High Level Design (HLD)

Insurance Premium Prediction

Revision 1.0.1

Date: 07/01/2023

Description: Initial revision HLD

Document Version Control

Date	Revision	Description	Author
06/01/2023	1.0.1	Initial revision(1.0.1)	Anand Chavan

Contents

1. Abstract

To give people an estimate of how much they need based on their individual health situation. After that, customers can work with any health insurance carrier and its plans and perks while keeping the projected cost from our study in mind. I am considering variables as age, sex, BMI, number of children, smoking habits and living region to predict the premium. This can assist a person in concentrating on the health side of an insurance policy rather than the ineffective part.

2. Introduction

2.1 Why this High-Level Design Document?

The purpose of this High-level document (HLD) is to describe the design of the project in detail which can be used as a reference manual.

The HLD will:

- Present all the design aspects and define them in detail.
- Describe the user interface being implemented.
- Describe the software interfaces.
- Describe the performance requirements.
- Include design features and the architecture of the project.

2.2 Scope

The HLD document presents the entire structure of the project in parts, such as the data ingestion, data pre-processing, solution development and the deployment part along with their respective architectures. This uses non- technical to mild technical terms which should be understandable to the administrators of the system.

2.3 Definitions

Term	Description
EDA	Exploratory Data Analysis

IDE	Integrated Development Environment
PaaS	Platform as a Service

3. General Description

3.1 Product Perspective

The insurance premium predictor is a machine learning based regression model which helps us to predict the insurance premium based on variables, which will be used to determine prediction and validation of plans offered by different companies.

3.2 Problem Statement

The goal of this project is to give people an estimate of how much they need based on their individual health situation. After that, customers can work with any health insurance carrier and its plans and perks while keeping the projected cost from our study in mind. This can assist a person in concentrating on the health side of an insurance policy rather than the ineffective part.

3.3 Proposed Solution

The solution proposed here is a web application, which takes details of the main variables of insurance which contributes to its premium and those details will be taken by a machine learning model in the backend, which predicts the premium in dollars and displays in the front-end page to the user.

3.4 Technical Requirements

I used python version 3.7 with some important libraries to develop a machine learning model, which accurately predicts the Insurance premium based on its details.

Then, the model is used as a back-end software for a front-end web application which can be used by the users.

3.5 Data Requirements

For training and testing the model, I used the public data set available in Kaggle, “Insurance Premium Prediction” by nursnaaz

URL: https://www.kaggle.com/noordeen/insurance-premium prediction

Data dictionary as follows:

Name	**Data Type**	**Description**
Age	Integer	Input variable
Sex	String	Input variable
BMI	Decimal	Input variable
Children	Integer	Input variable
Smoker	String	Input variable
Region	String	Input variable
Expenses	Decimal	Output variable

3.6 Tools and Technologies used

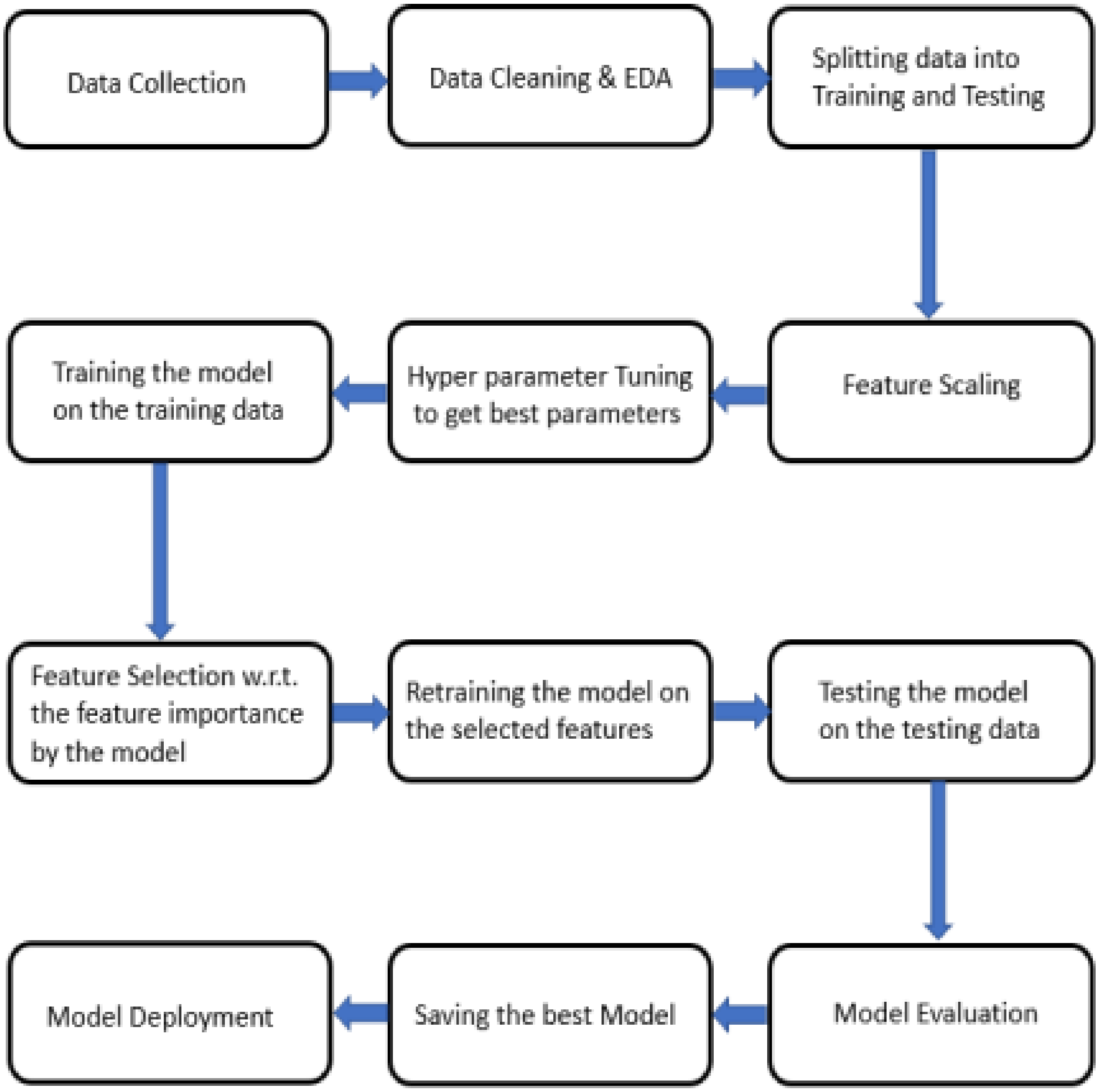
- Jupyter notebook is used for EDA and experimentation with various ML algorithms with the help of pandas, numpy, matplotlib, seaborn, scikit-learn, statsmodels and xgboost libraries.
- PyCharm is an IDE used for development and deployment of the solution with logging. Used python version 3.7 and libraries include logging ,pandas,numpy,scikit-learn, statsmodels, joblib, streamlit
- GitHub is used as a version control system.
- Deployed on the web using Render.

3.7 Constraints

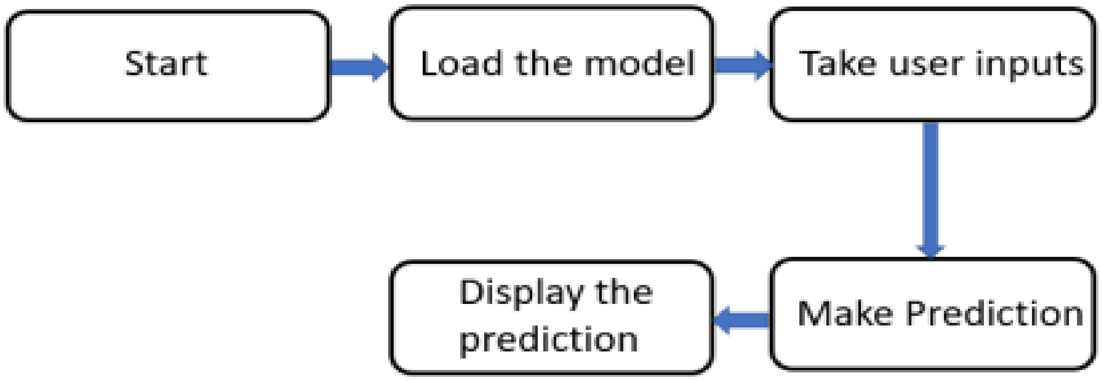
Prediction system must be user friendly, error free and users should not be required to know about any of the workings.

4. Design Details Data Collection

4.1 Process Flow


Data Collection
Data Cleaning & EDA
Splitting data into Training and Testing
Feature Scaling
Hyper parameter Tuning to get best parameters
Training the model on the training data
Feature Selection w.r.t. the feature importance by the model
Retraining the model on the selected features
Testing the model on the testing data
Model Evaluation
Saving the best Model
Model Deployment

4.2 Deployment Process



4.3 Event log

In this project, I used the “logging” library in both the development and deployment stages, which keeps logging the events at every step into the “.log” files. One of the advantages of event logging is, it makes debugging much easier, like we can directly go to that specific line of code, having errors.

4.4 Error handling

Used exceptions handling to catch the errors, so that they will be recorded in logs and ensure the smooth run, without getting terminated in the middle. Once the run gets completed, we can check the log files for the errors and can take an appropriate debugging action.

4.5 Performance

The ML based Insurance Premium prediction application is used for predicting the Insurance premium based on its age and other factors. So, it should be as accurate as possible, so that it will not mislead the user. Also, the model retraining is very important to keep it relevant if the new factors are added in future or to improve the performance.

4.6 Reusability

The code written and the components used have an ability to be reused without any problem.

4.7 Application compatibility

The different components or modules of this project use python version 3.7 as their interface between them. Each component has its own task to perform, and it is the job of the python version to ensure proper transfer of the information.

4.8 Resource utilization

In this project, any task may likely use all the processing power available in the system, until it is accomplished.

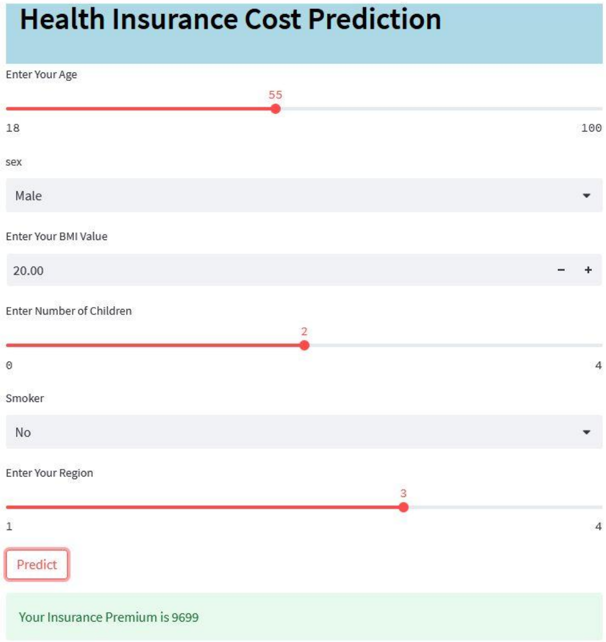
4.9 Deployment

Deployment of the application on the web using Render

URL: https://insurance-premium-prediction-pgzp.onrender.com

4.10 User Interface

Designed user interface using streamlit. It looks as per the below image.



Health Insurance Cost Prediction

Enter Your Age

55

sex

Male

Enter Your BMI Value

20.00

Enter Number of Children

2

Smoker

No

Enter Your Region

3

Your Insurance Premium is 9699

5. Conclusion

Insurance premium prediction is used to predict the premium of the personal for health insurance based on the given input entities, which enables the user to determine cost and save money on effective plans.

6. References

- https://scikitlearn.org/stable/modules/generated/sklearn.linear_mode.LassoCV.html
- https://www.google.com/url?sa=D&q=https://towardsdatascience.com/feature-selection-with-pandase3690ad8504b&ust=1637241300000000&usg=AOvVaw0pO 3x0h1T83PvMf_TFBEI7&hl=en
- https://towardsdatascience.com/3-techniques-to-avoidoverfitting-of-decision-trees-1e7d3d985a09
- https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html
- https://www.w3schools.com/