HIGH-LEVEL DESIGN (HLD)

Adult Census Income Prediction

Last date of revision: 03 September 2021

Document Version Control

Date issued	Version	Description	Author
28.08.2021	1	General Description	Sathappan PR
30.08.2021	2	Design Details	Susan Elizabeth Varghese
01.09.2021	3	Dashboard	Jebin R

CONTENTS

Abstract

In recent trends many employees are acquiring a salary in less than 50k due to excess of population and also, companies expense is going high.
Now a day's technology is improved compare to employee machine is working two times better and proceeding high productivity in less time.

1 Introduction

1.1 Why this High Level Design Documents?

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 model 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 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

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 use non-technical to mildly-technical terms which should be understandable to the administrators of the system

1.3 Definition

Database	Collection of all the information monitored by the system
IDE	Integrated Development Environment
ML	Machine Learning
ACIP	Adult Census Income Prediction

2 General Description

2.1 Product Perspective

The Adult Census income Prediction is done in Machine Learning (ML). The Decision Tree model, which help us to predict whether adult salary is less than 50k or more than 50k.

2.2 Problem Statement

To analyze the income prediction with the help of Machine Learning solution in Decision Tree model to implement the following use cases.

- 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 proposed here is an ACIP (Adult Census Income Prediction) based on ML algorithm can be implemented to perform above mention use cases in first case, if goal is to predict whether a person has an income of more than 50K a year or not. Further in the second case, this is basically a binary classification problem where a person is classified into the >50K group or <=50K group. These analysis will be done in Decision Tree model.

2.4 Further Improvements

ACIP can be added with more use cases like salary prediction, based on experience to analyze the adult income. For better and fast response or action, with help of ACIP and AAI to implement in all the sectors.

2.5 Data Requirements

Data requirement completely depend on a problem statement

- We need dataset based on the problem to get accurate solution and n-number of rows and columns
- We required at least one independent variable and one dependent variable to analyze the problem.
- In this data one dimensional and two dimensional array are intimated.
- 13 Column, 32561 rows are presented in this dataset.
- In this problem the data is analyzed to train and testing for building the model in machine learning algorithm and checking the accuracy for deploying.

2.6 Tools Used

Python programming language and frame works such as NumPy, Pandas, PyCharm, Scikit-learn, Heroku, Flask are used to build the whole model

- PyCharm is used as IDE,
- For visualization of the plots, matplotlib and seaborn are used,
- Heroku is used for deployment of the model,
- Tableau is used for creating dashboard ,
- HTML & CSS is used to front end development,
- Python and Flask is used to back end development ,

- Github is used for version control system.

2.7 Constraints

The ACIP based on salary for employees in the basis of independent variable and it comprised on machine learning model in the workings.

2.8 Assumptions

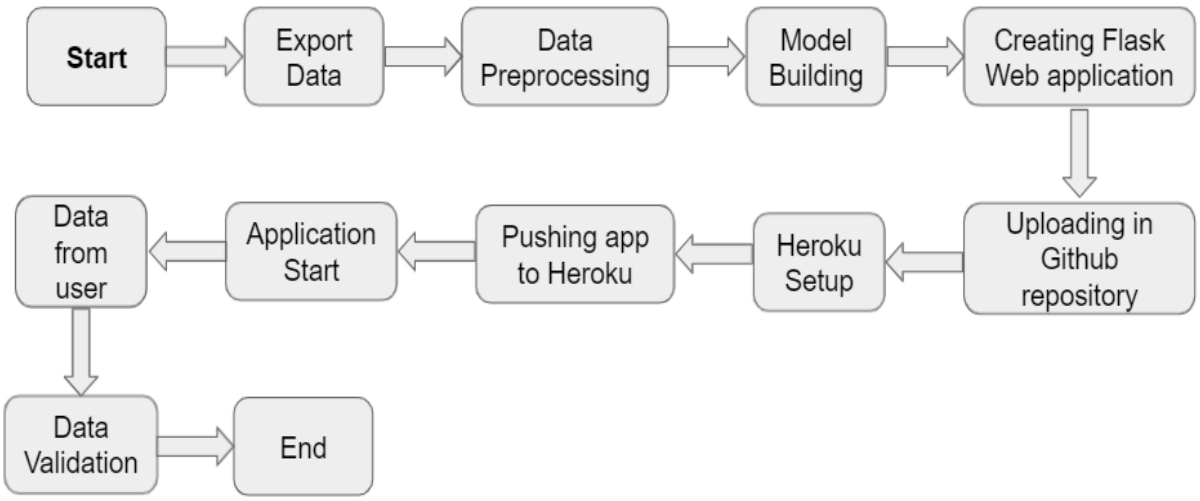
The main objective of the project is to implement the use cases as previously mentioned (2.2 Problem Statement) foe new dataset that comes through ACIP is to train and test the model for checking the accuracy the model is fit, overfit or underfit. If the model is fit proceeding with algorithm to get the prediction for deploying purpose. It is assumed that all aspects of this project has ability to work together in deploying and expecting.

3. Design Detail

3.1 Process Flow

For identifying the different type of anomalies, we will use a machine learning base model. Below is the process flow diagram is as shown below.

1. Proposed Methodology
2. Model Training Evaluation,
3. Deployment Process



3.2 Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong? If in case more missing values are comprised the data will be dropped or adding the dummy variable. An error will be defined as anything that falls outside the normal and intended usage

4. Performance

The ACIP solution is used for Adult Salary prediction less than 50k or more than 50k in the bases of independent variable. Training and testing in SkLearn are pre processing for building the model and checking the accuracy.

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 performed, it will likely use all the processing poer available until that function is finished

4.4 Deployment

5. Dashboard

Dashboards will be implemented to display and indicate certain KPIs and relevant indicators for the unveiled problems that if not addressed in time could cause catastrophes of unimaginable impact

As when the system start to capture the historical/periodic data for a user, the dashboards will be includes to display charts over times with progress on various indicators or factors

5.1 Key performance indicators (KPI)

1. Comparison of accuracy in model prediction.
2. Occupation and education is more important for model prediction.
3. Capital Gain and Capital Loss is conforming whether the salary is less are more than Rs.50k for adult.
4. The collected performance is good, but not highly optimized (e.g. Hyper parameters are tuned).
5. Implementing the pipe line, where the first step is a column transformer that applies a one hot encoder to just the categorical variables and numerical variable which helps to predict.

6. Conclusion

In this salary prediction to analyze (ACIP) will be less than 50k or greater than 50k is trained in algorithm and build a model. And also checking accuracy, the model to be fit for finding out the perfect algorithm to deploy the model and getting perfect prediction based on ACIP. The capital gain and capital loss are more important component in predicting the salary.