Contents

Document Version Control

Date Issued	Version	Description	Author
30/10/2021	1	Initial HLD – V1.0.1	Sonu Kumar Pal

Abstract

The prominent inequality of wealth and income is a huge concern especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improves the economic stability of a nation. Governments in different countries have been trying their best to address this problem and provide an optimal solution. This study aims to show the usage of machine learning and data mining techniques in providing a solution to the income equality problem. The UCI Adult Dataset has been used for the purpose. Classification has been done to predict whether a person's yearly income in US falls in the income category of either greater than 50K Dollars or less equal to 50K Dollars category based on a certain set of attributes.

High Level Design (HLD)

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.

2. General Description

2.1 Product Perspective

This application is a machine learning model which will help us to predict whether the person income is more than 50k or not.

2.2 Problem statement

The Goal is to predict 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.

2.3 Proposed Solution

The solution here is an Classification based Machine Learning model. It can be implemented by different classification algorithms (like Logistic Regression, Random Forest Classifier, Decision Tree Classifier, SVC and so on.).Here first we are performing Data preprocessing step, in which feature engineering, feature selection, feature scaling are performed and then we are going to build model.

2.4 Technical Requirements

In this project the requirements to get person income through various platforms. For that, in this project we are going to use different technologies. Here is some requirements for this project.

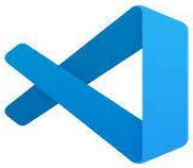
- Model should be exposed through API or user interface, so that anyone can test model.
- Cassandra database should be integrated in this project for any kind of user input.
- Model should be deployed on cloud (Azure, AWS, GCP).

2.5 Data Requirements

Data Requirement completely depends on our problem.

- For training and testing the model, we are using adult census income dataset from uci machine learning repository.
- From user we are taking following input :
    - Age
    - Workclass
    - Education
    - Maritial-Status
    - Occupation
    - Relationship
    - Race
    - Sex
    - Country
    - Capital Gain
    - Capital Loss
    - Hours-per-week

2.6 Tools Used



- Vscode is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- Azure is used for deployment of the model.
- Cassandra is used to retrieve, insert, delete, and update the database.
- Front end development is done using HTML/CSS, Flask is used for backend development and for API development.
- GitHub is used as version control system.

2.7 Constraints

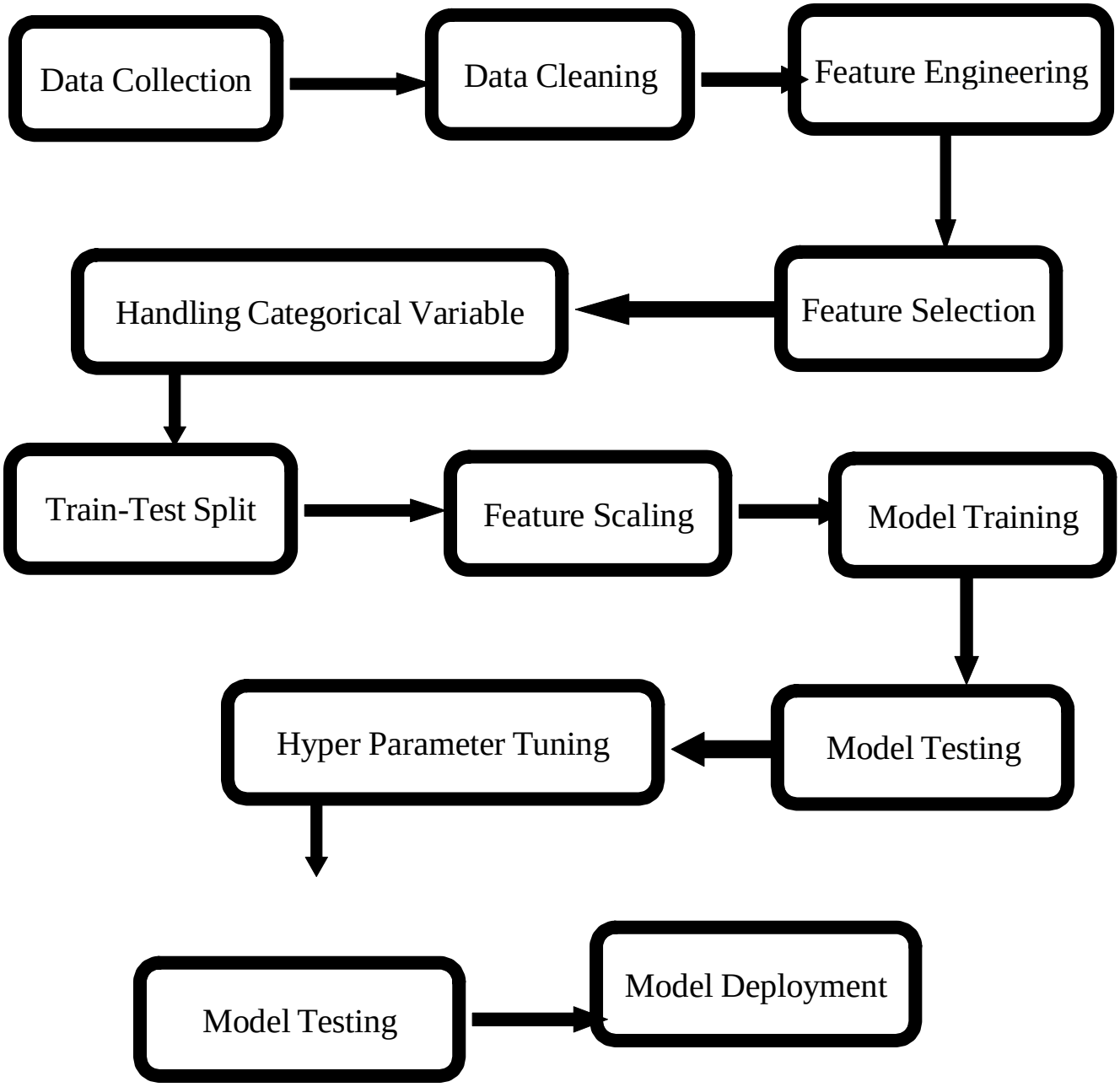
The income prediction system must be user friendly, errors free and users should not be required to know any of the back-end working.

2.8 Assumptions

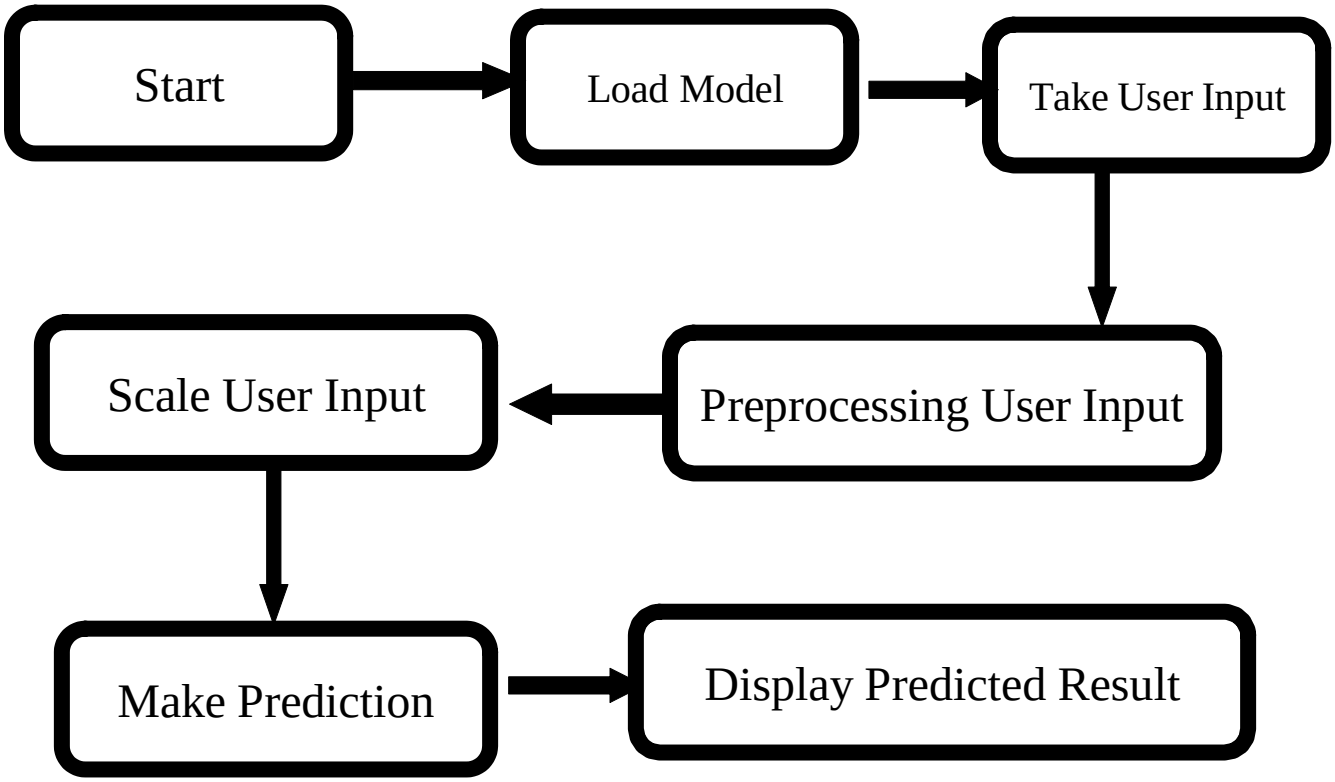
The main aim of this application is to predict whether the person income is more than 50k or not. So it is also assumed that whenever new data will be loaded the model will give accurate result.

3. Design Details

3.1 Process Flow



3.2 Deployment Process



3.3 Event Log

In this Project we are logging every process so that the user will know whatprocess is running internally.

Step-By-Step Description:

- In this Project we defined logging for every function, class.
- By logging we can monitor every insertion, every flow of data in database.
- By logging we are monitor every step which may create problem or everystep which is important in file system.
- We have designed logging in such a way that system should not hang even after so many logging's, so that we can easily debug issues which may arises during process flow.

3.4 Error Handling

We have designed this project in such a way that, at any step if error occur then our application should not terminate rather it should catch that error and displaythat error with proper explanation as to what went wrong during process flow.

4. Performance

Our application is built with best model with good accuracy and it will predict accurate income of user.

4.1 Re-usability

We have done programming of this project in such a way that it should be reusable. So that anyone can add and contribute without facing any problems.

4.2 Application Compatibility

This application is work same for all system. It's a platform independent application.

4.3 Resource Utilization

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

4.3 Deployment

Google cloud, AWS, Azure are used for deployment

5. Conclusion

Adult census income prediction web application predicts the person income is more than 50k or not. Accuracy of this application is good and it will predict 85-90% accurate results.

6. Reference

- https://archive.ics.uci.edu/ml/datasets/adult
- https://www.researchgate.net/publication/344196724_Machine-learning_analysis_for_adult_census_income_dataset
- https://ieeexplore.ieee.org/document/8748528
- https://acadpubl.eu/hub/2018-118-22/articles/22a/88.pdf