Low Level Design (LLD)

Adult Census Income Prediction

Revision Number: 1.5
Last date of revision: 07/7/2024

Jyoti Pandey

Document Version Control

Date Issued	Version	Description	Author
26th June 2024	1.1	First Draft	Jyoti Pandey
1st July 2024	1.2	Added Workflow charts	Jyoti Pandey
3rd July 2024	1.3	Added Proposed Solution	Jyoti Pandey
5th July 2024	1.4	Added Test Cases	Jyoti Pandey
7th July 2024	1.5	Added KPIs	Jyoti Pandey

Contents

Abstract

Data has always been the backbone of many important decisions. When an assumption is backed up by facts and numbers, the chances of incorrectness and bad decisions decrease and in today's world, Countless decisions in private and public sectors are based on Census data. Census data is the backbone of the democratic system of government, highly affecting the economic sectors. Census-related figures are used to distribute the federal funding by the government into different states and localities.

The above introduction had an aim to increase the awareness about how the income factor actually has an impact not only on the personal lives of people, but also an impact on the nation and its betterment. We will now have a look on the data extracted from the 1994 Census bureau database, and try to find insights about how different features have an impact on the income of an individual and also do some predictive analysis using the modern Data Science and Machine Learning techniques.

1 Introduction

1.1 Why this Low-Level Design Document?

The goal of LLD or a low-level design document is to give the internal logical design of the actual program code for Adult Census Income Prediction. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

1.2 Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work.

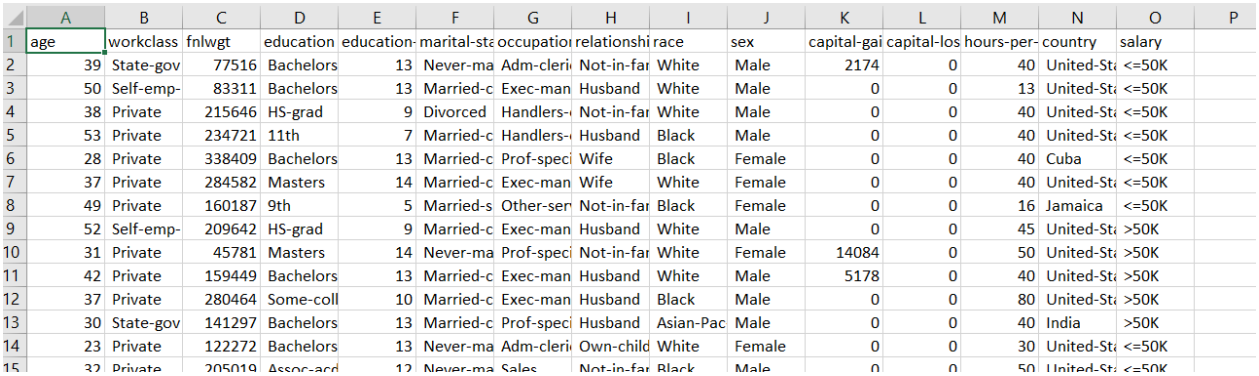
2 Technical specifications

2.1 Dataset

Features	Finalized	Source
Age	Yes	https://github.com/JyotiPandey111/Adult-Census-Income-Prediction/tree/main/data
Work Class	Yes	
Education	Yes	
Marital Status	Yes	
Occupation	Yes	
Relationship	Yes	
Race	Yes	
Sex	Yes	
Capital Gain	Yes	
Capital Loss	Yes	
Hours	Yes	
Country	Yes	

2.1.1 Income Dataset Overview

There are 32561 rows and 15 columns in the dataset.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	age	workclass	fnlwgt	education	education-	marital-sta	occupatior	relationshi	race	sex	capital-gai	capital-los	hours-per-	country	salary	
2	39	State-gov	77516	Bachelors	13	Never-ma	Adm-cleri	Not-in-far	White	Male	2174	0	40	United-Sta	<=50K	
3	50	Self-emp-	83311	Bachelors	13	Married-c	Exec-man	Husband	White	Male	0	0	13	United-Sta	<=50K	
4	38	Private	215646	HS-grad	9	Divorced	Handlers-	Not-in-far	White	Male	0	0	40	United-Sta	<=50K	
5	53	Private	234721	11th	7	Married-c	Handlers-	Husband	Black	Male	0	0	40	United-Sta	<=50K	
6	28	Private	338409	Bachelors	13	Married-c	Prof-speci	Wife	Black	Female	0	0	40	Cuba	<=50K	
7	37	Private	284582	Masters	14	Married-c	Exec-man	Wife	White	Female	0	0	40	United-Sta	<=50K	
8	49	Private	160187	9th	5	Married-s	Other-ser	Not-in-far	Black	Female	0	0	16	Jamaica	<=50K	
9	52	Self-emp-	209642	HS-grad	9	Married-c	Exec-man	Husband	White	Male	0	0	45	United-Sta	>50K	
10	31	Private	45781	Masters	14	Never-ma	Prof-speci	Not-in-far	White	Female	14084	0	50	United-Sta	>50K	
11	42	Private	159449	Bachelors	13	Married-c	Exec-man	Husband	White	Male	5178	0	40	United-Sta	>50K	
12	37	Private	280464	Some-coll	10	Married-c	Exec-man	Husband	Black	Male	0	0	80	United-Sta	>50K	
13	30	State-gov	141297	Bachelors	13	Married-c	Prof-speci	Husband	Asian-Pac	Male	0	0	40	India	>50K	
14	23	Private	122272	Bachelors	13	Never-ma	Adm-cleri	Own-child	White	Female	0	0	30	United-Sta	<=50K	
15	32	Private	205019	Assoc-acd	12	Never-ma	Sales	Not-in-far	Black	Male	0	0	50	United-Sta	<=50K	

2.2 Predicting Income

1. User enters the values for age, capital gain, capital loss, hours per week.
2. The User chooses the work class, education marital status, occupation, relationship, race sex, country by clicking on one of the available options.
3. The system presents the set of inputs required from the user.
4. The user gives required information.
5. The system should be able to predict whether income is greater than 50K dollars or not based on the user information.

2.3 Logging

We should be able to log every activity done by the user.

1. The System identifies at what step logging required
2. The System should be able to log each and every system flow.
3. Developers can choose logging methods. You can choose database logging/ File logging as well.
4. System should not be hung even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

2.4 Database

System needs to store every request into the database and we need to store it in such a way that it is easy to retrain the model as well.

1. User enters the values for age, capital gain, capital loss, hours per week.
2. The User chooses the work class, education marital status, occupation, relationship, race sex, country by clicking on one of the available options.
3. The system stores each and every data given by the user or received on request to the database. Database you can choose your own choice whether MongoDB/ MySQL. We have chosen MongoDB.

2.5 Deployment

1. AWS

3 Technology stack

Front End	HTML, Fast API
Backend	Python
Database	MongoDB
Deployment	AWS

4 Proposed Solution

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Try out different machine learning algorithms that's best fit for the above case.

The dataset named Adult Census Income is available in kaggle and UCI repository. This data was extracted from the 1994 census bureau dataset by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The prediction task is to determine whether a person makes over $50K a year or not.

5 Project Architecture Workflow

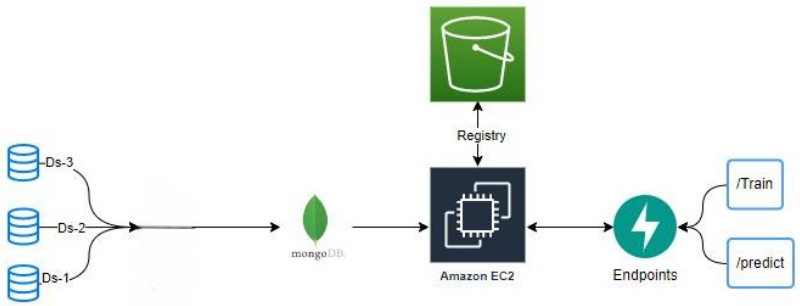


Fig 1: Data Ingestion

Data Ingestion:

Uploaded the data to MongoDB database and retrieved the data from MongoDB by creating connection in data ingestion component.

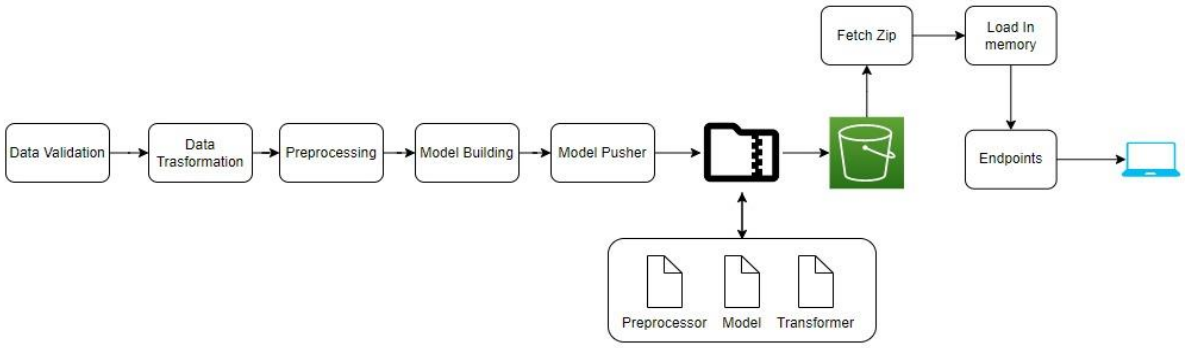


Fig 2 : Project Architecture

Data Validation:

In this component, data is validated by checking that all the columns are retrieved from MongoDB database.

Data Transformation:

In this component, data has checked and proper techniques are used to handle missing values, incorrect values, categorical feature handling, outliers handling, feature extraction, feature manipulation.

Model Builder:

In this component of project, feature selection has been done and a model is created which performs the best among all the classification Machine Learning algorithms.

Model Pusher:

This component is used after evaluating the best model in order to maintain the performance of model for near future and pushes the best model to system.

6 Deployment Architecture Workflow

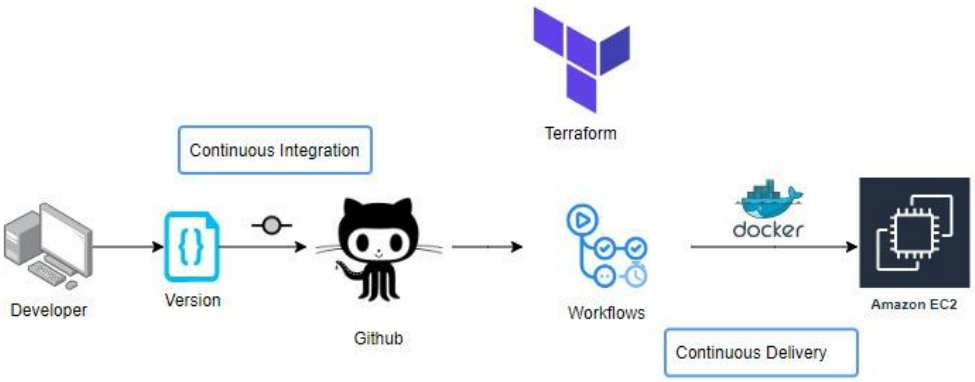


Fig 3: Deployment Architecture

7 Test cases

Test Case Description	Pre-Requisite	Expected Result
Verify whether the Application URL is accessible to the user	Application URL should be defined	Application URL should be accessible to the user
Verify whether the Application loads completely for the user when the URL is accessed	1. Application URL is accessible 2. Application is deployed	Application URL should be defined
Verify whether user is able to see input fields.	Application is accessible	User should be able to see input fields
Verify whether user is able to edit all input fields	Application is accessible	User should be able to edit all input fields
Verify whether user gets Submit button to submit the inputs	Application is accessible	User should get Submit button to submit the inputs
Verify whether user is presented with results on clicking submit	Application is accessible	User should be presented with results on clicking submit
Verify whether the results are in accordance to the selections user made	Application is accessible	The results should be in accordance to the selections user made

8 Key performance indicators (KPI)

- Comparison of F1 score of model prediction value and actual value.
- User enters the values for age, capital gain, capital loss, hours per week.
- The User chooses the work class, education marital status, occupation, relationship, race, sex, country by clicking on one of the available options.
- Capital Loss and Capital Gain present or not. If present, is it low or high.