High Level Design
Insurance Premium Prediction

Prepared By: Mangesh Devkate

Data Science Intern

iNeuron.ai

Document Version Control

Version	Date	Author	Comments
1	1-12-2022	Mangesh Devkate	

Contents

Abstract

Machine learning is a part of Artificial Intelligence and it is a process in which machine is having ability to learn without being explicitly programmed. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, agriculture, health, insurance and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. In this project, we will estimate the amount of insurance premium on the basis of personal health information. Taking various aspects of a dataset collected from people, and the methodology followed for building a predictive model.

We analysed the personal health data to predict health insurance premium of individuals. We have used multiple regression algorithms namely Linear Regression, KNN, Decision tree, Random Forest, Adaboost, Gradient Boost, XGBoost, SVM for model building. Training dataset was used for training model and we have predicted the result. Predicted result was compared with actual data to test and verify the model accuracy. Accuracies of all the models were compared. Among all the models Random Forest and Gradient Boost performed better than other models. Gradient boosting is best suited model for this case as it has given best evaluation score compare to other models.

1.0 Introduction

1.1 Why this High-Level Design Document?

High Level Design in short HLD is the general system design means it refers to the overall system design. It describes the overall description/architecture of the application. 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

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.

1.3 Definition

Term	Definition
Database	Collection of all the information
IDE	Integrated Development Environment
API	Application Programming Interface
KPI	Key Performance Indicators
VS Code	Visual Study Code
EDA	Exploratory Data Analysis
KNN	KNearest Neighbors

2.0 General Description

2.1 Product Perspective

The Insurance premium prediction is a machine learning based predictive model which will help us to predict the insurance premium of the personal for health insurance.

2.2 Problem Statement

To develop an API to predict the insurance premium using people individual health data by analysing the following:

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

2.3 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 mentioned use cases. In first case, analysing how BMI value affect 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 API to predict the premium.

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

- Programming language: Python, HTML, CSS
- Libraries: NumPy, Pandas, Matplotlib, Seaborn, Plotly, Scikit-learn

- Web framework: Flask
- Cloud platform: Heroku
- IDE: Jupyter Notebook
- Code hosting platform: GitHub

- NumPy is most commonly used package for scientific computing in Python.
- Pandas is an open-source Python package that is widely used for data analysis and machine learning tasks.
- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- Plotly is an open-source data visualization library used to create interactive and quality charts/graphs.
- Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- Flask is used to build API.
- Jupyter Notebook 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.

2.7 Constraints

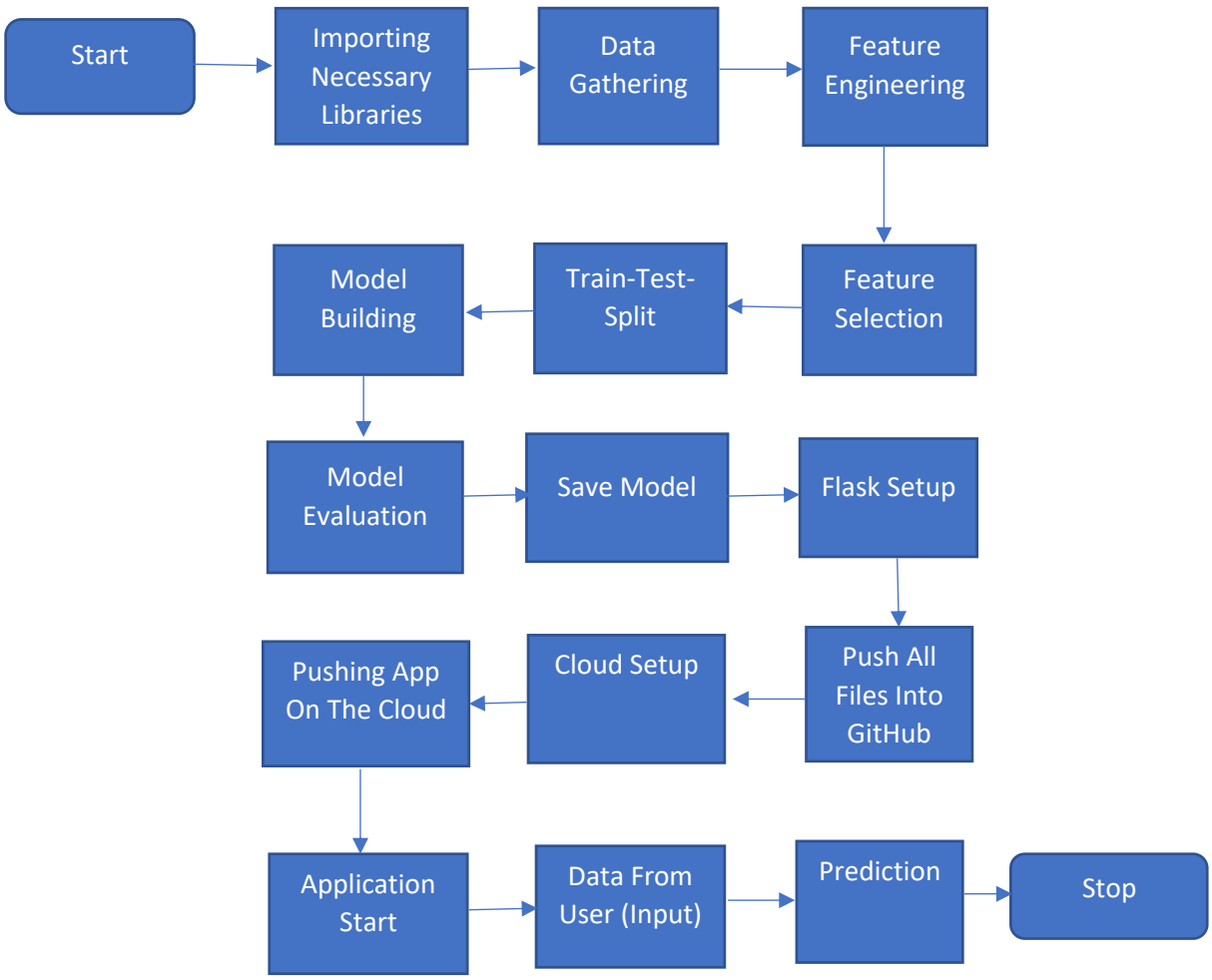
We only predict the expected estimating cost of insurance premium of customers based on their personal health information.

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

We have used Heroku cloud platform for our model deployment. Heroku is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud. Heroku is a cloud platform as a service (PaaS) supporting several programming languages. One of the first cloud platforms, Heroku has been in development since June 2007, when it supported only the Ruby programming language, but now supports Java, Node.js, Scala, Clojure, Python, PHP, and Go. For this reason, Heroku is said to be a polyglot platform as it has features for a developer to build, run and scale applications in a similar manner across most languages. Heroku was acquired by Salesforce in 2010 for $212 million.

5.0 Dashboards

In business computer information systems, a dashboard is a type of graphical user interface which often provides at-a-glance views of key performance indicators relevant to a particular objective or business process.

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.

Tableau: Tableau can help anyone see and understand their data. Connect to almost any database, drag and drop to create visualizations, and share with a click. Tableau offers drag and drop and other features such as multiple chart formats and mapping capabilities. The software is able to plot latitude and longitude coordinates and connect to spatial files

Microsoft Power BI: Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence.[1] It is part of the Microsoft Power Platform. Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. Data may be input by reading directly from a database, webpage, or structured files such as spreadsheets, CSV, XML, and JSON.

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 analysing 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.