INSURANCE PREMIUM PREDICTION
HIGH LEVEL DESIGN (HLD)

Document Version Control

Date Issued	Version	Description	Author
26.8.2024	V1.0	Initial LLD- V1.0	Sarthak Bansal
26.8.2024	V1.0	Initial LLD- V1.0	Jayant Bisht
26.8.2024	V1.0	Initial LLD- V1.0	Hardik Mudgal
26.8.2024	V1.0	Initial LLD- V1.0	Kartik
26.8.2024	V1.0	Initial LLD- V1.0	Mayank Singla

Contents

Abstract

This paper explores using machine learning to enhance insurance premium prediction, aiming to improve accuracy beyond traditional actuarial methods. By applying advanced algorithms like gradient boosting, random forests, and neural networks, the study addresses challenges such as data quality and feature engineering. The research involves pre-processing a detailed dataset, training various models, and evaluating their performance using metrics like Mean Absolute Error. The results show that machine learning models provide more precise premium forecasts compared to traditional methods. The paper also discusses the practical integration of these models, including regulatory compliance and ethical considerations. Overall, the study highlights the potential for machine learning to improve decision-making and customer satisfaction in the insurance industry.

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: o 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 **Insurance Premium Prediction** model is based on machine learning which helps us to determine the estimate the cost of insurance for individuals based on multiple factors such as their age, BMI, sex, etc.

2.2 Problem Statement

The goal is to develop a machine learning model that accurately predicts insurance premiums based on various factors such as age, gender, medical history, lifestyle habits, and geographic location. Rising healthcare costs and the complexity of premium calculations necessitate a reliable prediction model to help insurance companies set fair premiums and assist customers in understanding their potential costs. The model should leverage historical data to identify patterns and relationships, ensuring accurate and fair premium estimations.

2.3 Proposed Solution

The solution proposed here is to estimate the premium of insurance based on individual's health data. This can be implemented by using the various factors which are mentioned above which can effect the cost of insurance. The factors which can help us deduce the cost of premium are :-

- BMI is one the main factors as it indicates the overall fitness of individual and hence can be used for predicting the cost of insurance.
- Smoking, it is used so that if the individual is smoker the chances of him/her being healthy reduces which can negatively affect the premium for insurance.
- Region is also used for this as the environment we live in can affect our lives and ultimately it is used for determining the cost of insurance.

These are some of the factors which can be used for insurance premium prediction model.

2.4 Technical Requirements

The solution proposed here can be implemented as a cloud-based solution or as an application hosted on an internal server, or even on a local machine. To access this application, the following minimum requirements are necessary:

- A good internet connection.
- A web browser.

For training the model, the following system requirements are preferred:

- 4GB RAM or more.
- An operating system such as Windows, Linux or Mac.
- Visual Studio Code or Jupyter notebook.

2.5 Data Requirements

The data requirements for this project will depend on the specific problem statement. A CSV file will be used as the input file, and the feature/field names and sequence should be followed as decided. It's important to have a clear understanding of the problem statement and the data that is required to solve it, to design a suitable data pipeline, and to train the model effectively.

2.6 Tools Used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, TensorFlow, Keras, and Roboflow 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 the 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 machine learning.
- Flask is used to build API.
- VS Code is used as an IDE (Integrated Development Environment)
- GitHub is used as a version control system.
- Front-end development is done using HTML and CSS.
- Railway is used for the deployment of the model.

2.7 Constraints

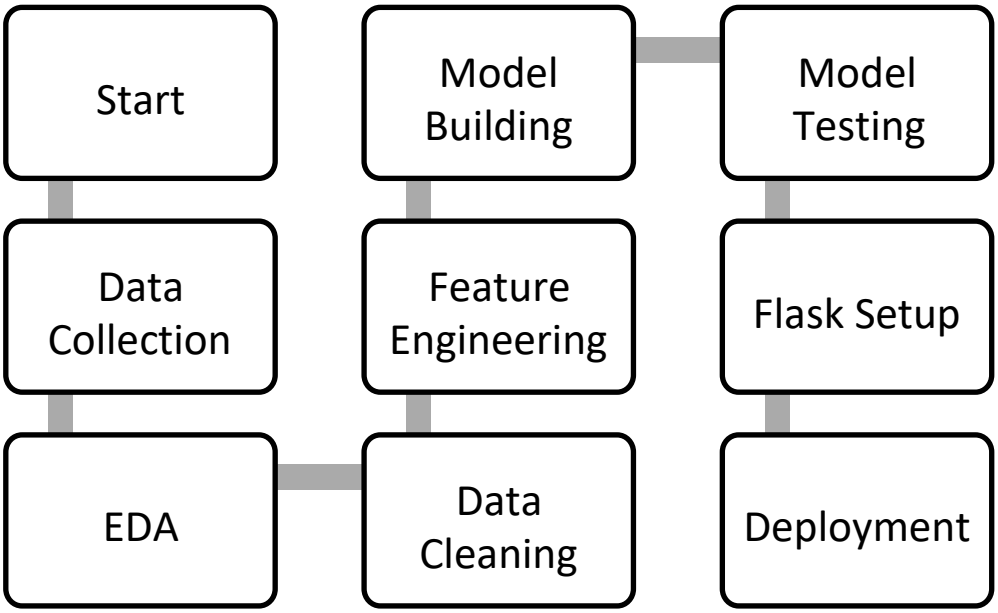
It is useful for the user by predicting Insurance prices based on their provided details for ex: - BMI, sex, smoker, yes/no, age, etc.

2.8 Assumptions

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

3 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.

3.3 Error Handling

Error handling in insurance premium prediction involves managing inaccuracies or issues that arise during the model building, evaluation, and deployment phases. An error is defined as anything that falls outside the intended usage and an explanation will be given as to what went wrong.

4 Performance

4.1 Reusability

The entire solution will be done in modular fashion and will be API oriented. So, the components are completely reusable and therefore, can be used for further scaling of application.

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 Dashboards

Dashboards are powerful tools in machine learning that provide visualizations and interactive elements to help monitor, evaluate, and manage ML models and their performance. They can be used for various purposes, including tracking model training, evaluating predictions, managing data, and facilitating collaboration.

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-tounderstand way.

6 Conclusion

The insurance premium prediction system has been successfully designed to enhance the accuracy and efficiency of premium calculations. Key highlights include:

- **Effective Data Integration:** The system consolidates diverse data sources, ensuring high-quality input for predictions.
- **Accurate Predictive Modeling:** Advanced machine learning algorithms, including Random Forest and Gradient Boosting, have been employed to achieve reliable forecasts.
- **User-Friendly Interface:** An intuitive design facilitates seamless data input and prediction retrieval.
- **Scalability and Security:** The system is built to scale with growing data and users, incorporating robust security measures to protect sensitive information.

Future enhancements could include integrating additional data sources and refining predictive models. Overall, the system significantly improves premium prediction processes and supports strategic decision-making.