AWS AI-ML VIRTUAL INTERNSHIP

AN INTERNSHIP REPORT

By

G. Urmila

(21JG5A0506)

Under the esteemed guidance

Mrs. D. Indu

Assistant Professor

CSE Department

Department of Computer Science and Engineering

GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN

[Approved by AICTE NEW DELHI, Affiliated to JNTUK Kakinada]

[Accredited by National Board of Accreditation (**NBA**) for B.Tech. CSE, ECE & IT – Valid from 2019-22 and 2022-25]

[Accredited by National Assessment and Accreditation Council (**NAAC**)– Valid from 2022-27]

Kommadi, Madhurawada, Visakhapatnam–530048

2023–2024

GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Internship report titled "AWS AI-ML Virtual Internship" is a bonafide work of following IV B.Tech students in the Department of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering for Women affiliated to JNT University, Kakinada during the academic year 2023-24 final semester

G. Urmila **21JG5A0506**

Mrs. D.Indu
Assistant Professor
(Internal Guide)
Dept. of CSE

Dr. P. V. S. L. Jagadamba
Professor
(Head of the department)
Dept. of CSE

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts with success.

We feel elated to extend our sincere gratitude to **Mrs. M. Aswini,** Assistant Professor for her encouragement during analysis of the project. Her annotations, insinuations, and criticisms are the key to the successful completion of the thesis and to providing us with all the required facilities.

We express our deep sense of gratitude and thanks to **Dr. P. V. S. Lakshmi Jagadamba**, Professor and Head of the Department of Computer Science and Engineering for her guidance and for expressing her valuable and grateful opinions on the project for its development and for providing lab sessions and extra hours to complete the project.

We would like to take this opportunity to express our profound sense of gratitude to Vice Principal, **Dr. G. Sudheer** for allowing us to utilize the college resources thereby facilitating the successful completion of our project. We are also thankful to both the teaching and non-teaching faculty of the Department of Computer Science and Engineering for giving valuable suggestions for our project.

We would like to take the opportunity to express our profound sense of gratitude to the revered Principal, **Dr. R. K. Goswami** for all the help and support towards the successful completion of our project.

ABSTRACT

This is a two-phase course comprising of CLOUD FOUNDATIONS and MACHINE LEARNING.

AWS Academy Cloud Foundations is intended for students who seek an overall understanding of cloud computing concepts, independent of specific technical roles. It provides a detailed overview of cloud concepts, AWS core services, security, architecture, pricing, and support.

We learned how to use AWS Cloud services basics and how the adoption framework works. Computational services like AMAZON EC2, LAMBDA etc. We learned about different storage types like AMAZON EBS, S3, EFS. This course also introduced about the AWS databases like DYNAMODB, REDSHIFT, RDS.

Machine learning is the use and development of computer systems that can learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.

In this course, we will learn how to describe machine learning (ML), which includes how to recognize how machine learning and deep learning are part of artificial intelligence**,** it also describes artificial intelligence and machine learning terminology. Through this we can identify how machine learning can be used to solve a business problem. We can also learn how to describe the machine learning process in detail and the list the tools available to data scientists to identify when to use machine learning instead of traditional software development methods. Implementation of a machine learning pipeline, which includes learning how to formulate a problem from a business request, obtain and secure data for machine learning, use Amazon Sage Maker to build a Jupyter notebook, outline the process for evaluating data, explanation of why data must be pre-processed. Using the open-source tools to examine and pre-process data. We can use Amazon Sage Maker to train and host a machine learning model.

It also includes in the use of cross validation to test the performance of a machine learning model, use of hosted model for inference and creating an Amazon Sage Maker hyperparameter tuning job to optimize a model's effectiveness. Finally, we will learn how to use managed Amazon ML services to solve specific machine learning problems in forecasting, computer vision, and natural language processing.

For this course we took a case study- "Unlocking Clinical Data from Narrative Reports". Objective of this case study is to evaluate the automated detection of clinical conditions described in narrative reports using "Natural Language Processing".

TABLE OF CONTENTS

COURSE: **AWS CLOUD FOUNDATIONS**

COURSE: **MACHINE LEARNING FOUNDATIONS**

COURSE: **AWS CLOUD FOUNDATIONS**

MODULE: 1 CLOUD CONCEPTS OVERVIEW

1. Introduction to Cloud Computing:

Cloud computing is the on-demand delivery of compute power, database, storage, applications, and other IT resources via the internet with pay-as-you-go pricing. These resources run on server computers that are in large data centres in different locations around the world.

CLOUD SERVICES: There are three main cloud service models. The below figure-1.1 explains about the control over IT resources.

Infrastructure as a service (IaaS): IaaS is also known as **Hardware as a Service (Haas)**. It is a computing infrastructure managed over the internet. The main advantage of using IaaS is that it helps users to avoid the cost and complexity of purchasing and managing the physical servers.

Platform as a service (PaaS): PaaS cloud computing platform is created for the programmer to develop, test, run, and manage the applications.

Software as a service (SaaS): SaaS is also known as "**on-demand software**". It is a software in which the applications are hosted by a cloud service provider. Users can access these applications with the help of internet connection and web browser.

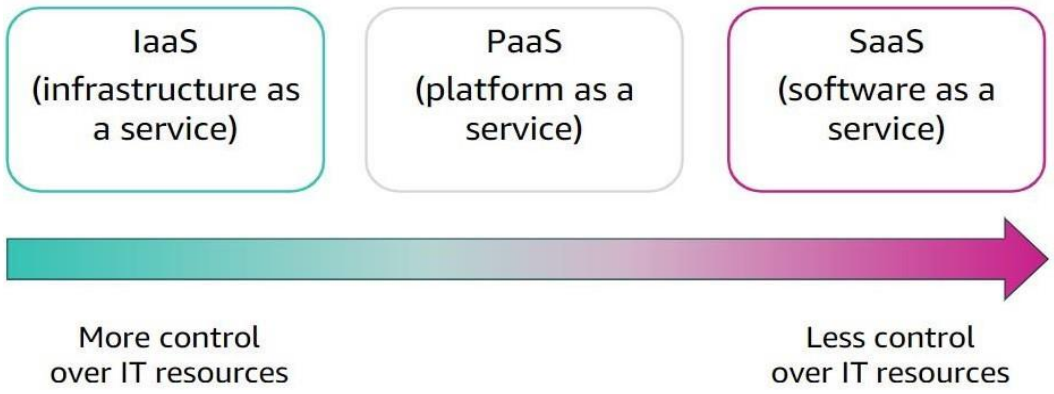


Figure-1.1: Cloud Services

2. Advantages of Cloud Computing:

1) Back-up and restore data
2) Improved collaboration
3) Excellent accessibility

4) iServices in the pay-per-use model

3. **Introduction to Amazon Web Services:**

Amazon Web Services (AWS) is a secure cloud platform that offers a broad set of global cloud-based products shown in below figure-1.2. Because these products are delivered over the internet, you have on-demand access to the compute, storage, network, database, and other IT resources that you might need for your projects—and the tools to manage them.

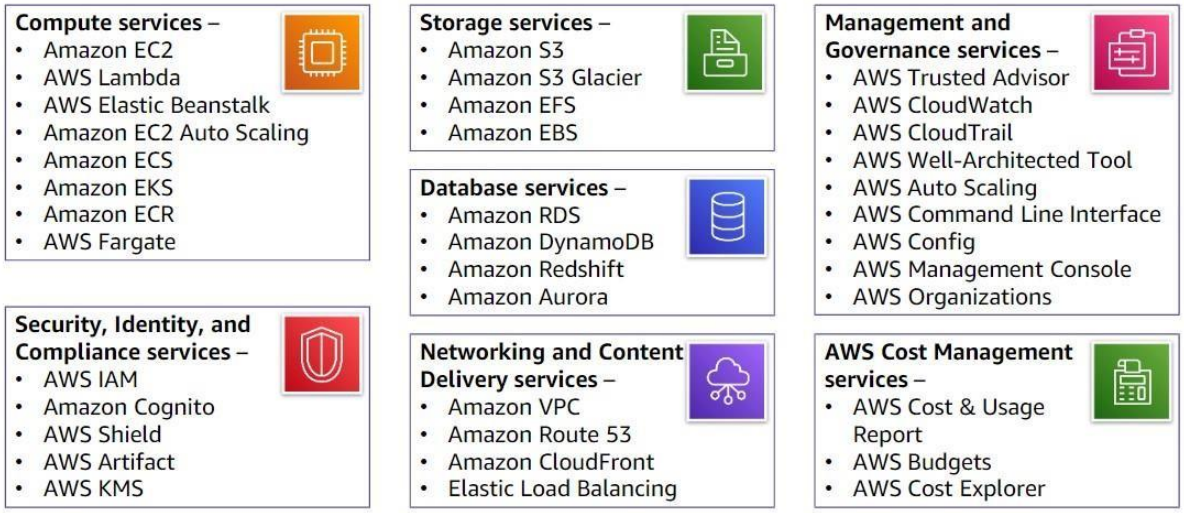


Figure-1.2- Amazon Web Services

4. **AWS Cloud Adoption Framework:**

The **AWS Cloud Adoption Framework (AWS CAF)** provides guidance and best practices to help organizations identify gaps in skills and processes. It also helps organizations build a comprehensive approach to cloud computing—both across the organization and throughout the IT lifecycle—to accelerate successful cloud adoption.

At the highest level, the AWS CAF organizes guidance into six areas of focus, called perspectives. Perspectives span people, processes, and technology. Each perspective consists of a set of capabilities, which covers distinct responsibilities that are owned or managed by functionally related stakeholders.

MODULE: 2 CLOUD ECONOMICS AND BILLING

1. Fundamentals of Pricing:

There are three fundamental drivers of cost with AWS: compute, storage, and outbound data transfer, figure-1.3 shows the fundamental pricing. These characteristics vary somewhat, depending on the AWS product and pricing model you choose.

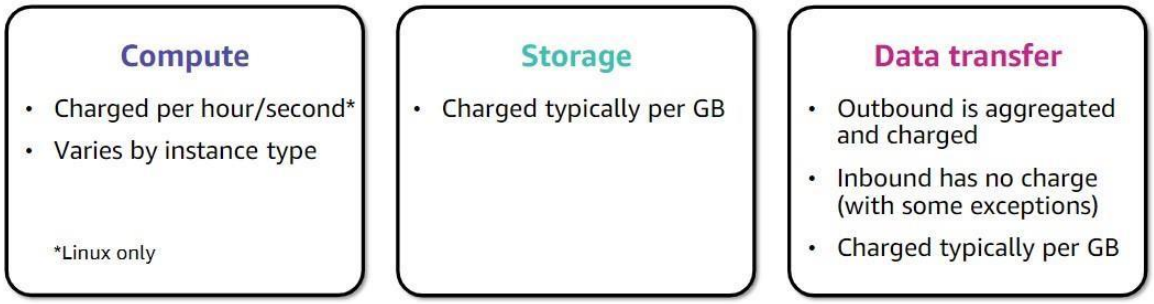


Figure-1.3-Fundamental Pricing

2. Total Cost of Ownership:

Total Cost of Ownership (TCO) is the financial estimate to help identify direct and indirect costs of a system.

- To compare the costs of running an entire infrastructure environment or specific workload on-premises versus on AWS

3. AWS Organisations:

AWS Organizations is a free account management service that enables you to consolidate multiple AWS accounts into an organization that you create and centrally manage.

AWS Organizations include consolidated billing and account management capabilities that help you to better meet the budgetary, security, and compliance needs of your business. The main benefits of AWS Organizations are:

- Centrally managed access policies across multiple AWS accounts. Controlled access to AWS services.

4. AWS Bill And Cost Management:

AWS Billing and Cost Management is the service that you use to pay your AWS bill, monitor your usage, and budget your costs. Billing and Cost Management enables you to forecast and obtain a better idea of what your costs and usage might be in the future so that you can plan.

MODULE: 3 GLOBAL INFRASTRUCTURE OVERVIEW

1. AWS Global Infrastructure:

The AWS Global Infrastructure is designed and built to deliver a flexible, reliable, scalable, and secure cloud computing environment with high-quality global network performance.

AWS Global Infrastructure Map: https://aws.amazon.com/about-aws/globalinfrastructure/#AWS_Global_Infrastructure_MapChoose a circle on the map to view summary information about the Region represented by the circle.

Regions and Availability Zones: https://aws.amazon.com/about-aws/globalinfrastructure/regions_az/Choose a tab to view a map of the selected geography and a list of Regions, Edge locations, Local zones, and Regional Caches.

2. AWS Service Overview:

AWS provides us 4 types of services they are classified as 1. Applications, Platform Services, Foundations Services, Infrastructure. Observe figure-1.4 below explains the subcategories that shared in each of the services provided by AWS. We can observe that foundation services contains Compute for virtual, automatic scaling and load balancing, contains networking and storage as object block and archive.

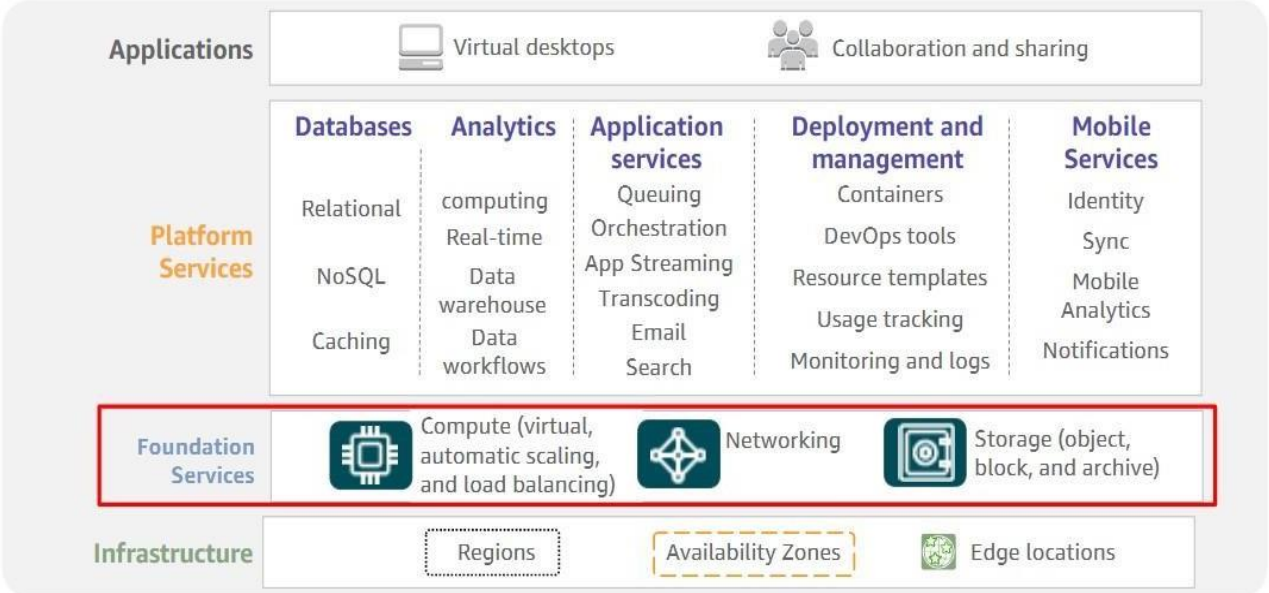


Figure-1.4-AWS services overview

MODULE: 4 CLOUD SECURITY

1. AWS Shared Responsibility Model:

AWS responsibility:

Security of the cloud AWS responsibilities:

- Physical security of data centres i.e., Controlled, need-based access. Virtualization infrastructure.

Customer responsibility:

Security in the cloud Customer responsibilities:

- Amazon Elastic Compute Cloud (Amazon EC2) instance operating system including patching, maintenance Applications: -Passwords, role-based access, etc.

2. AWS Identity and Access Management (IAM):

IAM is a no-cost AWS account feature.

Use IAM to manage access to AWS resources–

- A resource is an entity in an AWS account that you can work with Which resources can be accessed and what can the user do to the resource How resources can be accessed.

3. Securing a New AWS Account:

AWS account root user access versus IAM access

Best practice: Do not use the AWS account root user except when necessary.

- Access to the account root user requires logging in with the email address that you used to create the account.

Example actions that can only be done with the account root user:

- Change account settings **4: Securing Accounts:**

Security features of AWS Organizations:

- Group AWS accounts into organizational units (OUs) and attach different access policies to each OU.
- Use service control policies to establish control over the AWS services and API actions that each AWS account can access.

5.Securing Data on AWS:

Encryption encodes data with a secret key, which makes it unreadable.

- Only those who have the secret key can decode the data
- You can encrypt data stored in any service that is supported by AWS KMS, including: **Amazon S3, Amazon EBS, Amazon Elastic File System (Amazon EFS), Amazon RDS managed databases.**

MODULE: 5 NETWORKING AND CONTENT DELIVERY

1: Networking Basics:

Computer Network:

An interconnection of multiple devices, also known as hosts, that are connected using multiple paths for the purpose of sending/receiving data or media. Computer networks can also include multiple devices/mediums which help in the communication between two different devices; these are known as Network devices and include things such as routers, switches, hubs, and bridges. Computer network can better understandable using OSI INTERCONNECTION MODEL-Table1.1. It contains of 7 layers of networking, each layer confine with specifications and tasks.

Layer	Number	Function	Protocol/Address
Application	7	Means for an application to access a computer network	HTTP(S), FTP, DHCP, LDAP
Presentation	6	• Ensures that the application layer can read the data • Encryption	ASCI, ICA
Session	5	Enables orderly exchange of data	NetBIOS, RPC
Transport	4	Provides protocols to support host-to-host communication	TCP, UDP
Network	3	Routing and packet forwarding (routers)	IP
Data link	2	Transfer data in the same LAN network (hubs and switches)	MAC
Physical	1	Transmission and reception of raw bitstreams over a physical medium	Signals (1s and 0s)

Table-1.1-OSI INTERCONNECTION MODEL

2. Amazon Virtual Private Cloud (VPC):

Enables you to provision a logically isolated section of the AWS Cloud where you can launch AWS resources in a virtual network that you define.

Gives you control over your virtual networking resources, including:

- Selection of IP address range
- Creation of subnets
- Configuration of route tables and network gateways
- Enables you to customize the network configuration for your VPC Enables you to use multiple layers of security

3. VPC Networking:

There are several VPC networking options, which include:

- Internet gateway: Connects your VPC to the internet
- NAT gateway: Enables instances in a private subnet to connect to the internet
- VPC endpoint: Connects your VPC to supported AWS services
- VPC peering: Connects your VPC to other VPCs
- VPC sharing: Allows multiple AWS accounts to create their application resources into shared, centrally managed Amazon VPCs
- AWS Site-to-Site VPN: Connects your VPC to remote networks
- AWS Direct Connect: Connects your VPC to a remote network by using a dedicated network connection

4. VPC Security:

- Build security into your VPC architecture:
- Isolate subnets if possible.
- Choose the appropriate gateway device or VPN connection for your needs.
- Use firewalls.
- Security groups and network ACLs are firewall options that you can use to secure your VPC.

5. Amazon Router 53:

- Is a highly available and scalable Domain Name System (DNS) web service.
- Is used to route end users to internet applications by translating names (like www.example.com) into numeric IP addresses (like 192.0.2.1) that computers use to connect to each other. Is fully compliant with IPv4 and IPv6.

6. Amazon Cloud Front:

- Fast, global, and secure CDN service.
- Global network of edge locations and regional edge caches.
- Self-service model.
- Pay-as-you-go pricing.

MODULE: 6 COMPUTE

1. Compute Services Overview: Compute services like Amazon EC2, Lambda, ECS, EKS, Elastic Bean Stalk. Below table-1.2 describes key concepts of each service and characteristics.

Services	Key Concepts	Characteristics	Ease of Use
• Amazon EC2	• Infrastructure as a service (IaaS) • Instance-based • Virtual machines	• Provision virtual machines that you can manage as you choose	A familiar concept to many IT professionals.
• AWS Lambda	• Serverless computing • Function-based • Low-cost	• Write and deploy code that runs on a schedule or that can be triggered by events • Use when possible (architect for the cloud)	A relatively new concept for many IT staff members, but easy to use after you learn how.
• Amazon ECS • Amazon EKS • AWS Fargate • Amazon ECR	• Container-based computing • Instance-based	• Spin up and run jobs more quickly	AWS Fargate reduces administrative overhead, but you can use options that give you more control.
• AWS Elastic Beanstalk	• Platform as a service (PaaS) • For web applications	• Focus on your code (building your application) • Can easily tie into other services—databases, Domain Name System (DNS), etc.	Fast and easy to get started.

Table-1.2-compute services

2. Amazon EC2:

Amazon Elastic Compute Cloud (Amazon EC2):

- Provides virtual machines—referred to as EC2 instances—in the cloud.
- Gives you full control over the guest operating system (Windows or Linux) on each instance.
- Launch instances with a few clicks or a line of code, and they are ready in minutes. You can control traffic to and from instances.

3. Amazon EC2 Cost Optimization:

- On demand Instances- Pay By Hour
- Reserved Instances- Payment for instance reservation.
- Spot Instances-Instances run as long as the services runs.
- Dedicated Hosts- Physical servers with EC2 instance fully used.

4. Container Services:

Containers are a method of operating system virtualization. Benefits are:

- Repeatable.
- Faster to launch and stop or terminate than virtual machines.

5. Introduction to AWS Lambda:

It is a serverless computing service.

- It supports multiple programming languages.
- Completely automated administration.
- Built-in fault tolerance It supports the orchestration of multiple functions.

1.

MODULE: 7 STORAGE

Amazon Elastic Block Store (Amazon EBS):

Amazon EBS enables you to create individual storage volumes and attach them to an Amazon EC2 instance:

- Amazon EBS offers block-level storage.
- Volumes are automatically replicated within its Availability Zone.
- It can be backed up automatically to Amazon S3 through snapshots.

Uses include –

- Boot volumes and storage for Amazon Elastic Compute Cloud (Amazon EC2) instances. Enterprise applications.

2. Amazon Simple Storage Service (Amazon EBS):

- Backup and storage –Provide data backup and storage services for others
- Application hosting –Provide services that deploy, install, and manage web applications
- Software delivery –Host your software applications that customers can download

3. Amazon Elastic File System (EFS):

File storage in the AWS Cloud:

- Works well for big data and analytics, media processing workflows, content management, web serving, and home directories.
- Petabyte-scale, low-latency file system.
- Shared storage.
- Elastic capacity.
- Compatible with all Linux-based AMIs for Amazon EC2.

4. Amazon Simple Storage Service Glacier:

- Amazon S3 Glacier is a data archiving service that is designed for security, durability, and an extremely low cost.
- Amazon S3 Glacier is designed to provide 11 9s of durability for objects.
- It supports the encryption of data in transit and at rest through Secure Sockets Layer (SSL) or Transport Layer Security (TLS).
- The Vault Lock feature enforces compliance through a policy.
- Extremely low-cost design works well for long-term archiving.
- Provides three options for access to archives—expedited, standard, and bulk—retrieval times range from a few minutes to several hours.

1.

MODULE: 8 DATABASES

Amazon Relational Databases Service:

Amazon RDS is a web service that makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while managing timeconsuming database administration tasks so you can focus on your applications and your business. Amazon RDS is scalable for compute and storage, and automated redundancy and backup is available. Supported database engines include Amazon Aurora, PostgreSQL, MySQL, MariaDB, Oracle, and Microsoft SQL Server.

2. Amazon Dynamo DB:

Fast and flexible NoSQL database service for any scale.

- Virtually unlimited storage.
- Items can have differing attributes.
- Low-latency queries.
- Scalable read/write throughput.

The core DynamoDB components are tables, items, and attributes.

- A table is a collection of data.
- Items are a group of attributes that is uniquely identifiable among all the other items.

3. Amazon Redshift:

Usage Case:1

Enterprise data warehouse (EDW)

- Migrate at a pace that customers are comfortable with
- Experiment without large upfront cost or commitment
- Respond faster to business needs

Big data

- Low price point for small customers
- Managed service for ease of deployment and maintenance Focus more on data and less on database management.

Usage Case:2

Software as a service (SaaS)

- Scale the data warehouse capacity as demand grows
- Add analytic functionality to applications

4. Amazon Aurora:

- Enterprise-class relational database.
- Compatible with MySQL or PostgreSQL.
- Automate time-consuming tasks (such as provisioning, patching, backup, recovery, failure detection, and repair).

1.

MODULE:9. CLOUD ARCHITECTURE

AWS Well Architected Framework:

A guide for designing infrastructures that are:

- Secure
- High performing
- Resilient
- Efficient

A consistent approach to evaluating and implementing cloud architectures

2. AWS Trusted Advisors:

Cost Optimization–AWS Trusted Advisor looks at your resource use and makes recommendations to help you optimize cost by eliminating unused and idle resources, or by making commitments to reserved capacity.

Performance–Improve the performance of your service by checking your service limits, ensuring you take advantage of provisioned throughput, and monitoring for overutilized instances.

Security–Improve the security of your application by closing gaps, enabling various AWS security features, and examining your permissions.

Fault Tolerance–Increase the availability and redundancy of your AWS application by taking advantage of automatic scaling, health checks, multi-AZ deployments, and backup capabilities. **Service Limits**–AWS Trusted Advisor checks for service usage that is more than 80 percent of the service limit. Values are based on a snapshot.

RELIABILITY:

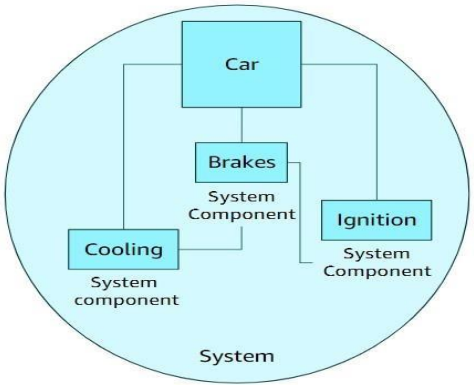


Figure-1.5-Relaibility of System

- A measure of your system's ability to provide functionality when desired by the user can be defined by reliability shown in figure-1.5.
- System includes all system components: Hardware and Software.
- Probability that your entire system will function as intended for a specified period.
- Mean Time Between Failures (MTBF)= total time in service/number of failures.

MODULE: 10 AUTO SCALING AND MONITORING

Elastic Load Balancing: Elastic Load Balancing automatically distributes your incoming traffic across multiple targets, such as EC2 instances, containers, and IP addresses, in one or more Availability Zones as shown in figure-1.6. It monitors the health of its registered targets, and routes traffic only to the healthy targets.

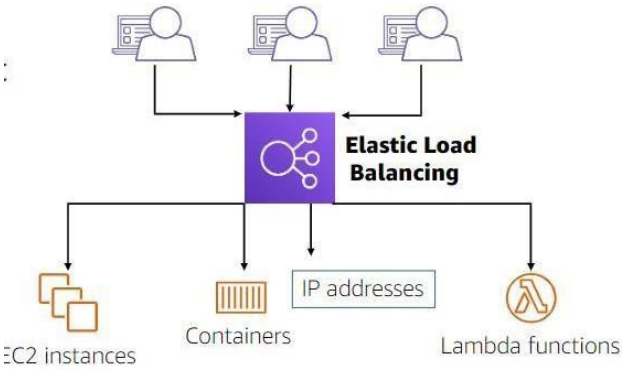


Figure-1.6-Elastic Load Balancing

Types of Elastic Load Balancing: There are 3 types of elastic load balancing consisting of Application Load Balancer, Network Load Balancer, Classic Load Balancer. Each of the 3 types has their better features and upgrades than the previous ones as shown in table-1.3.

Application Load Balancer	Network Load Balancer	Classic Load Balancer (Previous Generation)
• Load balancing of HTTP and HTTPS traffic	• Load balancing of TCP, UDP, and TLS traffic where extreme performance is required	• Load balancing of HTTP, HTTPS, TCP, and SSL traffic
• Routes traffic to targets based on content of request • Provides advanced request routing targeted at the delivery of modern application architectures, including microservices and containers	• Routes traffic to targets based on IP protocol data • Can handle millions of requests per second while maintaining ultra-low latencies • Is optimized to handle sudden and volatile traffic patterns	• Load balancing across multiple EC2 instances
• Operates at the application layer (OSI model layer 7)	• Operates at the transport layer (OSI model layer 4)	• Operates at both the application and transport layers.

Table-1.3-Types of Load Balancers

1. Amazon CloudWatch:

CloudWatch enables you to –

- Collect and track standard and custom metrics.
- Define rules that match changes in your AWS environment and route these events to targets for processing.

2. Amazon EC2 Auto Scaling:

- Helps you maintain application availability.
- Enables you to automatically add or remove EC2 instances according to conditions that you define.
- Detects impaired EC2 instances and unhealthy applications and replaces the instances without your intervention.

COURSE: MACHINE LEARNING FOUNDATIONS

MODULE 1: INTRODUCING TO MACHINE LEARNING

1. What is Machine Learning?

Machine learning is the scientific study of algorithms and statistical models to perform a task by using inference instead of instructions. Below Figure-2.1 represents machine learning flow

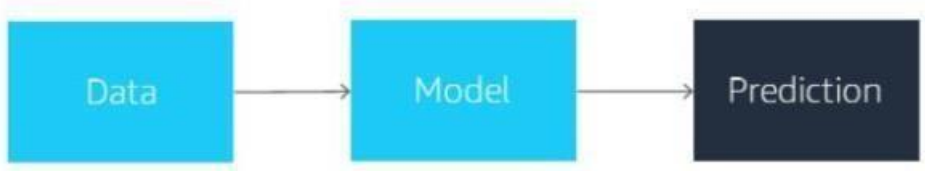


Figure-2.1-Machine Learning Flow

- Artificial intelligence is the broad field of building machines to perform human tasks.
- Machine learning is a subset of AI. It focuses on using data to train ML models so the models can make predictions.
- Deep learning is a technique that was inspired from human biology. It uses layers of neurons to build networks that solve problems.

2. Business Problems Solved with Machine Learning

Machine learning is used throughout a person's digital life. Here are some examples:

- Spam –Your spam filter is the result of an ML program that was trained with examples of spam and regular email messages.
- Recommendations –Based on books that you read or products that you buy, ML programs predict other books or products that you might want. Again, the ML program was trained with data from other readers' habits and purchases.

Machine learning problems can be grouped into –

- Supervised learning: You have training data for which you know the answer.
- Unsupervised learning: You have data, but you are looking for insights within the data.
- Reinforcement learning: The model learns in a way that is based on experience and feedback.

3. Machine Learning Process

The machine learning pipeline process can guide you through the process of training and evaluating a model.

The iterative process can be broken into three broad steps –

- Data processing
- Model training
- Model evaluation

ML PIPELINE: A machine learning pipeline is the end-to-end construct that orchestrates the flow of data into, and output from, a machine learning model as in figure-2.2 (or set of

multiple models). It includes raw data input, features, outputs, the machine learning model and model parameters, and prediction outputs.

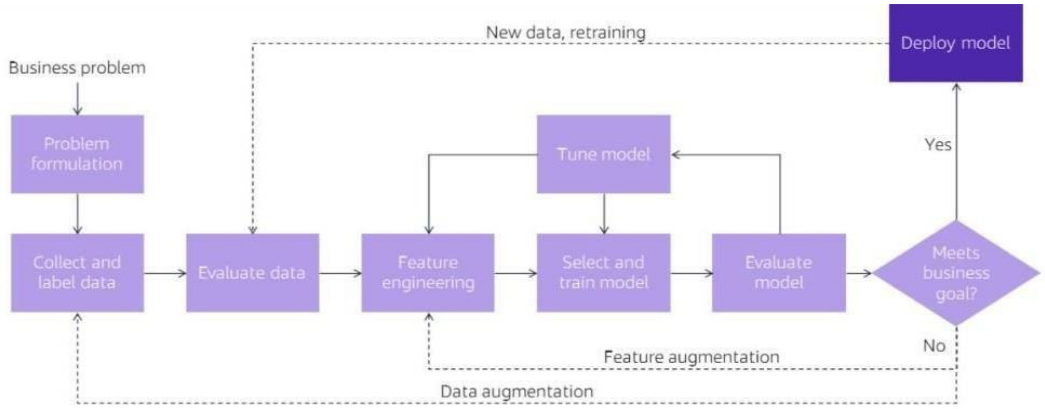


Figure-2.2-Ml Pipeline

4.Machine Learning Tools Overview

- Jupyter Notebook is an open-source web application that enables you to create and share documents that contain live code, equations, visualizations, and narrative text.
- Jupyter Lab is a web-based interactive development environment for Jupyter notebooks, code, and data. Jupyter Lab is flexible.
- pandas is an open-source Python library. It's used for data handling and analysis. It represents data in a table that is similar to a spreadsheet. This table is known as a panda Data Frame.
- Matplotlib is a library for creating scientific static, animated, and interactive visualizations in Python. You use it to generate plots of your data later in this course.

5. Machine Learning Challenges

NumPy is one of the fundamental scientific computing packages in Python. It contains functions for N-dimensional array objects and useful math functions such as linear algebra, Fourier transform, and random number capabilities. -learn is an open-source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data pre-processing, model selection and evaluation, and many other utilities.

MODULE 2: IMPLEMENTING A MACHINE LEARNING PIPELINE WITH AMAZON SAGEMAKER

1. Formulating Machine Learning Problems

Business problems must be converted into an ML problem. Questions to ask include –

- Have we asked why enough times to get a solid business problem statement and know why it is important?
- Can you measure the outcome or impact if your solution is implemented?

Most business problems fall into one of two categories –

- Classification (binary or multi): Does the target belong to a class?
- Regression: Can you predict a numerical value?

2. Collecting and Securing Data

- **Private data** is data that you (or your customers) have in various existing systems. Everything from log files to customer invoice databases can be useful, depending on the problem that you want to solve. In some cases, data is found in many different systems.
- **Open-source data** comprises many different open-source datasets that range from scientific information to movie reviews. These datasets are usually available for use in research or for teaching purposes. You can find open-source datasets hosted by AWS, Kaggle, and the UC Irvine Machine Learning Repository.
- **Securing Data** can be managed using AWS Cloud Trial that tracks the user activity and monitor the API and detects if it is unusual.

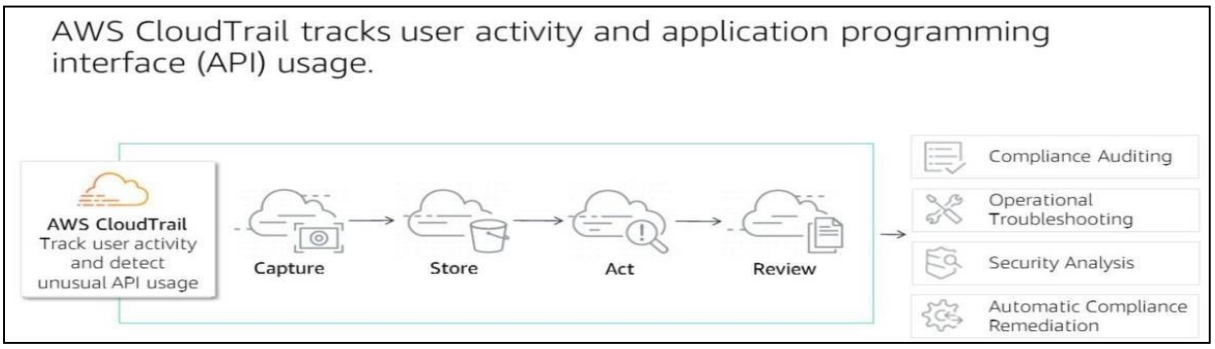


Figure-2.3-AWS Cloud Trail

3. Evaluating Data

- **Descriptive statistics** can be organized into different categories. Overall statistics include the number of rows (instances) and the number of columns (features or attributes) in your dataset. This information, which relates to the dimensions of your data, is important. For example, it can indicate that you have too many features, which can lead to high dimensionality and poor model performance.
- **Attribute statistics** are another type of descriptive statistic, specifically for numeric attributes. They give a better sense of the shape of your attributes, including properties like the mean, standard deviation, variance, minimum value, and maximum value.
- **Multivariate statistics** look at relationships between more than one variable, such as correlations and relationships between your attributes.

4. Feature Engineering

Feature selection is about selecting the features that are most relevant and discarding the rest. Feature selection is applied to prevent either redundancy or irrelevance in the existing features, or to get a limited number of features to prevent overfitting.

Feature extraction is about building up valuable information from raw data by reformatting, combining, and transforming primary features into new ones. This transformation continues until it yields a new set of data that can be consumed by the model to achieve the goals.

Outliers:

During feature engineering. You can handle outliers with several different approaches. They include, but are not limited to:

- **Deleting the outlier:** This approach might be a good choice if your outlier is based on an artificial error. Artificial error means that the outlier isn't natural and was introduced because of some failure—perhaps incorrectly entered data.
- **Imputing a new value for the outlier:** You can use the mean of the feature, for instance, and impute that value to replace the outlier value. Again, this would be a good approach if an artificial error caused the outlier.

Feature Selection: Filter Methods

Filter methods (figure-2.4) use a proxy measure instead of the actual model's performance. Filter methods are fast to compute, and they still capturing the usefulness of the feature set. Common measures include:

- **Pearson's correlation coefficient** –Measures the statistical relationship or association between two continuous variables.
- **Linear discriminant analysis (LDA)** –Is used to find a linear combination of features that separates two or more classes.

Feature Selection: Wrapper Methods

- **Forward selection** starts with no features and adds them until the best model is found. (figure-2.5)
- **Backward selection** starts with all features, drops them one at a time, and selects the best model.

Feature Selection: Embedded Methods

Embedded methods(figure-2.6) combine the qualities of filter and wrapper methods. They are implemented from algorithms that have their own built-in feature selection methods.

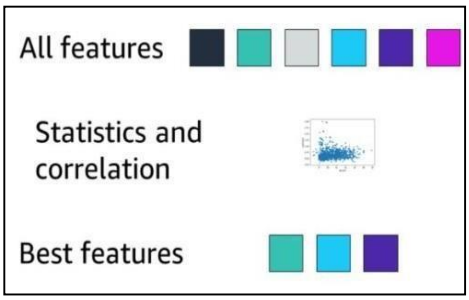


Fig-2.4 Filter Method

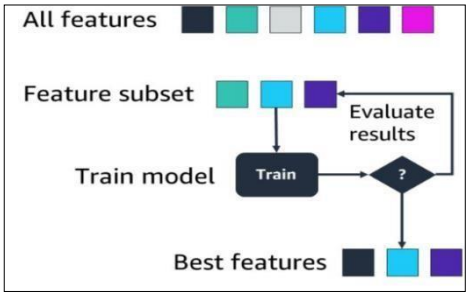


Fig-2.5Wrapper Method

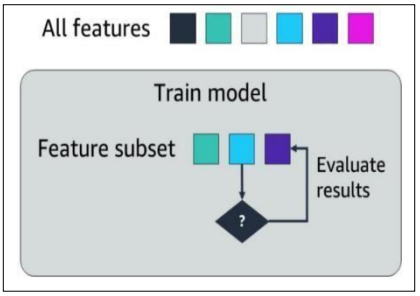


Fig-2.6Embedded Method

5. Training

Holdout technique(figure-2.7) and k-fold cross validation(figure-2.8) methods are the most used ones when the data is to be classified as test set and training set.

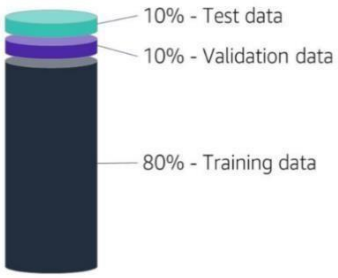


Figure-2.7-Holdout

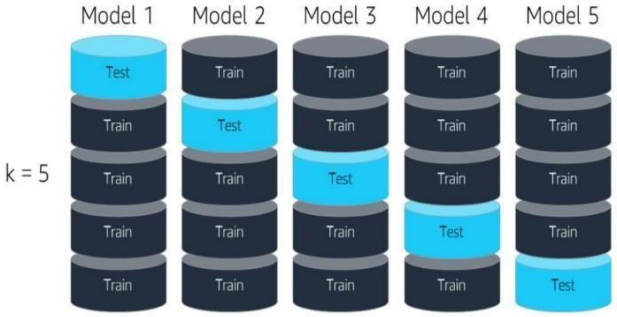


Figure-2.8 K-fold cross validation

LINEAR LEARNER: The Amazon SageMaker linear learner algorithm provides a solution for both classification and regression problems. The Amazon SageMaker linear learner algorithm compares favourably with methods that provide a solution for only continuous objectives. It provides a significant increase in speed over naive hyperparameter optimization techniques.

6. Hosting and Using the Model

- You can deploy your trained model by using Amazon SageMaker to handle API calls from applications, or to perform predictions by using a batch transformation.
- Use Single-model endpoints for simple use cases and use multi-model endpoint support to save resources when you have multiple models to deploy.

7. Evaluating the Accuracy of the Model

Confusion Matrix Terminology: Confusion Matrix is a performance measurement for machine learning classification. An example for classification(figure-2.9).

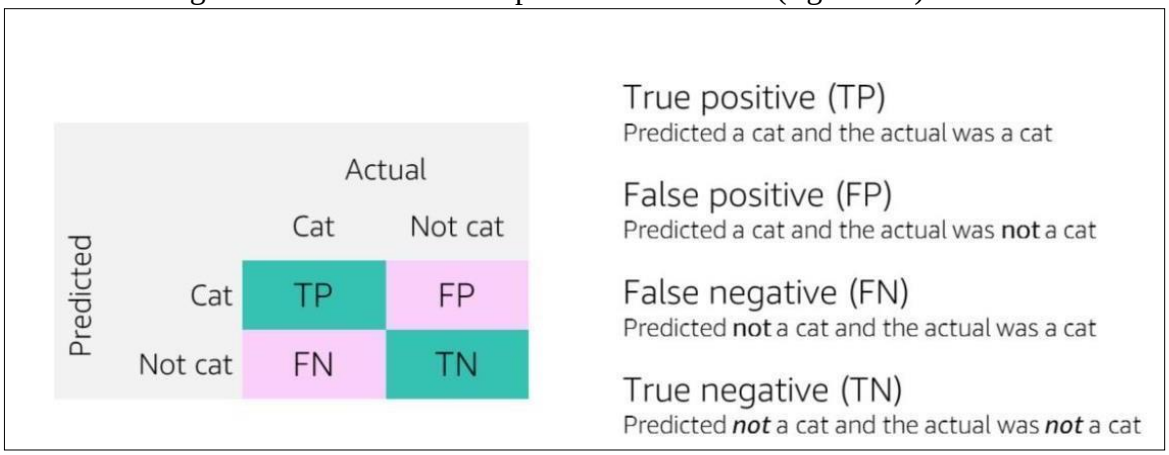


Figure-2.9-Confusion Matrix

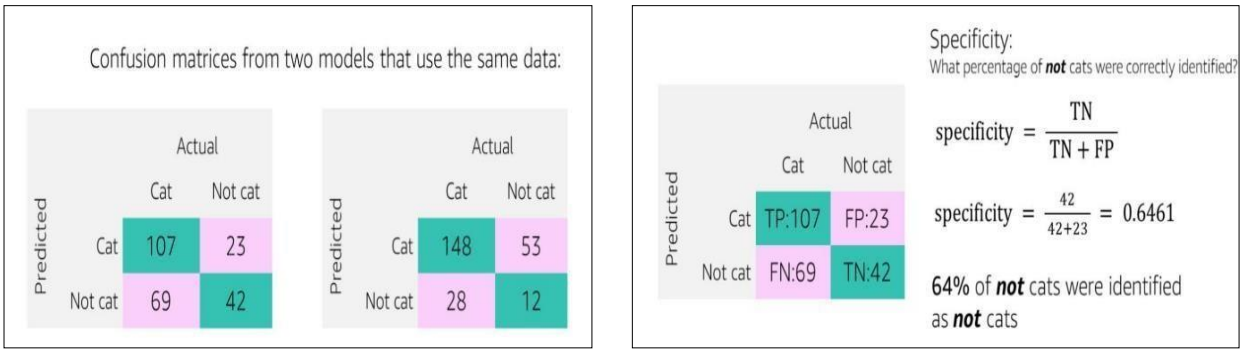


Figure-2.10- Specificity

8. Hyperparameter and model tuning

HYPERPARAMETER TUNING:

- Tuning hyperparameters can be labour-intensive. Traditionally, this kind of tuning was done manually.
- Then, they would train the model and score it on the validation data. This process would be repeated until satisfactory results were achieved.
- This process is not always the most thorough and efficient way of tuning your hyperparameters. It helps the model to define and filtering, and optimizer for finding patterns and defining the attributes of data by itself(figure-2.11).

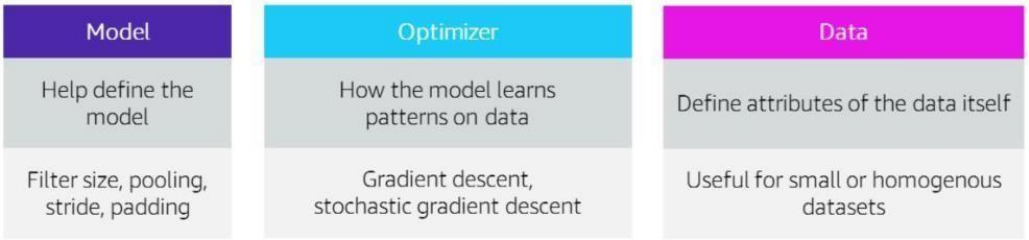


Figure-2.11-Model tuning

LAB: Implementing a Machine Learning pipeline with Amazon SageMaker.

- **Amazon SageMaker, Creating and Importing Data.**

1. Launch an Amazon SageMaker notebook instance.
2. Launch a Jupyter notebook.
3. Run code in a notebook.
4. Download data from an external source.
5. Upload and download a Jupyter notebook to your local machine.

Exploring Data

1. From the uploaded data use some functions in python like dtypes() -for describing the data types of each variable used.
2. describe() function is used to find the statistical insights like “mean”, “standard deviation”, “min, max”, “count”, “quartiles” etc..
3. dataframe.plot() is used for the visualization.

Encoding Categorical Data

1. df.info() for the getting the dtypes and df.columns() for column names
2. For encoding ordinal features firstly we use df[“column name”].value_counts(). Apply mapper using replace method i.e. df[“new col”]=df[“col1”].repalce(col2)
3. Use get_dummies() to add binary features to the required columns of the data frame.
4. Now using info() to get the encoded data.

Training a Model

1. Step:1 import data that is required to train.
2. Step:2 import boto3 and from sagemaker.image_uris import retrieve and import sagemaker and apply format changes to the imported data. Now explore the data.

Deploying Model

1. Import the necessary libraries, perform predictions using df.predict() function to the test part of the row.
2. To delete the end point of the predictor use the function df.delete_endpoint(test data)
3. Now perform batch transform using the botos3 library and apply the key value pairs as the dictionary.
4. Convert the values to binary features using “.apply(binary_convert)” function to the transformed data.

MODULE: 3. INTRODUCING FORECASTING

1. OVERVIEW OF FORECASTING

Forecasting is an important area of machine learning. It is important because so many opportunities for predicting future outcomes are based on historical data. It's based on time series of data.

Time series data as falling into two broad categories.

The first type is univariate, which means that it has only one variable. The second type is multivariate.

In addition to these two categories, most time series datasets also follow one of the following patterns:

- Trend –A pattern that shows the values as they increase, decrease, or stay the same over time.
- Seasonal –A repeating pattern that is based on the seasons in a year.
- Cyclical –Some other form of a repeating pattern.
- Irregular –Changes in the data over time that appear to be random or that have no discernible pattern.

2. PROCESSING TIME SERIES DATA

During the time series data processing we need to check whether the data behaviour as forward filled, moving average, backward fill or interpolation. (Figure-2.12)

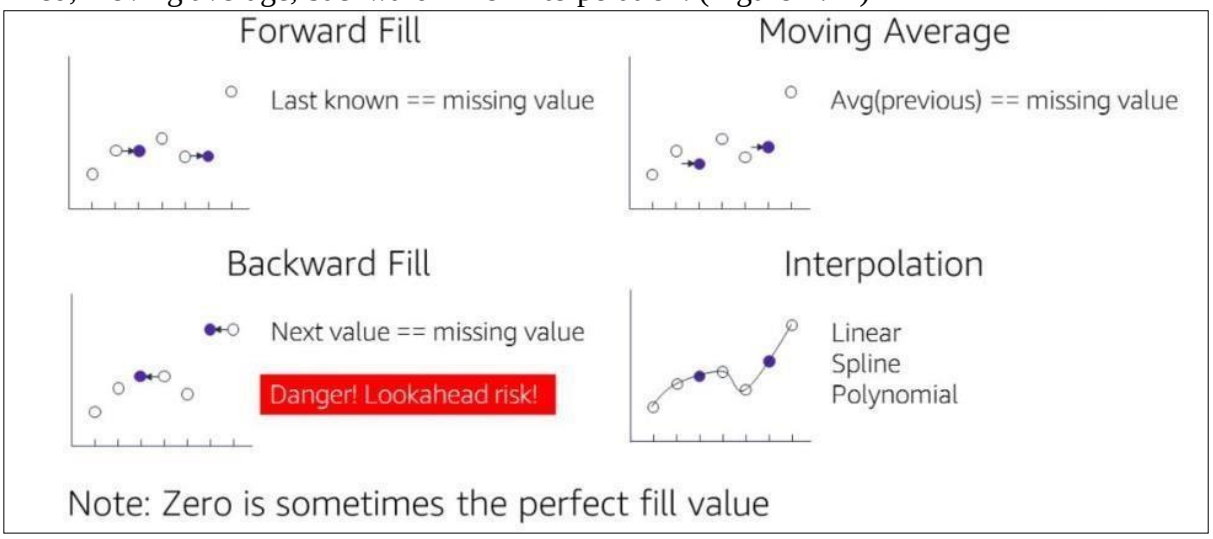


Figure-2.12-Time series data processing

- Time Series Data Handling: Smoothing of Data: Smoothing your data can help you deal with outliers and other anomalies. You might consider smoothing for the following reasons.
- Data preparation –Removing error values and outliers.
- Visualization –Reducing noise in a plot.

Time Series Data Algorithms: There are 5 types of Time Series Data Algorithms consisting of ARMA, DeepAR+, ETS, NPTS, Prophet as shown in figure-2.13.

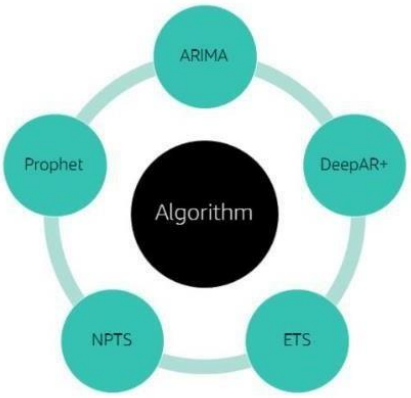


Figure-2.13-Time series data algorithms

- **Autoregressive Integrated Moving Average (ARIMA):** This algorithm removes autocorrelations, which might influence the pattern of observations.
- **Deep AR+: A** supervised learning algorithm for forecasting one-dimensional time series. It uses a recurrent neural network to train a model over multiple time series.
- **Exponential Smoothing (ETS):** This algorithm is useful for datasets with seasonality. It uses a weighted average for all observations. The weights are decreased over time.
- **Non-Parametric Time Series (NPTS):** –Predictions are based on sampling from past observations. Specialized versions are available for seasonal and climatological datasets.
- **Prophet:** A Bayesian time series model. It's useful for datasets that span a long time period, have missing data, or have large outliers.

3. Using Amazon Forecast

Below flowchart (figure-2.14) describes about the forecasting steps:

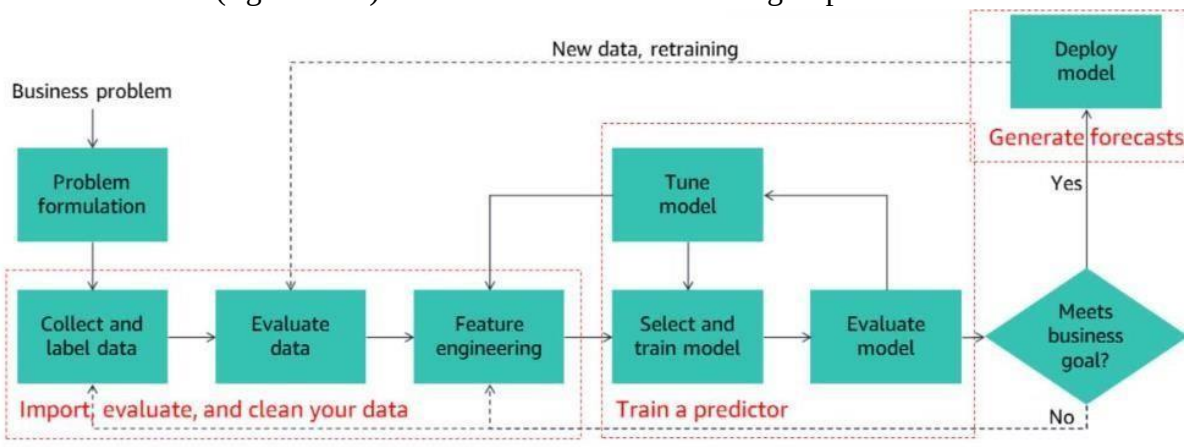


Figure-2.14-Amazon forecast flowchart

Import your data –You must import as much data as you have—both historical data and related data. You should do some basic evaluation and feature engineering before you use the data to train a model.

Train a predictor –To train a predictor, you must choose an algorithm. If you are not sure which algorithm is best for your data, you can let Amazon Forecast choose by selecting Auto ML as your algorithm. You also must select a domain for your data, but if you're not sure which domain fits best, you can select a custom domain. Domains have specific types of data that they require. For more information, see Predefined Dataset Domains and Dataset Types in the Amazon Forecast documentation.

Generate forecasts –As soon as you have a trained model, you can use the model to make a forecast by using an input dataset group. After you generate a forecast, you can query the forecast, or you can export it to an Amazon Simple Storage Service (Amazon S3) bucket. You also have the option to encrypt the data in the forecast before you export it.

LAB: Creating Forecast with Amazon Forecast.

Importing python packages in Jupyter notebook

1. Import boto3 (AWS SDK for python) and import warnings add use function warnings.filterwarnings('ignore').
2. Import pandas for data frames and matplotlib for the visualization and for plotting functions. 3. Import helper functions like "time", "sys", "os", "io" "json".

Read the file formats like .csv/.xlsx etc. and convert into time series.

1. Using pd.read_excel('file.xlsx'). Now use df.dropna() to remove the missing values from the dataset or use XGBOOST algorithm to deal with missing values.
2. Now convert dataset using column that contains dates to time series data set. Using "pd.to_datetime()".

Cleaning and reducing the size of the data

1. In this task we need to select the data that is unique. Let the unique data be x (through column). 2. Now using the function "x.unique()" , now the redundancy data gets deleted.

Examining the required code and removing anomalies

1. Using data.requiredcode.describe() we can quickly verify the dataset.
2. Use describe() and plot() for changes in the metrics.

Splitting the data

1. Split the data into 2 or more samples that contains columns that are correlated.
2. Each split parts of pairs should be assigned to separate variables.

Down sampling and Forecasting

1. Using the resample function from pandas we can make the cumulative summation.
2. Using the groupby() function and using the .create_predictor we create the "predictor" and using "create_forecast " we create forecast : >predictor_arn = create_predictor_response['PredictorArn']

Forecast completion

1. Using forecast we just need to use the service name in the query_forecast(forecasrArn=forest_arn). And plot the results using stock code.
2. Now we can forecast to our datasets.

MODULE 4: INTRODUCING COMPUTER VISION

1. Computer Vision enables machines to identify people, places, and things in images with accuracy at or above human levels, with greater speed and efficiency. Often built with deep learning models, computer vision automates the extraction, analysis, classification, and understanding of useful information from a single image or a sequence of images. The image data can take many forms, such as single images, video sequences, views from multiple cameras, or three-dimensional data.

Applications of Computer Vision:

Public safety and home security (figure-2.15).

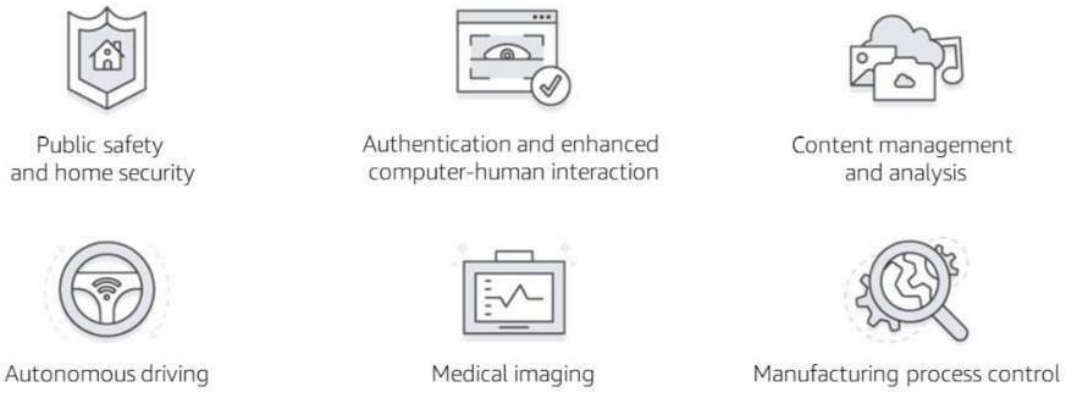


Figure-2.15-CV Applications

Computer vision with image and facial recognition can help to quickly identify unlawful entries or persons of interest. This process can result in safer communities and a more effective way of deterring crimes.

Authentication and enhanced computer-human interaction:

Enhanced human-computer interaction can improve customer satisfaction. Examples include products that are based on customer sentiment analysis in retail outlets or faster banking services with quick authentication that is based on customer identity and preferences.

Content management and analysis:

Millions of images are added every day to media and social channels. The use of computer vision technologies—such as metadata extraction and image classification—can improve efficiency and revenue opportunities.

Autonomous driving:

By using computer-vision technologies, auto manufacturers can provide improved and safer self-driving car navigation, which can help realize autonomous driving and make it a reliable transportation option.

Medical imaging:

Medical image analysis with computer vision can improve the accuracy and speed of a patient's medical diagnosis, which can result in better treatment outcomes and life expectancy. Manufacturing process control Well-trained computer vision that is incorporated into robotics can improve quality assurance and operational efficiencies in manufacturing applications. This process can result in more reliable and cost-effective products.

Computer vison problems:

Problem 01: Recognizing food & state whether it's breakfast or lunch or dinner
As the CV classified the objects as milk, peaches, ice cream, salad, nuggets, bread roll thus it's a breakfast (figure-2.16).
Problem 02: Video Analysis
2. Image and Video Analysis: Amazon Recognition is a computer vision service based on deep learning. You can use it to add image and video analysis to your applications (figure-2.17) Amazon Recognition enables you to perform the following types of analysis:

- **Searchable image and video libraries**–Amazon Recognition makes images and stored videos searchable so that you can discover the objects and scenes that appear in them.

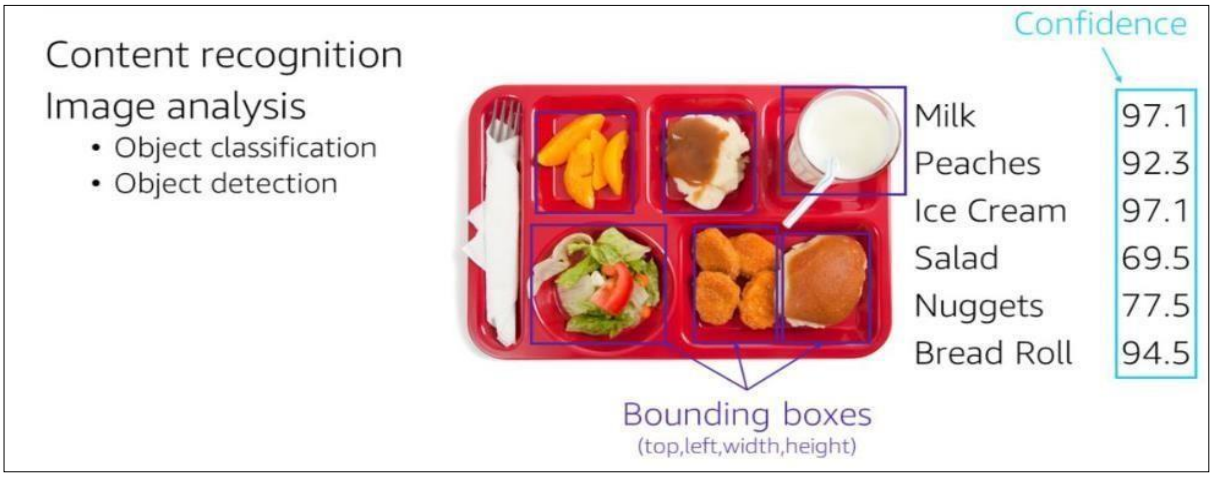


Figure-2.16-Problem 1

Figure-2.17-Problem 2

CASE 01: Searchable Image Library

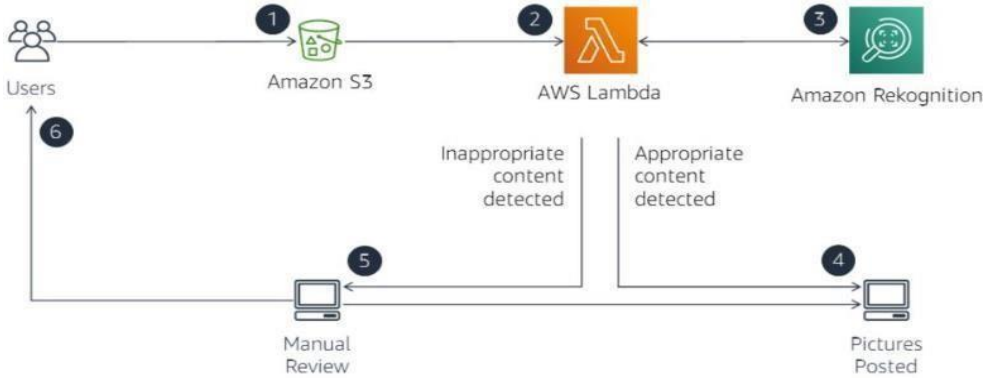


Figure-2.18

CASE 02: Sentiment Analysis

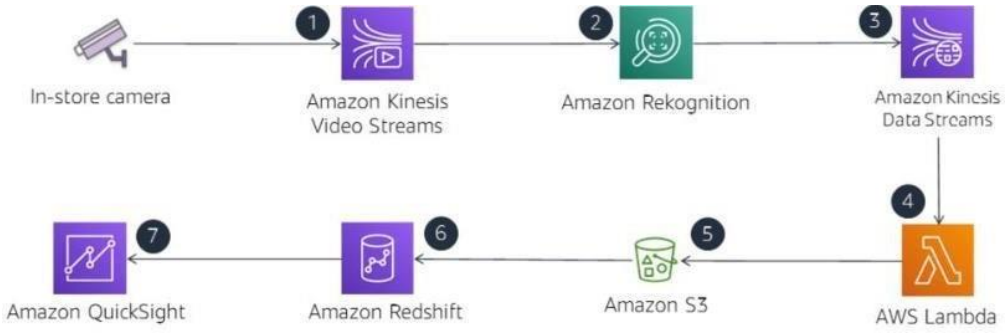
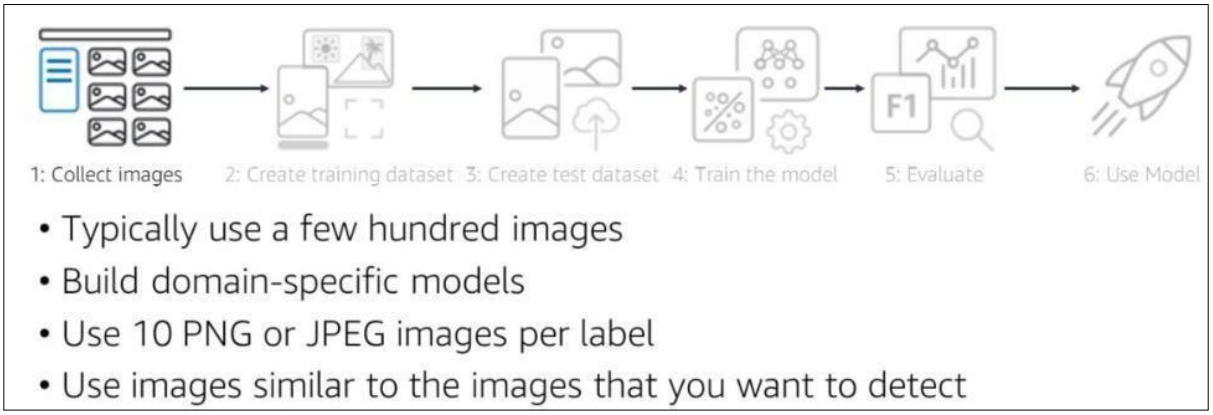


Figure-2.19

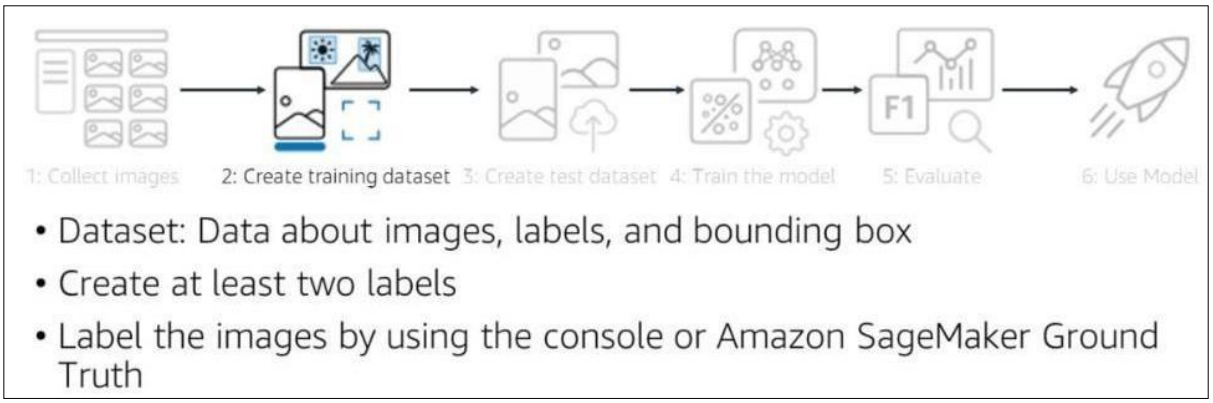
4. Preparing Customs Dataset for Computer Vison

There are 6 steps involved in preparing customs data:
Each step has it's functionalities like collection of images, creating training dataset, create test dataset, train the model, evaluate and then use the model.

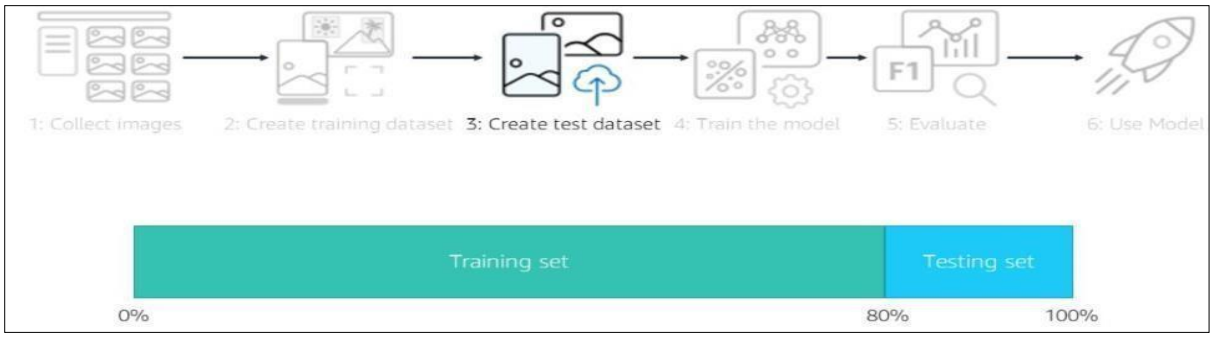
STEP 01: Collect Images



STEP 02: Create Training Dataset

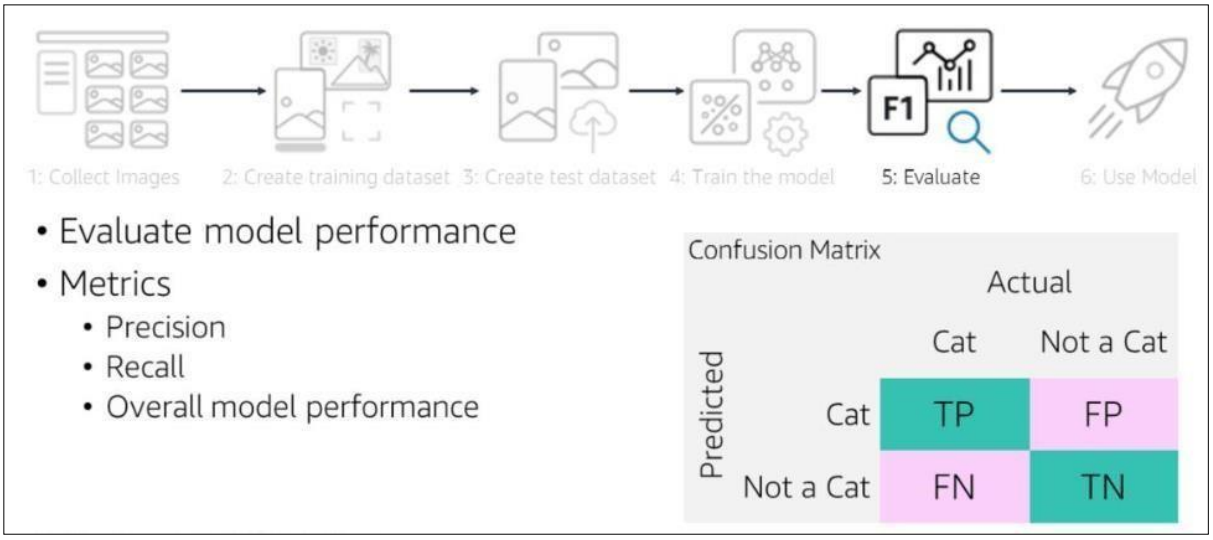


STEP 03: Create Test Dataset

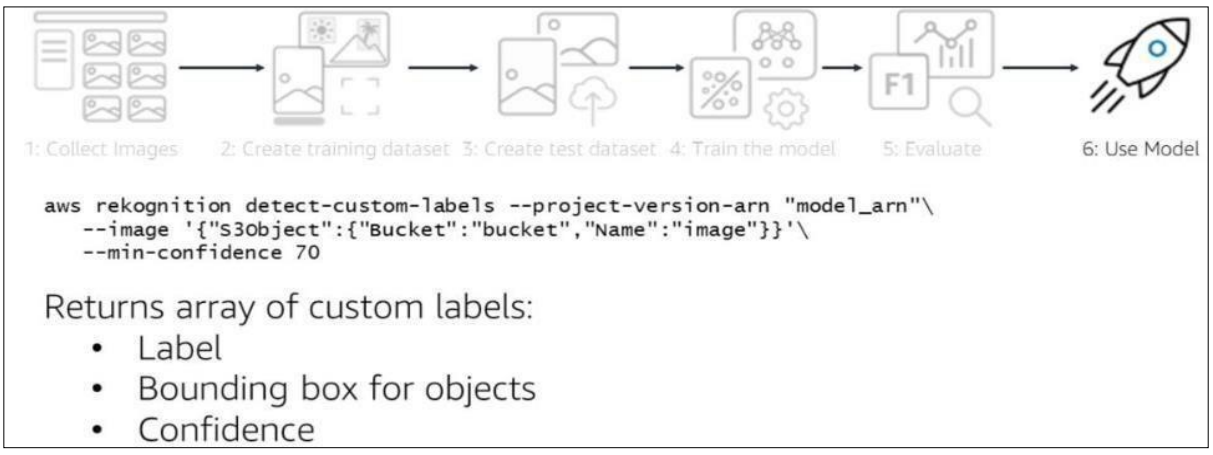


STEP 04: Train the Model

STEP 05: Evaluate



STEP 06: Use Model



LAB: Facial Recognition

Importing required libraries

1. Import necessary libraries like “from skimage import io”, “from skimage.transform import rescale”, “from matplotlib import pyplot as plt”.
2. Now “import boto3”, “import numpy as np”, “from PIL import Image, ImageDraw, ImageColor, ImageOps.

Creating a collection

1. Client =boto3.client(‘recognition’)
2. Collection_id = ‘collection’
3. Response = client.create_collection(CollectionId=collection_id)

Uploading an image to search

1. Use the io.imread(“image file”) and use im.show(“filename”) to show image.
2. Rescale the image size using “filename=rescale(filename,0.50, mode=’constant’)”

Adding image to the collection

1. Using stock code add the image data to the collection.
2. Now the objects are created.

Viewing the bounding box of the detected face

1. Set a variable img as image.open(filename)
2. Add a variables like imgwidth and imgheight =img.size
3. Use a loop to set the bounding box with imgwidth, imgheight for top, left, width, height.

Listing and finding the faces in the collection

1. Use the forecast code from the client-service name.
2. Set the “targetimage” for the images in the collection excluding the search image. Then threshold value and search images to count at once.
3. Draw a box around the discovered face among the collection suing the same stock code for bounding.
4. To reset and delete the collection data from the client use the “delete_collection” in exceptions handling functions (try and except) for the clear display of status code.

MODULE 6: INTRODUCING NATURAL LANGUAGE PROCESSING

1. Overview of Natural Language Processing

NLP develops computational algorithms to automatically analyse and represent and represent human language. By evaluating the structure of language, machine learning systems can process large sets of words, phrases, and sentences(figure-2.20)

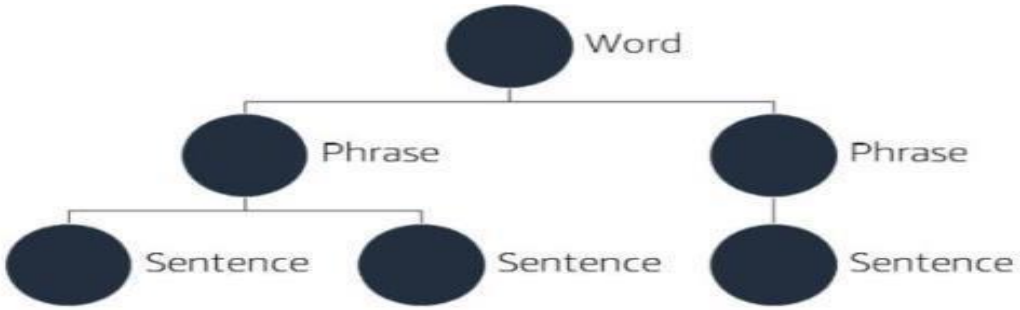


Figure-2.20-Structure of NLP

Some challenges of NLP:

- **Discovering the structure of the text** –One of the first tasks of any NLP application is to break the text into meaningful units, such as words, phrases, and sentences.
- **Labelling data** –After the system converts the text to data, the next challenge is to apply labels that represent the various parts of speech. Every language requires a different labelling scheme to match the language's grammar.
- **Representing context** –Because word meaning depends on context, any NLP system needs a way to represent context. It is a big challenge because of the large number of contexts.
- **Applying grammar** –Dealing with the variation in how humans use language is a major challenge for NLP systems.

NLP FLOW CHART: NLP flow chart starts with collection of test database as shown in figure-2.21. Then the test data gets tokenize using word vector coding and further it gets analysed and use model for prediction of results.

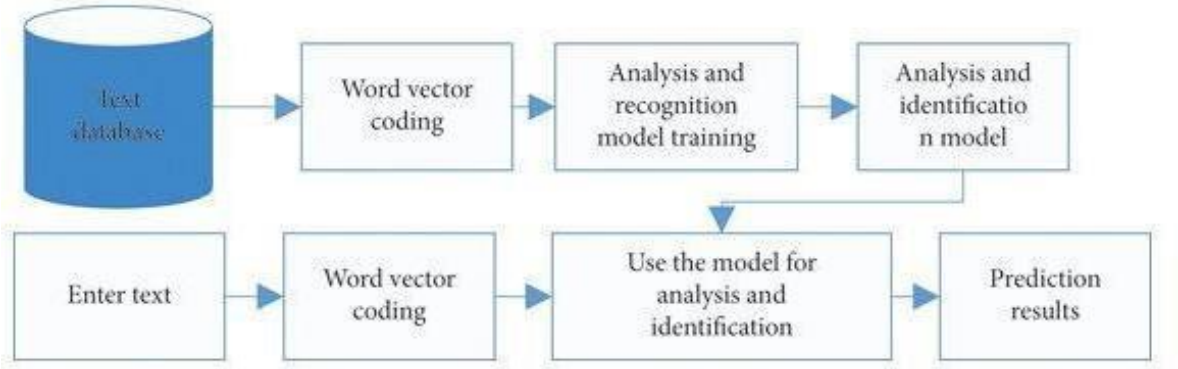
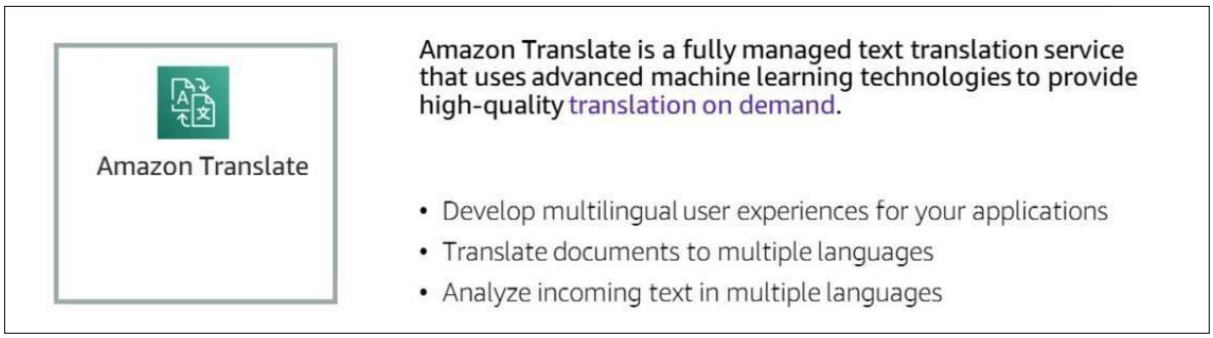


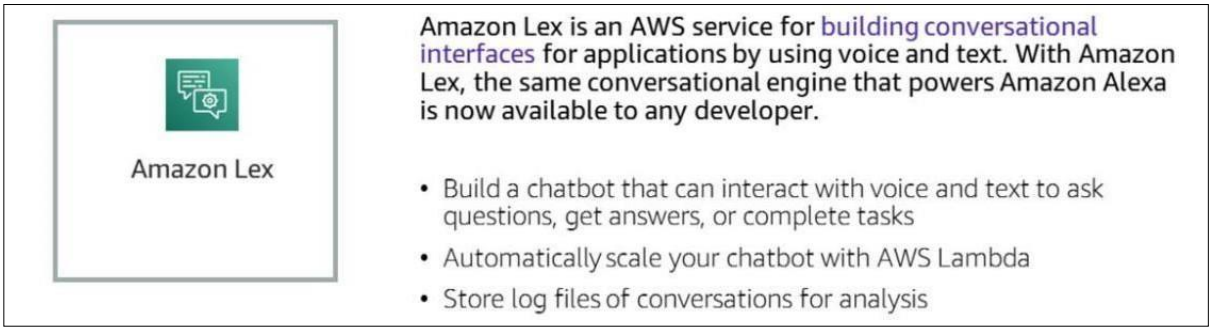
Figure-2.21-NLP flow chart

2. Natural Language Processing Managed services

Uses:



- International Websites
- Software Localisation



- Interactive Assistants
- Database Queries

LAB: Natural Language Processing

1. To create an Amazon Lex bot (console)
2. Sign into the AWS Management Console and open the Amazon Lex console at https://console.aws.amazon.com/lex/.
3. If this is your first bot, choose Get Started; otherwise, on the Bots page, choose Create.
   - On the Create your Lex bot page, provide the following information, and then
   - default bot name (Order Flowers).
   - For COPPA, choose No.
   - For User utterance storage, choose the appropriate response.
   - Choose Create.
4. The console makes the necessary requests to Amazon Lex to save the configuration.
5. The console then displays the bot editor window.
6. Wait for confirmation that your bot was built.
7. Test the bot**.**

CASE STUDY

UNLOCKING CLINICAL DATA FROM NARRATIVE REPORTS

Objective:

To evaluate the automated detection of clinical conditions described in narrative reports using Natural Language Processing. **Design:**

Using the automated methods NLP analysing 200 patients' clinical records data. **Study Subjects:**

A computerized, general-purpose natural language processor; 6 internists; 6 radiologists; 6 lay persons; and 3 other computer methods. **Introduction:**

Though there really are no limits to how NLP can support the healthcare industry, let's look at three primary use cases:

- **Improving Clinical Documentation:** Rather than waste valuable time manually reviewing complex EHR (Electronic Health Report), NLP uses speech-to-text dictation and formulated data entry to extract critical data from EHR at the point of care. This not only enables physicians to focus on providing patients with the essential care they need, it also ensures that clinical documentation is accurate and kept up to date.
- **Accelerating Clinical Trial Matching:** Using NLP, healthcare providers can automatically review massive quantities of unstructured clinical and patient data and identify eligible candidates for clinical trials. Not only does this enable patients to access experimental care that could dramatically improve their condition — and their lives — it also supports innovation in the medical field.
- **Supporting Clinical Decisions:** NLP makes it fast, easy, and efficient for physicians to access health-related information exactly when they need it, enabling them to make more informed decisions at the point of care.

Main Outcome Measures:

Intrasubject disagreement was quantified by using specificity (patients without disease) and sensitivity (patients with disease) differences between NLP and physicians' analysis.

Results:

Using a majority vote, physicians detected 101 conditions in the 200 reports (0.51 per report); the most common condition was acute bacterial pneumonia (prevalence, 0.14), and the least common was chronic obstructive pulmonary disease (prevalence, 0.03). Pairs of physicians disagreed on the presence of at least 1 condition for an average of 20% of reports. The average intrasubject distance among physicians was 0.24 (95% CI, 0.19 to 0.29) out of a maximum possible distance of 6. No physician had a significantly greater distance than the average. The average distance of the natural language processor from the physicians was 0.26 (CI, 0.21 to 0.32; not significantly greater than the average among physicians). Lay persons and alternative computer methods had significantly greater distance from the physicians (all >0.5). The natural language processor had a sensitivity of 81% (CI, 73% to 87%) and a specificity of 98% (CI,
97% to 99%); physicians had an average sensitivity of 85% and an average specificity of 98%.

Conclusions:

Physicians disagreed on the interpretation of narrative reports, but this was not caused by outlier physicians or a consistent difference in the way internists and radiologists read reports. The natural language processor was not distinguishable from the physicians and was superior to all other comparison subjects. Although the domain of this study was restricted (six clinical conditions in chest radiographs), natural language processing seems to have the potential to extract clinical information from narrative reports in a manner that will support automated decision-support and clinical research.

Proposed Solution:

Clinical Assertion Model:

Clinical assertion modeling enables healthcare providers to analyze clinical notes and identify whether a patient is experiencing a problem, and whether that problem is present, absent, or conditional. For this reason, clinical assertion models are often used to help diagnose and treat patients.

For example, a patient might tell her doctor that she's experienced a headache for the past two weeks and feels anxious when she walks fast. After examining the patient, the doctor might note that she has no symptoms of alopecia and that she doesn't appear to be in any pain.

The doctor could later use a combination of NER and text classification to analyze their clinical from that appointments and flag "headache," "anxious," "alopecia," and "pain" as PROBLEM entities. From there, the doctor could further categorize those problems by making assertions as to whether they were present, conditional, or absent — in this case, the headache would be present, anxiousness would be conditional, and alopecia and pain would be absent.

As you can see based on this example, this application of NLP in healthcare enables physicians to optimize patient care by identifying which problems are most pressing and administering immediate treatment.

CONCLUSION

These modules described how model explain ability relates to AI/ML solutions, giving customers insight to explain ability requirements when initiating AI/ML use cases. Using AWS, four pillars were presented to assess model explain ability options to bridge knowledge gaps and requirements for simple to complex algorithms. To help convey how these models explain ability options relate to real-world scenarios, examples from a range of industries were demonstrated. It is recommended that AI/ML owners or business leaders follow these steps when initiating a new AI/ML solution:

- Collect business requirements to identify the level of explain ability required for your business to accept the solution.
- Based on business requirements, implement an assessment for model explain ability.
- Work with an AI/ML technician to communicate model explain ability assessment and find the optimal AI/ML solution to meet your business objectives.
- After the solution is completed, revisit the model explain ability assessment to evaluate that business requirements are continuously met.
- By taking these steps, we will mitigate regulation risks and ensure trust in our model. With this trust, when the event comes to push your AI/ML solution into an AWS production environment, we will be ready to create business value for our use case

REFERENCES

1. Machine learning on AWS https://aws.amazon.com/machine-learning/?nc2=h_ql_sol_use_ml
2. Amazon _AWS _EC2
https://aws.amazon.com/ec2/
3. Amazon _AWS_S3
https://aws.amazon.com/s3/
4. Amazon_AWS_sagemaker
https://aws.amazon.com/sagemaker/
5. GitHub machine learning scikit learn
https://github.com/scikit-learn/scikit-learn.git
6. AWS_forecast
https://aws.amazon.com/forecast 7. Case_study_on_ML_architecture-uber
https://pantelis.github.io/cs634/docs/common/lectures/uber-ml-arch-case-study/
8. AWS Global Infrastructure Map: https://aws.amazon.com/about-aws/globalinfrastructure/#AWS_Global_Infrastructure_MapChoose a circle on the map to view summary information about the Region represented by the circle.
9. Regions and Availability Zones: https://aws.amazon.com/about-aws/globalinfrastructure/regions_az/Choose a tab to view a map of the selected geography and a list of Regions, Edge locations, Local zones, and Regional Caches.
10. Clinical Narrative reports.
https://www.acpjournals.org/doi/abs/10.7326/0003-4819-122-9-199505010-00007

CERTIFICATE