Adult Census Income Prediction

High Level Design (HLD)

Domain	: Machine Learning
Creator	: Meghana Munjeti
Mail-ID	: meghameghana297@gmail.com
Date issued	: 07/10/2023

Contents

Abstract

This High level design document outlines the Key aspects of Adult Census Dataset project, including its Scope, problem statement, proposed solution, data requirements, tools used and other relevant details. The document serves as a road map for the design and implementation of the project.

1.Introduction

This dataset contains information from the 1994 census bureau database, and it’s commonly used for Classification and prediction tasks, particularly for income prediction.

1.1 What is High-Level Design Document?

The goal of this HLD or a high-level design 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 design aspects and define them in detail
- Describe all user interfaces being implemented
- Describe the hardware and software interfaces

1.2 Scope

The HLD documentation presents the structure of the system, such as 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

Adult-Census Dataset : A dataset containing demographic and employment related information of individuals.

Income Level : The Annual income of the individual categorized as ‘<=50k’ and ‘>=50k’

2. General Description

The system will be implemented as a web application. The application will be developed using a variety of open source technologies, including Python, Flask. The application will be hosted on a cloud platform, such as AWS.

2.1 Product Description

The project involves developing a machine learning model to predict an individual’s income level based on the dataset feature.

2.2 Problem Statement

To create an ML based solution for predictive analysis of a person’s annual income and also deploy it in the form of a UI.

The aim to predict whether a person earns ‘<=50k’ or ‘>=50k’. This is basically classification problem.

2.3 Proposed Solution

Using all the standard techniques used in the life cycle of a Data Science project starting from Data Exploration, Data Cleaning, Feature Engineering, Model Selection, Model Building and Model Testing and also building a frontend where a user can fill their information in the form input and get the output instantly.

2.4 Further Improvements

Further enhancements may focus on model accuracy, incorporation additional data and improving the interface.

2.5 Data requirements

The system will require the Adult Census dataset to be processed and analyzed. The dataset is available in a CSV format.

2.6 Tools Used

- For visualization tasks, matplotlib, seaborn and plotly were used
- Flask were used for building the web application and server to run the code
- Apache Cassandra was used to storage and retrieval of data
- GitHub is used as version control system
- NumPy and Pandas were used to clean and interpret data
- Scikit-learn was used to cross validate and compare different models

Hardware requirements

No specific hardware requirements are necessary, as project can be executed on standard personal computer.

2.7 Constraints

The front-end must be user friendly and should not need any one to have any prior knowledge in order to use it.

2.8 Assumptions

The following assumptions are made about the project:

- The Adult Census dataset will be available in a usable format.
- The system will be hosted on a cloud platform.
- Users will have access to a web browser.

3.Design Details

Process Flow:

The process flow involves following steps:

1. Data Collection
2. Data Preprocessing
3. Feature Engineering
4. Model Building
5. Model Evaluation
6. Deployment

3.1 Event Log

The system should log every event so that the user will know what process is running internally. Initial step-by-step description:

1. The system identifies at what level logging is required

2. The system should be able to log each and every system flow

3. Developer can choose logging method. You can choose database logging/ File logging as well

4. System should not hang even after so many loggings. Logging just because we can easily debug issues, so logging is mandatory to do.

3.2 Error Handling

Errors should be encountered, an explanation will be displayed as to what went wrong .

An error will be defined as anything that falls outside the normal intended usage.

3.3 Performance

Performance Consideration include:

1.Model Accuracy

2.Model training time

3.Resource utilization

3.4 Reusability

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

3.5 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 Python to ensure proper transfer of information.

3.6 Resource Utilization

Efficient utilization of computing resources will be a priority to ensure scalability

3.7 Deployment

We will Deploy our Model in AWS Cloud Platform

4. Dashboards

Interactive dashboards may be developed to visualize project results and insights.

4.1 KPIs(Key Performance Indicators)

- Key Performance Indicators of Adult Census
- Latency or the amount of time the application takes to display results for some specific input.
- The processing power our application takes to run
- The memory and RAM our application takes to run on a web server.

5. Conclusion

All in all, overall project architecture, design details, used technologies and performance were explained in detail.

The system will be implemented as a web application using a variety of technologies. The application is hosted on a cloud platform, which will make it accessible to users from anywhere in the world.