HIGH LEVEL DESIGN (HLD)

ADULT CENSUS INCOME PREDICTION

Revision Number: 1.0

Last date of revision: 23/09/2023

Document Version Control

Date Issued	Version	Description	Author
24/06/2023	1	Initial HLD - VI .0	C Sindhuja
23/09/2023	2	Updated HLD - VI .0	C Sindhuja

Table of Contents

Abstract

Predicting income levels is an important task in a variety of sectors, from financial planning to social policy creation. Understanding a person's income can provide important insights on their economic well-being, consumption habits, and general financial stability. Binary classification is a fundamental problem in machine learning in which each instance is assigned to one of two classes based on its attributes. In this project, we tackle the binary classification problem of predicting whether a person earns more than $50,000 per year or not. The two classes are >50K and =50K, which reflect persons with high and low earnings, respectively. To complete this task, we will investigate and study a wide range of factors that may influence an individual's income, including education level, occupation, work hours, marital status, age, and others. These characteristics can assist us in identifying relevant patterns and connections, providing useful insights into the causes that contribute to greater income levels.

1 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- 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 to mildly-technical terms which should be understandable to the administrators of the system.

1.3 Definitions

Term	Description
Database	Collection of all the information monitored by this system
IDE	Integrated Development Environment
UI	User Interface

2 General Description

2.1 Product Perspective

Adult Census Income Prediction is a binary classification problem that predicts whether a person earns more than $50,000 per year or not.

2.2 Problem statement

- To predict whether a person has an income of more than 50K a year or not.
- It is basically a binary classification problem where a person is classified into the >50K group or <=50K group.

2.3 PROPOSED SOLUTION

The proposed solution is a web application written in Python that uses the Flask microservices framework. The programme is hosted on AWS web services and has a user interface (UI) that is interactive and engaging. This application's underlying model is constructed using several linear classification models from the scikit-learn machine learning toolkit. The NumPy and pandas libraries are used to manage data processing efficiently.

2.4 FURTHER IMPROVEMENTS

The approach aims to make considerable advances by significantly increasing the dataset size and applying a cutting-edge deep learning algorithm. This method claims unrivalled accuracy and insight into estimating income levels. The web application will revolutionise the field of income prediction by combining a large dataset with cutting-edge technology, leaving traditional methods behind and setting new benchmarks for excellence in machine learning.

2.5 Technical Requirements

This document addresses the requirements at the early stages:

- Development of a web application in Python using the Flask microservices framework.
- Hosting the application on AWS web services for scalability and reliability.
- Designing an interactive and engaging user interface (UI) to enhance user experience.
- Implementation of a robust underlying model based on multiple linear classification models from the scikit-learn machine learning toolkit.
- Integration of the NumPy and pandas libraries for efficient data processing.
- Ensuring compatibility with various web browsers and mobile devices for wider accessibility.

- Implementing secure authentication and authorization mechanisms to protect sensitive user data.
- Incorporating error handling and logging features for better debugging and maintenance.
- Conducting thorough testing and performance optimization to ensure the application's responsiveness and stability.
- Providing documentation for installation, usage, and maintenance to aid future development and updates.

2.6 Data Requirements

Data requirement completely depend on our problem statement.

- The data needs to be maintained in tabular format
- Since there are only two labels there is a lot of instances for each class
- The Dataset has the following features
    - Age
    - Work class
    - Final weight
    - Education
    - Education-Num
    - Marital-status
    - Occupation
    - Relationship
    - Race
    - Sex
    - Capital-gain
    - Capital-loss
    - Hours-per-week
    - Country
    - Salary
- Here the following are numerical columns
    - age
    - Final weight
    - Capital-gain
    - Capital-loss
    - Hours-per-week
- The rest of the columns are categorical columns

2.7 Tools Used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, render and Flask are used to build the whole model.

- IDE: Visual Studio Code (VSCode) is utilized for Python programming.
- Python: The core programming language used for the entire model development.
- Scikit-learn: Employed to implement various linear classification models.
- AWS: Utilized for hosting and deploying the web application.
- Pandas and NumPy: Integrated for efficient data processing tasks.
- Flask: Used as the microservices framework for building the web application.

2.8 Constraints

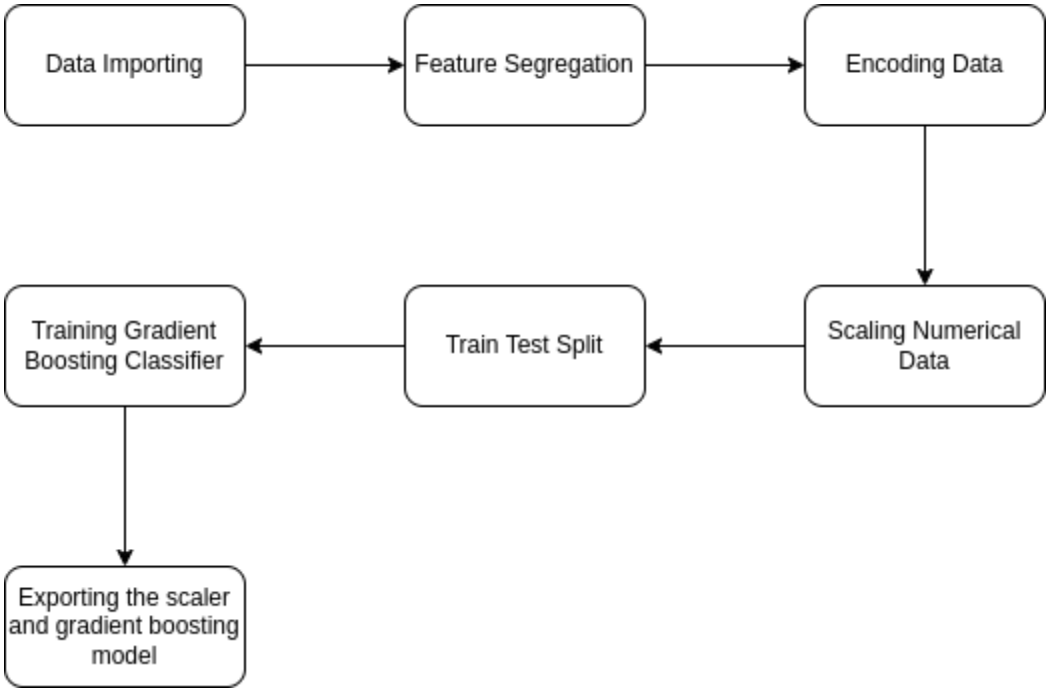
The system is developed with a heavy emphasis on usability in mind, guaranteeing that users may easily acquire the desired outcomes. The user interface and interaction have been optimised to give a smooth experience, allowing users to traverse the programme effortlessly and achieve the required outcomes without any complexity.

2.9 Assumptions

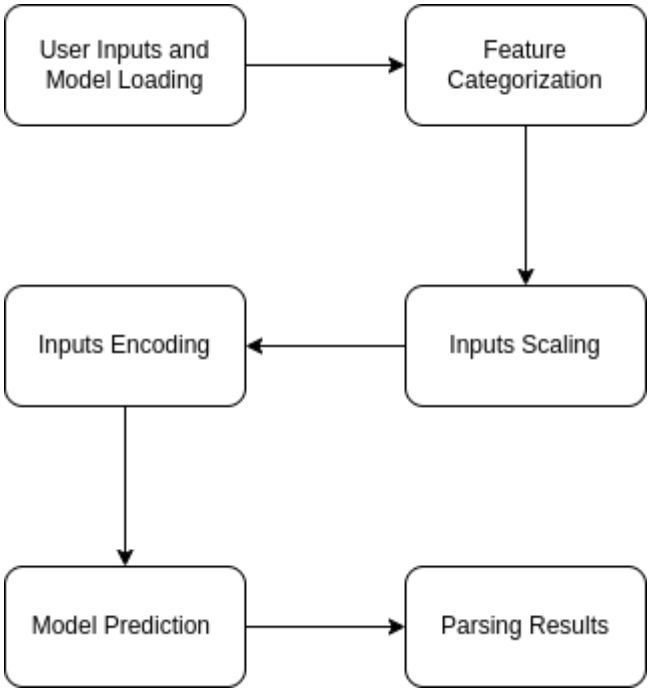
The main objective of the project is to implement the use cases as previously mentioned (2.2 Problem Statement) for making predictions, the system employs a Machine Learning-based model that has been trained on an existing dataset and can also handle fresh data supplied by the user. This allows the system to create correct predictions for both the known dataset and any new data points, assuring the system's adaptability and effectiveness in delivering consistent outcomes. The project is designed to run smoothly since all of its components have been meticulously prepared to fit.

3 Design Details

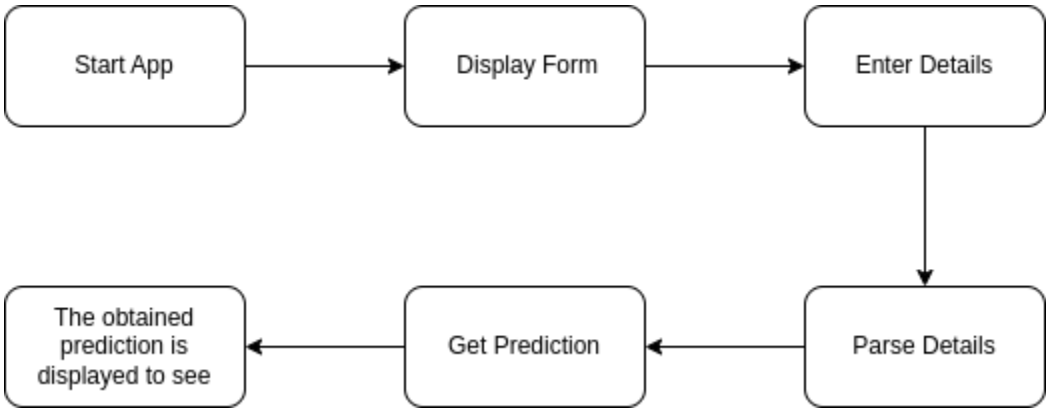
3.1 Training Flow



3.2 Prediction Flow



3.3 Application Flow



4 Performance

The Adult Census Income Prediction solution is used to classify individuals based on numerous variables such as education, age, skills, and so on. persons with no formal education and self-taught persons are also viable data points, therefore the model must be retrained on a regular basis. The importance of regular model retraining is emphasised in order to continuously improve the system's performance and avoid misleading results. The system can effectively discover patterns and generate educated predictions with an accurate underlying model based on linear classification models from the scikit-learn machine learning toolbox.

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

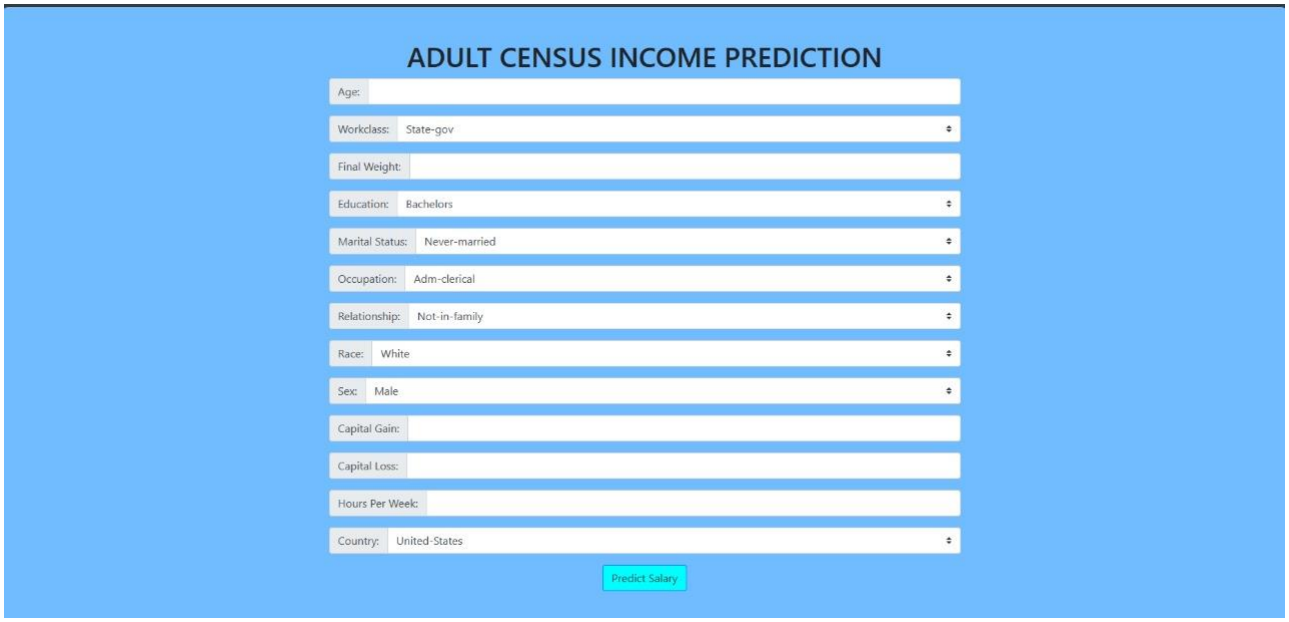
4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

4.4 Deployment

render

4.5 Deployment



5 Conclusion

In conclusion, the Adult Census Income Prediction project uses machine learning to predict income levels. It features a user-friendly web application hosted on AWS. The project focuses on accuracy, security, and future enhancements, aiming to provide valuable insights into income levels for financial planning and policy-making.