next project guys you can see this is the next project image secer and Auto image classification Library so this Library I implemented if you if I show you the uh GitHub implementation so this is the GitHub guys this is the official GitHub uh I implemented this package long back around two years back I implemented this package see this is the package so this is one Auto image classification package I implemented okay let's say if you want

to perform image classification but you don't know about coding you don't want to write too much of code what you can do you can install uh IM seeker and you can launch it after that you can upload your data and it will automatically do the image classification for you see everything I have mentioned here you can check it out and here is the pii website as well see it is already hosted in the piie you can also install this package guys okay pep install image seeker and you can see I'm the author okay this package I implemented and here is the demo how to use it and all okay

everything I have added here so that's how if you have any uh open source work if you have any public work this kinds of work you can mention here okay again this is a very uh I mean uh I mean important let's say I mean things you will be doing here got it now here I mentioned image Seeker reduces developer time of writing too much code for any image classification task so that they focus on the parts of the problem that really matters created a python package hosted on pii website so that people can easily install uh pep package man manager use

tensorflow kasas in the back end of the library kasas adopts the principle of progressive disclosure of the complexity simple workflow should be quick and easy that's why MC is a flexible okay that's how you have to write about your project got it guys so that's how you have to mention around two to three project at least two to three project you have to mention in the resume see I have added four project here four project actually I have added you can see four different different projects I have added got

it now you can ask me how to how to mention the projects see whenever you are mentioning the project first of all see the job description let's say if I search for any jobs let's say if I go to the job section let's say here I will search the job rule let's a data scientist let's say India India or us whatever location you prefer you just try to search here now let's say this company is hiring data scientist so what I'll do I'll just click here now see they have mentioned the job description okay they have mentioned the

job description now try to uh carefully read the job description okay see what kinds of responsibility they're looking for machine learning okay then linear regression optimization forecasting technique okay so just try to read the job description properly based on that try to decide what are the project you should mention let's say you know machine learning deep learning computer vision gen everything but you are interested about this job no need to mention about like gener TBI project no need to mention about let's say computer vision projects only you just need to mention related machine learning project because they're looking for machine

learning developer here okay machine learning related job Ro they're looking for that time what you will do you will only mention two to three project which is relevant to the machine learning not the computer visual or not the natural language processing okay but what I have seen people will mention everything people will mention everything they will mention machine learning deep learning NLP computer vision and everything okay so always try to create two to three copy of your resume always two to three copy just try to create and try to upgrade okay try to upgrade this

resume based on the job description if they looking for machine learning developer try to mention uh let's see skills and project related machine learning if they're looking for gni if they're looking for let's say computer vision try to mention with respect to that only but if they're looking for everything the complete data scientist that time you can mention everything there is no issue okay so this thing you have to always I mean remember guys okay this is the super important things that's how gets rejected actually okay don't do this do this mistake this kinds

of mistake first of all read the job description properly based on that try to mention your projects and skills in your resume okay this is the idea got it then you have to mention your education whatever education you have done just try to mention then if you have any research and publication you can also mention see I already have one publication I have mentioned here now if you have any awards you can also mention if you have any certification you can mention if you have any online lectures you can also mention okay so these are the extra curriculum if you're not mentioning it's completely fine if

you don't have but if you have it try to mention okay this would be a plus point for you got it so this is how actually we can create a resume for a data science not a data science for any any any kinds of domain okay any kinds of domain you can follow this technique now for freshers you have to complete in just one place and for experienced person actually you have to complete in two two pages or two to three pages okay this is what actually you have to remember okay and here I have already uh let's say collected R of resume template and I

created one repo so let me share with you so this resume is also available inside this GitHub you can see um buy resume I think yeah this one so you can download and you can check it out and also try to restart this resume guys all of you this uh I mean repository so that um I mean people will get to know okay this kinds of let's say repository is there so that they can also get the template of the resume fine now you can ask me if you have any question guys question related resume question related carer guidance you can let me know in the chat I'll just try to answer your question or else

I think I can end the session right now because I have explained each and everything and um some of the tips I want to give let's say whenever you are attending for any interview if they're asking the question uh related to your project first of all try to I mean give the proper let's say problem statement like the way I started the project I think if you if you check my lecture now from day one I think you saw how I explained my project there project introduction setup whatever tools and Technology I'm using okay that's how you also need to like explain about your

project first of all give the proper problem statement then what is the solution you are proposing what is your job responsibility there okay what kinds of tools and Technology you are using why you are using this to tools and Technology okay these kinds of question they will be asking and you have to prepare for that and if you have implemented the project definitely you will be able to give the answer okay instead of copy pasting from any other let's say person project and this is this is one good project guys the project we have implemented this project you can mention

in your resume but I'll suggest you guys try to P the kaggle.com all of you and from there you can see like real time problem statement all of you so pick up one data set and try to follow the same Pipeline and Implement one project Implement one good project if I go to the data set see thousands of data sets are available so sech for let's say machine learning I did machine learning related data see lots of data set you will be getting okay and lots of actually let's say uh competition is also going on if you go to the competition

lots of realtime competition is also going on okay you can also attend this competition even you will also get the prize money got it so that's how you can pick different different problem statement and you can implement the solution and you can mention in your resume okay so this would be considered as a real well product instead of implementing Titanic project okay or let's say uh I mean U Mist classification project don't need to do like that try to pick up any real-time problem statement always fine any question anyone you can let me know

and uh soon you will get the next announce announcement guys uh I will be bringing another live S Series okay for you I will announce on my YouTube channel for this you can uh I mean uh let's say check my community okay Community section so there I will give the update okay I think no question so thank you guys for joining the session so now I think we can end the session if you don't have any query okay you are watching my video and doing the coding that's great yeah definitely it's a good practice so you should write the code okay although you are watching my code but you are writing at least now you are

getting the entire understanding so this is super important got it fine any question guys let me know and please try to subscribe to the channel guys let's hit 14k subscriber and please share and hit the like button all of you okay so so thank you guys for joining the session uh it is possible for a candidate with a very bad academic per performance for four years of Gap graduation get the definitely nikil if you have the proper skill set definitely will also crack

this kinds of job R because I've have seen many people they have already done so no need to worry about the Gap and I mean your academic results always try to focus on the skills because companies will always look for the skills okay not academic performance so this kinds of project you have to implement with end to end manner uh build a proper resume with all the skills projects you are having then keep on applying different different job position okay definitely you will get a call all right okay fine so thanks for joining joining the session guys uh bye everyone good night thank you for

Show thinking

This was the final session of the "MLOps Production Ready Machine Learning Project" series. It started by resolving the deployment issue from the previous session and then transitioned into a detailed career guidance segment on how to build an effective resume and discuss projects in interviews.

**🚀 Part 1: Resolving the Deployment Issue & Successful CI/CD**

The session began by successfully deploying the application, fixing the authentication error that occurred previously.

* **The Fix:** The issue was a missing permission in the presenter's GitHub account settings. For a Personal Access Token (PAT) to modify workflow files, the **"workflow" scope** must be enabled. After enabling this, the commit was successful.
* **Live Deployment Walkthrough:**
  1. The presenter quickly re-created the necessary AWS resources: an **IAM User**, an **ECR (Elastic Container Registry)** repository, and an **EC2 (Elastic Compute Cloud)** instance.
  2. He configured the EC2 instance as a **self-hosted runner** for GitHub Actions, linking the AWS server to the GitHub repository.
  3. He committed the Dockerfile and the GitHub Actions workflow file (.github/workflows/aws.yaml).
  4. The push automatically triggered the **GitHub Actions CI/CD pipeline**. The pipeline successfully built the Docker image, pushed it to ECR, and deployed it on the EC2 instance.
  5. The final step was **Port Mapping**: opening port 8080 in the EC2 instance's security group to make the FastAPI application accessible to the public.
* **Result:** The presenter navigated to the EC2 instance's public IP address on port 8080 and showed the **live, running web application**. He also demonstrated how to terminate all AWS resources to avoid costs.