HIGH LEVEL DESIGN (HLD)

Adult Census Income Prediction

Abhishek Upadhyay
iNeuron Intelligence Pvt Ltd

Document Version Control

Date Issued	Version	Description	Author
08.05.2024	V1.0	Initial HLD-V1.0	Abhishek Upadhyay

Contents

Abstract

We analyze the personal health data to predict insurance premium of individuals. Seven regression models naming Logistics Regression, Decision Tree Classification, Random Forest Classification, Gradient Boosting Classification, KNN, SVC have been used to compare and contrast the performance of these algorithms.

Training dataset was used for training model and that training model helped to come up with some predictions. Then the predicted amount was compared with actual data to test and verify the model accuracy. Later accuracies of all these models were compared. It was gathered that Gradient Boosting and Random Forest algorithms performed better than the remaining models.

Gradient boosting is best suited in this case because it gives best evaluation score comparable to other models.

1.0 Introduction

1.1 Why this High-Level Design Document?

The purpose of this High-Level document is to add necessary details to current project description to represent a suitable model for coding. This document is used as a reference manual for how the model interact at a high-level.

The HLD will

- Presents all 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 feature and the architecture of the project.

1.2 Scope

The HLD document presents the structure of the system, such as the database architecture, application architecture, and technology architecture. The HLD uses non-technical to middle-technical terms which should be understandable to the administrators of the system.

1.3 Definitions

Term	Description
Database	Collection of all the information
IDE	Integrated Development Environment
API	Application Programming Interface
KPI	Key Performance Indicator
VS Code	Visual Studio Code
EDA	Exploratory Data Analysis
KNN	XGB Classification Type your text Type your text Type your text Type your text Type your text Type your text Type your text Type your text

2.0 General Description

2.1 Product Perspective

The Goal is to predict whether a person has an income of more than 50K a year or n
This is basically a binary classification problem where a person is classified into the
>50K group or <=50K group.

Type your text Type your text

2.2 Problem Statement

To develop an API interface to predict the Adult Income Census income using people individual data and analyzing the following:

To detect Age value affects the prediction.

To detect education affects the Income of the Person.

- To create API interface to predict the premium

2.3 Proposed Solution

The solution proposed here is an estimating income of Adult person based on people data and this can be implemented to perform above mention use cases. In first case, analyzing how Age value affect the people income as well as income of the person. In the second case, if model detects the smoking affecting the premium, we will inform that to people. And in the last use case, we will be making an interface to predict the income.

2.4 Further Improvements

2.5 Technical Requirements

The solution can be a cloud-based or application hosted on an internal server or even be hosted on a local machine. For accessing this application below are the minimum requirements:

- Good internet connection.
- Web Browser.

For training model, the system requirements are as follows:

- +4 GB RAM preferred
- Operation System: Windows, Linux, Mac
- Visual Studio Code / Jupyter notebook

2.6 Data Requirements

Data requirements completely depends on out problem statement.

- Comma separated values (CSV) file.
- Input file feature/field names and its sequence should be followed as per decided.

2.7 Tools Used

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

- Pandas is an open-source Python package that is widely used for data analysis and machine learning tasks.
- NumPy is most commonly used package for scientific computing in Python.
- Plotly is an open-source data visualization library used to create interactive and quality charts/graphs.
- Scikit-learn is used for a machine learning.
- Flask is used to build API.
- VS Code is used as IDE (Integrated Development Environment)
- GitHub is used as version control system.
- Front end development is done using HTML/CSS.
- Heroku is used for deployment of the model.

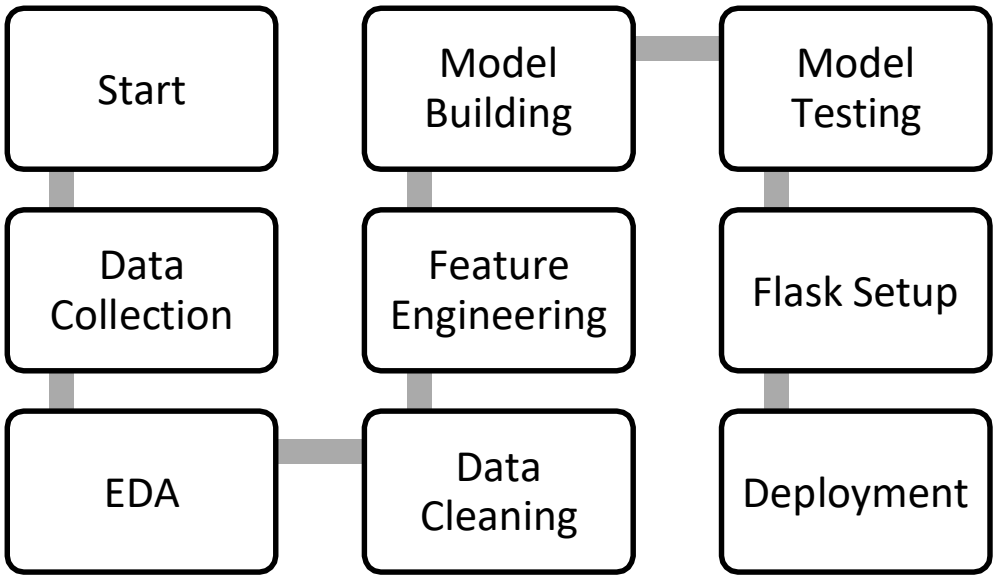
2.8 Constraints

2.9 Assumptions

The main objective of the project is to develop an API to predict the premium for people on the basis of their health information. Machine learning based regression model is used for predicting above mentioned cases on the input data.

3.0 Design Details

3.1 Process Flow



3.2 Event Log

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

Initial Step-By-Step Description:

- The system identifies at what step logging required.
- The system should be able to log each and every system flow.
- Developer can choose logging method. You can choose database logging.

System should not hang out even after using so many loggings.

4.0 Performance

4.1 Reusability

The entire solution will be done in modular fashion and will be API oriented. So, in the case of the scaling the application, the components are completely reusable.

4.2 Application Compatibility

The interaction with the application is done through the designed user interface, which the end user can access through any web browser.

4.3 Deployment

5.0 Dashboards

A dashboard is a data visualization and analysis tool that displays on one screen the status of key performance indicators (KPIs) and other important business metrics.

As a high-level reporting mechanism, dashboards provide fast 'big-picture' answer to critical business questions and assist and benefit decision making in several ways:

- Communicating how premium is varies with BMI value.
- Visualizing relationship of gender with premium in easy-to-understand way.

6.0 Conclusion

This system shows us that the different techniques that are used in order to estimate the how much amount of premium required on the basis of individual health situation. After analyzing it shows how a smoker and non-smokers affecting the amount of estimate. Also, significant difference between male and female expenses. Accuracy, which plays a key role in prediction-based system. From the results we could see that Gradient Boosting turned out to be best working model for this problem in terms of the accuracy. Our predictions help user to know how much amount premium they need on the basis of their current health situation.