HIGH LEVEL DESIGN (HLD)

Insurance Premium Prediction

Written By - Dhananjay K. Gurav

Document Version Control:

Date Issued	Version	Description	Author
10/12/2023	1.0	Initial HLD- V1.0	Dhananjay
26/12/2023	1.1	Updated requirements	Dhananjay

Abstract

Health insurance is an agreement in which an insurance company agrees to pay your medical expenses in exchange for an insurance premium payment. Most people do not buy Health insurance thinking that it will cost more money for premium but, the insurance premium price depends on several factors such as age, gender, body mass index (BMI), region, and other special factors like smoking. In this project we have built a model that can predict medical expenses based on several features of an individual such as age, physical/family condition and location so that, customers can work with any health insurance carrier and its plans and perks while keeping the projected cost from our study in mind. Also, it will help medical insurance companies to make decisions on charging the premium for particular customers.

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 interacts 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
EDA	Exploratory Data Analysis
IDE	Intergrated Development Environment
API	Application Programming Interface
VS Code	Visual Studio Code

2.0 General Description

2.1 Product Perspective:

The Insurance premium Prediction is a model that will help us to get estimate of Money/cost to spend on health, medical expenses so that person can choose suitable health insurance plan.

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 create API interface to get estimated premium price.
- To detect how BMI value affects the premium price.
- To detect how smoking affects the health expenses of an individual.

2.3 Proposed Solution:

The solution proposed here is an estimating premium of insurance or how much money they will need for medical expenses based on person's health condition and this can be implemented to perform above mentioned use cases. In the second case, if the model detects the BMI or smoking affects the premium, we will make people aware of it. And lastly, we will be creating an interface to predict the premium.

2.4 Further Improvements

2.5 Technical Requirements

The solution can be a cloud-based 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
- VS Code/ PyCharm IDE / Jupyter notebook

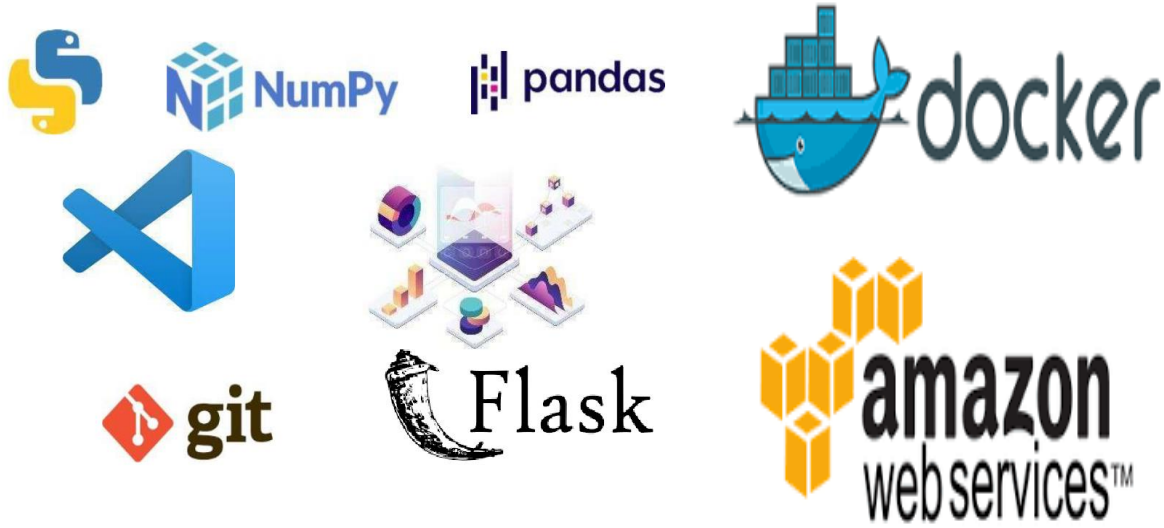
2.6 Data Requirements

Data requirements completely depend on our problem statement.

- Comma separated values (CSV) file.
- We will be required data of people with following attributes,

1. Age: Age of person
2. Sex: Gender/Sex of person
3. BMI: Body mass Index for Individual
4. Children: No. of children person has.
5. Smoker: Whether person smokes or not.
6. Region: The geographical region where a person lives.

2.7 Tools Used



Python programming language and libraries such as Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn as well as Flask Framework will be used to build the end-to-end solution for problem statement.

- Pandas is an open-source Python package that is widely used for data analysis, data preprocessing and machine learning tasks.
- NumPy is most used package for scientific computing in Python.
- Matplotlib and Seaborn are open-source data visualization libraries used to Create quality charts/graphs visualizations.
- Scikit-learn is widely used for a preprocessing, transforming data and to build machine learning models.
- VS Code is used as IDE (Integrated Development Environment)

• Jupyter notebook is used for writing and running python code in blocks for developing logic step by step.

• GitHub is used as code versioning control system.

• GitHub actions are used for continuous integration and continuous deployment of source code.

• Front end development can be done using HTML/CSS.

• Flask is microframework that can be used to build an interactive API.

• Docker is used to containerize the application

• AWS is used for deployment of the model and storing data artifacts.

2.8 Constraints

The Insurance Premium Price Prediction system must be correct enough that it does not mislead any estimate and as automated as possible and users should not be required to know any of the workings.

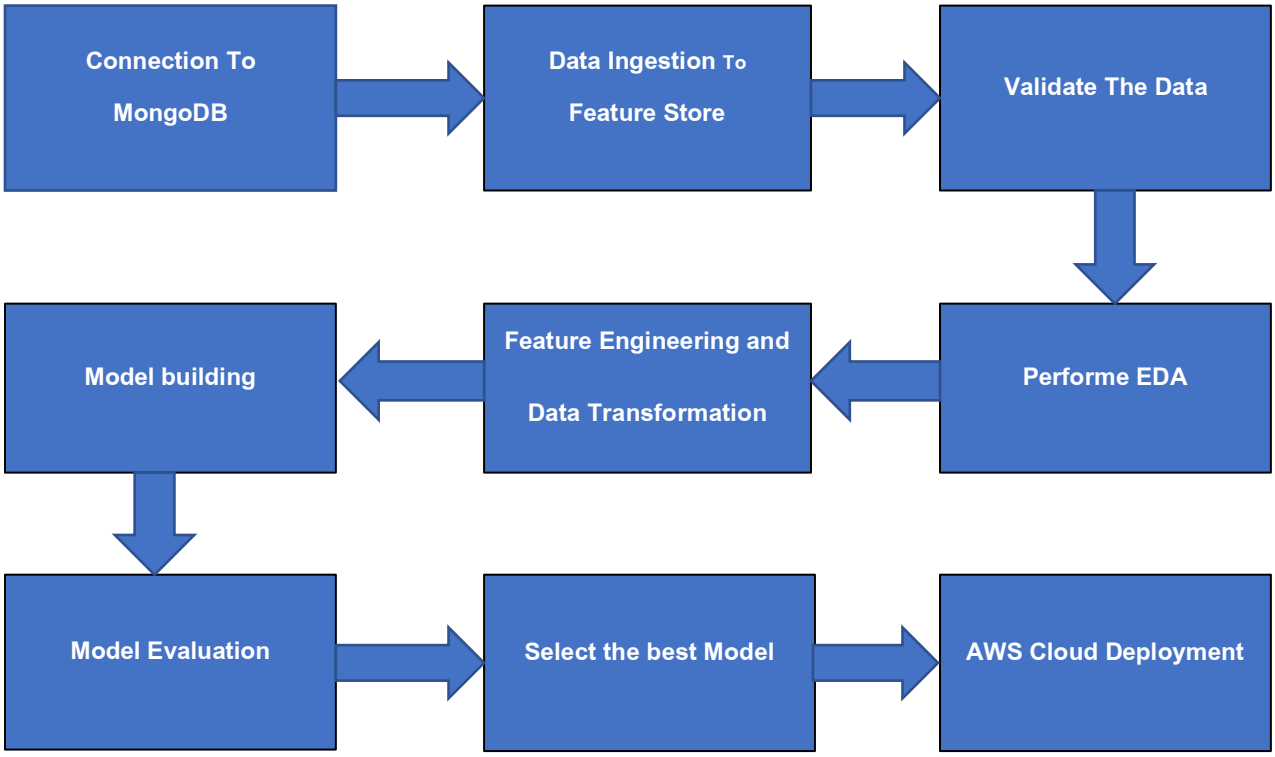
2.9 Assumptions

The main objective of the project is to develop an API to predict the premium for people based on 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:

For estimating health expenses, we will use machine learning based model. Below is the process flow diagram is as shown below



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

Errors handling should be done using the CustomException module so that after incurring the error, an explanation will be displayed as to what went wrong. An error will be defined as anything that falls outside the normal and intended usage

4.0 Performance

Performance of the machine learning based solution will be evaluated using Accuracy matrix(R2 score) and RMSE So that necessary action will be taken ASP if needed. Also, model retraining is very important to improve performance.

4.1 Reusability

The entire solution will be done in modular fashion and will be API oriented. So, in the case of 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 or local machine.

4.3 Deployment

The deployment of the app will be done using AWS cloud platform. CICD pipeline will get created by using GitHub actions to apply any modifications from source code to deployed model.

5.0 Conclusion

This system shows us the different techniques that are used to estimate the amount of premium required based on individuals' health condition. After analyzing it shows how the habit of smoking affects the amount of estimate premium. Also, significant difference between male and female expenses. Accuracy, which plays a key role in prediction-based systems from the results we will select the best working model for this problem in terms of accuracy/ R2 Score. Our predictions will help users to know how much premium they need based on their current health situation.