High-Level Design (HLD)

Adult Census Income Prediction

Written By	SUDHANSU GOUDA
Version	1.0
Date	25/02/2023

Document Change Control Record

Version	Date	Author	Comments

Review

Version	Date	Reviewer	Comments

Approval

Version	Review date	Reviewed By	Approved By	Comments

Abstract

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted. Certain attributes of each product and store have been defined. The aim is to build a predictive model and find out whether the income of the person is greater than or less than $50,000. Using this model we will get insights as well as we get to determine the income group.

1. Introduction

1.1 Why these High-Level Design Documents?

What does a high level document mean?

A high-level design document (HLDD) describes the architecture used in the development of a particular software product. It usually includes a diagram that depicts the envisioned structure of the software system. Since this is a high-level document, non-technical language is often used. The HLD will be:

- Present all of the design aspects and define them in detail.
- Describe the user interface being implemented.
- Describe the needed Python libraries for the coding.
- Describe the performance requirements.
- Include design features and the architecture of the project.
- List and describe the non-functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application Compatibility
    - Resource Utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture(layers), application flow (Navigation), and technology architecture, The HLD uses non-technical and mildly-technical terms which should be understandable to the administrators of the system

1.3 Definition

TERM	DESCRIPTION
DB	Database, the cloud platform where the data will be stored. Can be considered cloud storage.
ML	Machine Learning
API OR APIS	Application Programming Interface can be considered a website link from there can extract information.

2. General Description

2.1 Product Perspective

Adult Census income prediction gives us the ability to predict the category of income when given in the features that helps the ML algorithm to predict the income category of the given entities.

2.2 Problem Statement

The aim is to build a predictive model and find out the category of income of each person. Using this model, we will try to understand the properties of persons age, job role, marital status and working hours which play a key role in understanding how the income varies based on these properties.

2.3 Proposed Solution

We will use Exploratory Data Analysis (EDA) to find the Key relations between different attributes and will use a ML algorithm to predict the future sales demand. We can tell the individual using this to gain an insight about which job is in demand or which work is high paying based on work hours and which job they take for the long term.

2.4 Data Requirements

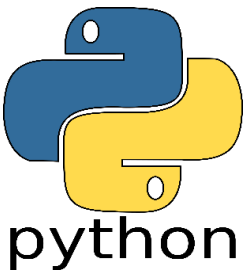
The data was provided by the iNeuron team in their dashboard. The Adult Census Income Prediction data recorded many people's job descriptions along with other insights. For building the ML model we will use the dataset that is given. The data is consisted of 32561 rows and various information about products like

- age: Age
- workclass: Work sector
- fnlwgt: Final weight
- eduction: Highest level of education
- education-num: Count of the value of education occurred

- marital-status: Current marital status
- occupation: Current job
- relationship: Relation in married life or else not in family
- race: Race of the person
- sex: Gender
- capital-gain: Profit earned on the sale of an asset like stocks, bonds or real estate.
- capital-loss: Loss incurred on the sale of an asset like stocks, bonds or real estate.
- hours-per-week: Sales of the product in the particular store.
- country: The country where the person is working currently
- salary: The salary category either more than $50,000 or less than $50,000

2.5 Tool Used

The programming language we used is python as Python is known to be the best programming language for data science, and it is commonly used by big tech companies for data science tasks. We are going to use some other python-based libraries such as NumPy and pandas for data Manipulation data cleaning and for some preprocessing tasks. To perform EDA, we will be switching between seaborn and matplotlib library. For model training we will use various classification-based Machine learning Algorithms such as random forest classifier,Logistic regression, KNeighbors classifier, Decision Tree classifier from the very famous Sci-kit learn library. After reaching a decent/good evaluation score we will then save the model using pickle Library. Now, for creating an app which we are further going to deploy we will be using Streamlit as our web framework , and for deployment we will then use Heroku cloud service to deploy our ML model.



2.6 Constraints

The System should be user-friendly, the user should get all proper messages while using the web app. The user also should get a proper error message if he/she has done something wrong on the web-app page. All the errors and results should be delivered in the easiest possible way and all the buttons that are going to be inserted on the web page should be labelled properly, so the user did not get confused using the system.

2.7 Assumptions

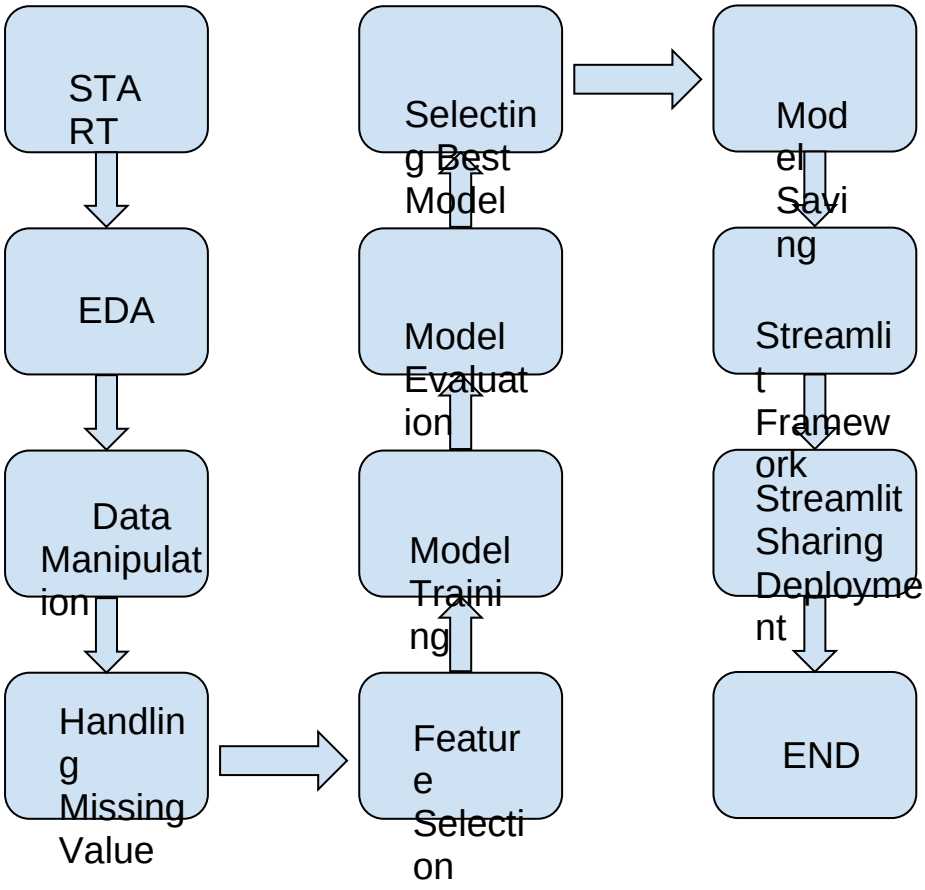
The main objective is to implement a system that will produce approximate future demand for a job role in his/her career.

3. Design Details

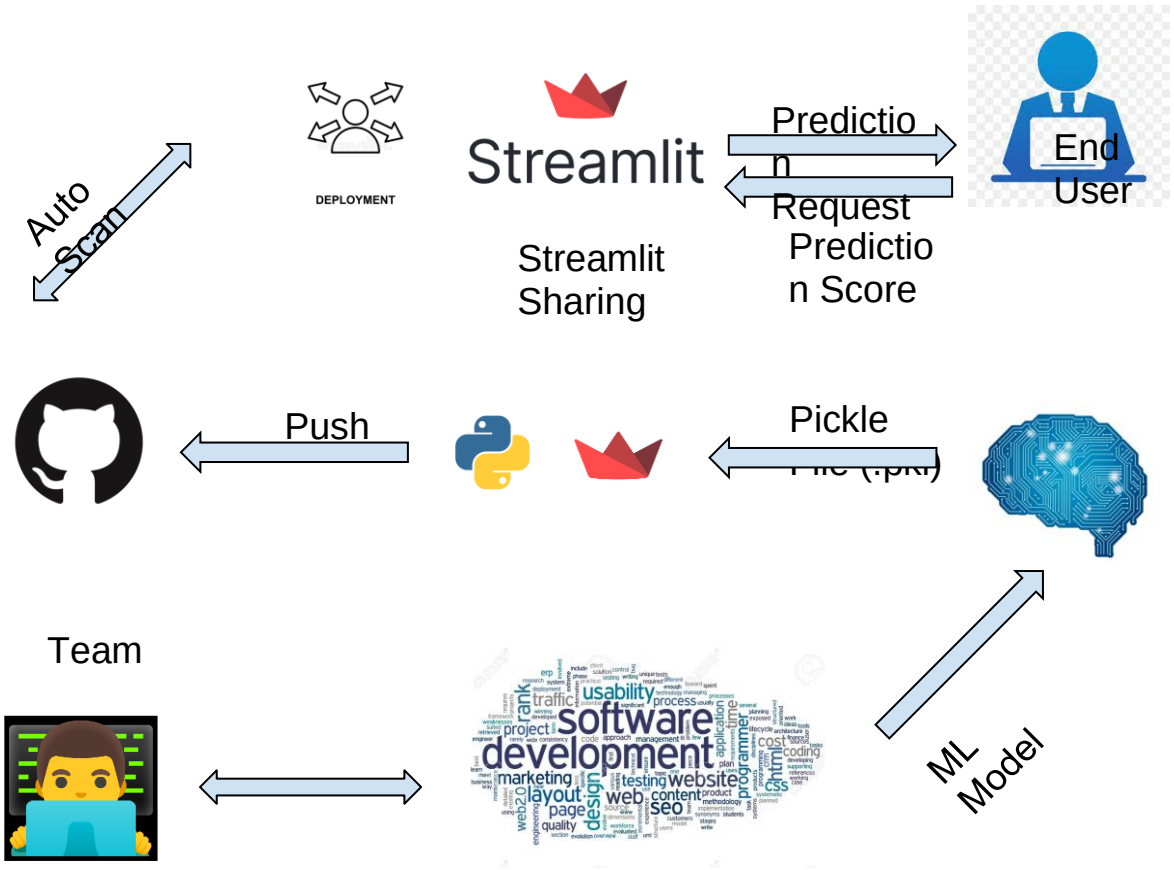
3.1 Process Flow

We will be using the following process flow for this project. The process flow will be based on End-To-End Machine learning workflow right from data gathering to model deployment.

Process Flow



3.2 Deployment Process



3.3 Error Handling

If any error occurred in the processing way, then the error message should be shown to the user in a completely non-technical way that can be understandable by any person. And Meaningful error messages should be shown, so the user can spot his mistake and re-run the process with improvement. All the errors that will occur should be handled properly. And we have to log every error for our application and have to manage the same.

4. Performance

The Adult Census Income Prediction is dependent on a particular ML algorithm that we choose. We will train various ML algorithms and will Evaluate the best fitting algorithm for predicting the target. Our system performance will be based on the data we are going to feed to the algorithms. And the performance will depend on the finalized model. and the web application and the deployment server. With all of these components, our program should run properly.

4.1 Reusability

The code and the module are created during the time of building the project should maintain all coding guidelines and full project code is written in a Modular fashion. Our system should have the flexibility to work properly from any location. And it should handle any improper input value from the user and should give a meaningful error message so the user can correct his/her mistake and enter valid input to get the result. And the system should be reusable in every manner with different types of inputs values that are all are it has been trained.

4.2 Application Compatibility

The different libraries and Python programming languages are used to build the system. Every library has its own functionality and it should work properly with our fluctuate system. Streamlit will be used for making the web app. All the components of the application should work properly and it should produce a result without any interpretation.

4.3 Resource Utilization

Our application should utilize the given resource properly and it should use a minimal amount of internet to work and call the web app. Our system should not use much computational resources hence it will make the application slow. Our application will be deployed on a cloud platform and it should utilize the resource given on the cloud and work properly.

5. Deployment

For the deployment process, we will be using Streamlit Sharing cloud to Deploy our model. Streamlit is used to make web apps as well as a cloud platform that lets companies build, deliver, monitor and scale apps.

6. Conclusion

The Income Predictor can tell whether a person with certain details can have a salary more than $50K annually or not. Income Predictor can further tell which job and how much work is needed to earn more than $50K annually and up until which age. According to which a candidate can plan on which career to pursue and what to expect from the work and how the candidate will manage his work life according to the desired job. Random forest Model had the highest accuracy among the other models we used, trained and after checking the accuracy.

7. Reference

Google images reference to showcase the framework/library used.

Google docs for drawing process flow and deployment process (flowcharts).