Project:

Insurance Premium Predictor

By-Mr.Aniket Dumbre

Prepared By	Approved By
Mr.Aniket Dumbre	Mr.Sunny Savita

Abstract:

We analyze the personal health data using various EAD techniques and experimented with various ML approaches like regression, ensemble techniques (bagging & boosting), clustering etc. We have used ML algorithms naming Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, KNN, AdaBoost Regression, XGB Regression, etc to compare and contrast the performance of these algorithms and select one which has higher accuracy. Random Forest is best suited in this case because it gives best evaluation score comparable to other models. Here we design pipeline for model training and prediction in modular programming way. Git is used for source code version control. Apache Airflow is used for the scheduling and orchestration of data pipelines in docker.

1. Introduction

1.1 Why this High – Level Design Document?

The purpose of this High-Level Design Document is to add the necessary details to the current project description to represent a suitable model for coding. This document is also intended to help to 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 details.
- 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:
  - Security
  - Reliability
  - Maintainability
  - Portability
  - Reusability
  - Application Compatibility
  - Resource utilization
  - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as database 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.

1.3 Definitions

Terms	Descriptions
Database	Collections of all information monitored by the system
IDE	Integrated Development Environment
AWS	Amazon Web Services
KPI	Key Performance Indicator
VS Code	Visual Studio Code
EDA	Exploratory Data Analysis
KNN	KNearest Neighbors

2. General Description

2.1 Product Perspective

The insurance premium predictor is a ML based model which will help us to predict the insurance premium based on individual health situation.

2.2 Problem Statement

To develop an API interface to predict the premium of insurance using people individual health data and analyzing the following:

- To detect BMI value affects the premium.
- To detect smoking affects the premium of the insurance.
- To create API interface to predict the premium

2.3 Dataset Description

Dataset have 3 numerical features (age, bmi, children etc), 3 categorical features (sex, smoker, and region) and target feature 'expenses' as numerical feature.

2.4 Attribute Information

1. age – age of individuals in years
2. sex – Male/Female
3. bmi – body mass index
4. no of children – no. of children person have
5. smoker – whether person smokes or not
6. region – 4 region
7. expenses – expenses of person on health insurance

2.5 Proposed solution

The solution proposed here is an estimating premium of insurance based on people health data and this can be implemented to perform above mention use cases. In first case, analyzing how BMI value affects the people health as well as premium of the insurance. 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 premium

2.6 Further Improvements

2.7 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.
- Desktop/Laptop
- Web Browser.

For training model, the system requirements are as follows:

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

2.8 Data Requirements

- We need balanced data in .csv format
- Input file feature/field names and its sequence should be followed as per decided.

2.9 Tools used

- **Python 3.8** – Programming language
- **NumPy** – NumPy is most commonly used package for scientific computing in Python.
- **Jupyter Notebook** – tool to do EDA
- **Scikit learn** – Machine learning library
- **Pandas** – Pandas is an open-source Python package that is widely used for data analysis and machine learning tasks.
- **Matplotlib, seaborn** – Data visualization
- **VS code** – IDE
- **Docker and apache airflow-** For continuous integration and depoyment

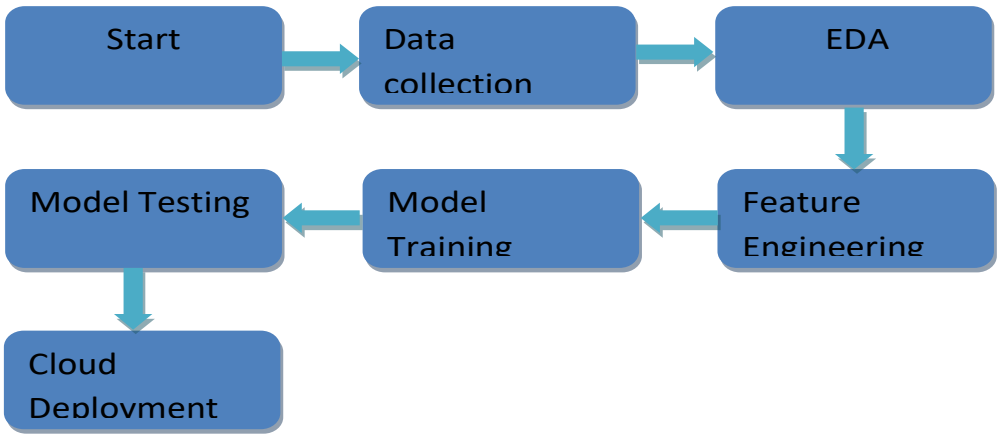
- **AWS** - Cloud deployment of application
- **MongoDB Atlas** – to retrieve, insert, delete and update the database
- **GitHub** – Source Code Version Control System

2.10 Constraints-nil

2.11 Assumptions- nil

3. Design Details

3.1 Application 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

If error encountered it should be displayed to user what went wrong.

3.4 Performance

The application is going to give user an estimation of premium, so it should be as accurate as possible otherwise it will lead to lose the business. Also model retraining is very important to improve performance.

3.5 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.

3.6 Application Compatibility

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

3.7 Deployment

Deploy application on AWS.

4 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.

5 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

6 Document Version Control

Date Issued	Version	Description	Author
14/11/2022	00	Initial Release-Draft created	Aniket Dumbre
18/11/2022	01	Modification in architecture	Aniket Dumbre