INSURANCE PREMIUM PREDICTION

High Level Design

FEBRUARY 19, 2023

DOCUMENT VERSION CONTROL

Date Issued	Version	Description	Author
19.02.2023	V1.0	Initial HLD-V1.0	Shikha

Contents

ABSTRACT

Our study aimed to predict insurance premiums for individuals by analyzing their personal health data. To accomplish this, we employed different regression models, including Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, Extra Tree Regressor, XG Boost and KNN, to compare their performances. We used a training dataset to train the models and generate predictions, which we then compared with actual data to test their accuracy. Finally, we compared the accuracies of all the models and found that Gradient Boosting and Random Forest algorithms outperformed the others. Among these, Gradient Boosting proved to be the most suitable for this case, as it yielded the highest evaluation score compared to the 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

2.0 General Description

2.1 Product Perspective

A machine learning predictive model is utilized to estimate insurance premiums, which aids in predicting the cost of health insurance premiums for individuals.

2.2 Problem Statement

The objective is to create an API interface for premium insurance prediction by analyzing an individual's health data and considering the effects of their BMI value and smoking habits on the insurance premium. This interface is designed to accurately predict the premium cost and provide valuable insights for insurers and policyholders.

2.3 Proposed Solution

The proposed solution involves utilizing individuals' health data to estimate insurance premiums, which can be implemented to address the aforementioned use cases. Firstly, the solution involves analyzing the impact of BMI values on an individual's health and the resulting insurance premium. Secondly, if the model detects that smoking habits affect the premium, we will provide information to the individuals. Finally, an interface will be developed to predict the premium, serving as the third use case.

2.5 Technical Requirements

To utilize the solution, you may access either a cloud-based or application-hosted service. Minimal requirements are needed to access the software, including a stable internet connection and a web browser. However, for training the model, specific system requirements must be met. These include a preferred 4 GB of RAM, compatibility with common operating systems such as Windows, Linux, or Mac, and software such as Visual Studio Code or Jupyter notebook.

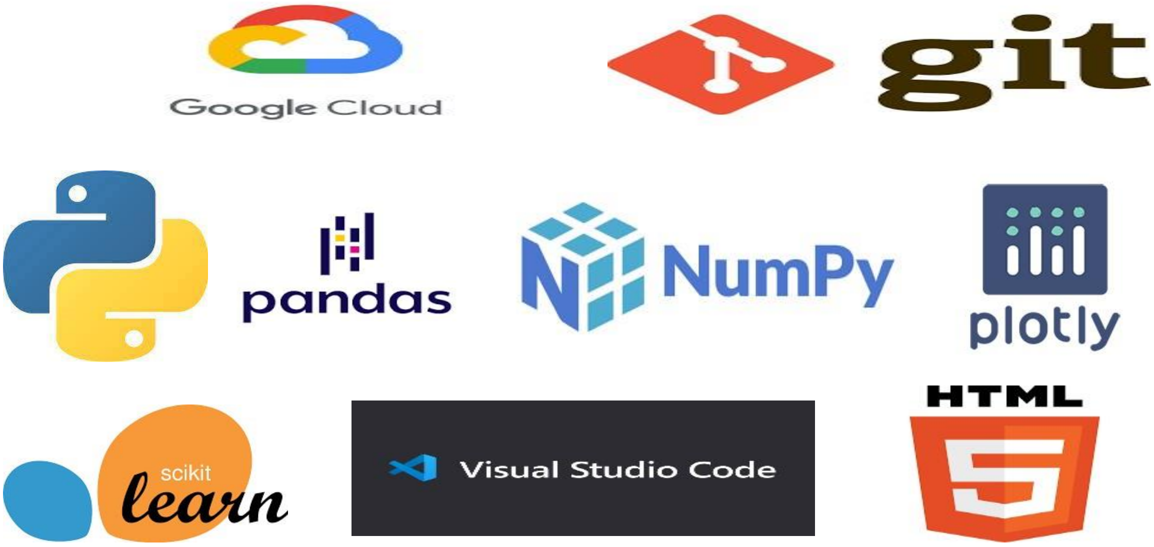
2.6 Data Requirements

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, Gradio 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.
- Gradio 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.
- Gradio is used for deployment of the model.

2.8 CONSTRAINTS

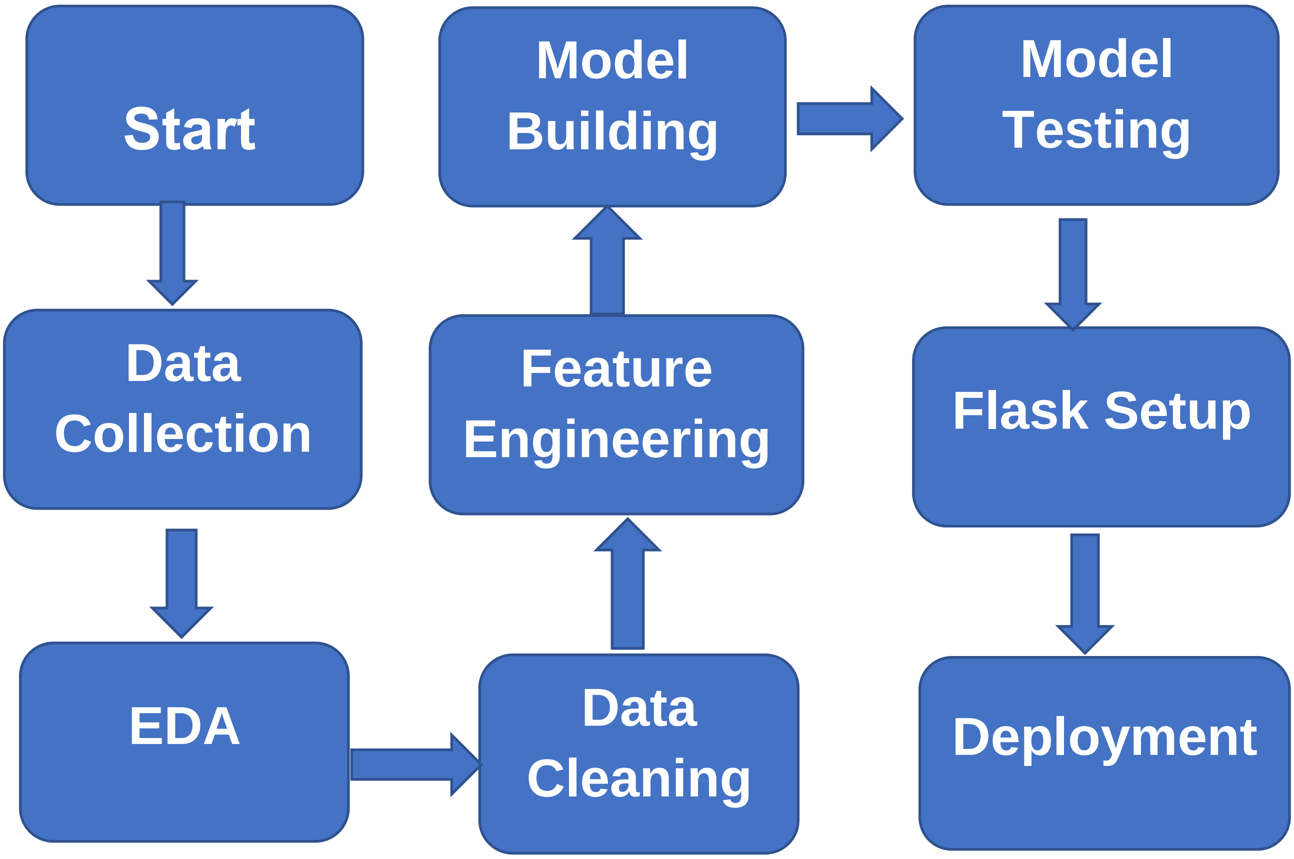
The application has the capability to furnish personalized insurance quotes based on various user-specific data points, such as their Body Mass Index (BMI), gender, and smoking status.

2.9 ASSUMPTIONS

The primary goal of this project is to create an API that can predict individuals' insurance premiums based on their health information. To achieve this, a machine learning-based regression model is utilized to forecast the insurance premium based on various inputs.

3.0 Design Details

3.1 Process Flow



3.2 Event Log

The system should keep track of every event so that the user will be aware of what processes are running inside it.

- Determine the appropriate step for the logging process.
- Ensure that the system can capture and record all system activities.
- Developers can choose from different methods for database logging.
- Limit excessive logging and avoid continued logging beyond a certain point.

4.0. Performance

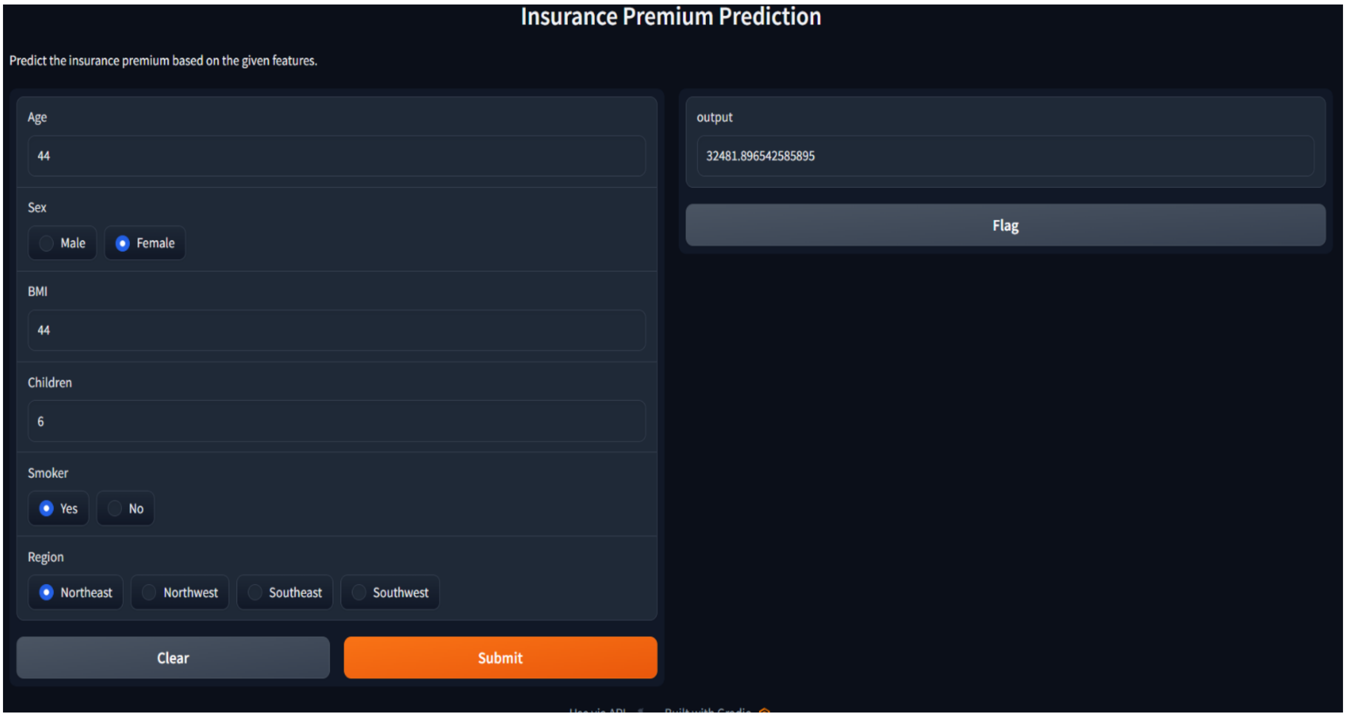
4.1 Reusability

The solution will be developed using a modular design, which will facilitate future scaling of the application and all solution components will be available through an API, providing a flexible way to modify the application as needed.

4.2 Application Compatibility

End users can access the application via any web browser by utilizing the provided user interface.

4.3 Deployment



Insurance Premium Prediction

Predict the insurance premium based on the given features.

Age

44

Sex

○ Male ● Female

BMI

44

Children

6

Smoker

● Yes ○ No

Region

● Northeast ○ Northwest ○ Southeast ○ Southwest

output

32481.896542585895

5.0 CONCLUSION

The system provides various techniques for estimating the amount of premium required based on an individual's health situation. The analysis shows how smoking and non-smoking habits can affect the estimated amount, and highlights the significant differences in expenses between males and females. Accuracy is a crucial factor for a prediction-based system, and the results indicate that Gradient Boosting is the best performing model in terms of accuracy. The predictions generated by the system enable users to determine the amount of premium required based on their current health situation.