HIGH LEVEL DESIGN (HLD)

ADULT CENSUS INCOME PREDICTION

Revision Number: 1.0

Last date of revision: 28/12/2022

Document Version Control

Date Issued	Version	Description	Author
28/12/2022	1	Initial HLD – V1.0	Panuganti Arun Kumar

Contents

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 using the following conditions: ((Age>16) and (AGI > 100) and (AFNLWGT>1) and (HRSWK>0)).

The Goal is to classify whether a person has an income of more than 50K a year or not. This is basically a binary classification problem where a person is classified into the >50K group or <=50K group.

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.

2 General Description

2.1 Product Perspective

The Adult Census Income Classification is a supervised machine learning based Price classifyion model which help us to classify the income by taking the input from the user in the web application in the form of input.

2.2 Problem statement

The Goal is to classify whether a person has an income of more than 50K a year or not based on different factors available in the provided dataset.

2.3 Proposed solution

The solution is here is the Adult Census Income Classification solution can be implemented to solve the problem statement i.e. to classify the income of the census based on different factors available in the provided dataset. At first case the model should be validate the data and then process the data in the suitable format, In second case the transformed data is to be trained using machine learning algorithms and produce the model. In third case the evaluation of the model should be done and finally the model will be ready for the estimation of new data input.

2.4 Further improvements

Adult Census Classification model can be improved by adding new features like to generate Exploratory Data Analysis on the data which is given for training to the model this will be helpful for taking the insights of the real data. By connecting this model to the database which can give real time updated data we can get more accurate results by continue training.

2.5 Technical requirements

This document addresses the requirements for the model building for classifying the flight fares. And recommending the necessary tools for the process of model building.

- Local machine (computer) with good processor because Machine learning needs good computational power.
- If you are using the cloud computing resources choose the best configuration.

- VS Code editor / PyCharm software for programming.
- Python version 3.9
- Good Internet connection for pushing the code to Git hub.
- Docker desktop application to create container.

2.6 Data Requirements

Data requirements completely depends on our problem statement.

- We need a data that is balanced and must have at least 10000 records.
- We require at least 8000 records for training the model and 2000 records for testing/Validating the model.
- The data should be in CSV file format.
- The data must be in text format with Unicode UTF-8

2.7 Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Flask, gunicorn, PyYAML, evidently and dill



- VS code editor is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- Front end HTML and CSS.
- Flask is used for the backend development.

- GitHub used for repository.
- Git is used for version control and deployment.

2.8 Constraints

The Adult Census Income Classification web application must be user friendly for that purpose a video on how to use must be provided. The major constraint is that if the dataset is changed the model will throw an error or stop working.

2.9 Assumptions

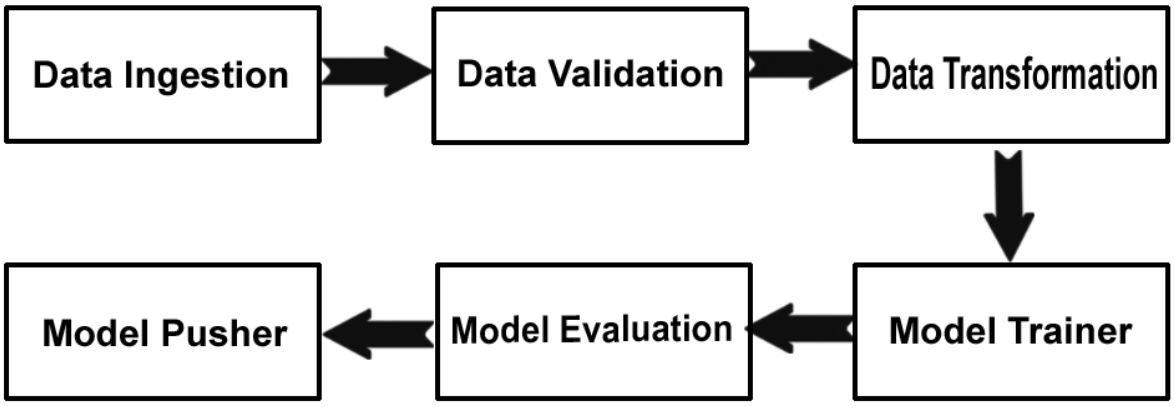
The main objective of the project is to implement the use cases as previously mentioned (2.2 Problem Statement) for new dataset that comes through the pipeline of model must get trained by the data and give classification. It is also assumed that all aspects of this project have the ability to work together in the way the developer is expecting.

3 DESIGN DETAILS

3.1 Process Flow

For classifying the income of the census, we will use 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.

Proposed Methodology



3.2 Event log

The system should log every event so that the suer will know what process is running internally.

Initial Step-By-Step Description:

1. The System identifies at what step logging 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 using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

3.3 Error Handling:

Should errors be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

3.4 Performance

The adult census income classification web application is used for classifying income of people. The performance is good at present and can also need to be improve more in future and also new features can be implemented to become more user friendly and useful.

3.5 Reusability

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

3.6 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.

3.7 Resource Utilization

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

3.8 Deployment

Deployment is done on the **Heroku** platform.

4 CONCLUSIONS

The Adult Census Income Classification web application can help the real-world users by giving the estimation of Prices. By this user can plan their travelling when they have enough budget.