HIGH LEVEL DESIGN (HLD)

INSURANCE PREMIUM PREDICTION

Barnalikka Pradhan, Punith B C
iNeuron

Document Version Control

Date	Version	Description	Author
25.09.2022	1	Initial HLD – V1.0	Barnalikka Pradhan, Punith B C

Contents

Abstract

In today's age and time, health insurances has become a guarded necessity. There is awareness for buying heath insurances but very often the vision, planning, implementation and capital do not go hand in hand for finding the most suitable health insurance for themselves. Therefore, having a basic idea, of the predicted cost of their health insurance individually will save a lot of money and time.

This system created with the help of machine learning algorithms will give the people this very idea of predicted cost based on personal features of BMI, age, sex, children, smoker, region they belong to & their current expenses catering soley to the individual at large.

In this scenario, we have firstly trained and then tested the model using three different machine learning algorithms namely Decision Tree Regressor, Random Forest Regressor, Gradient Boost Regressor out of which the Random forest Regressor provides the highest accuracy test score proving its niche in being the best working algorithm of the three.

1. 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 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

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

Terms	Definitions
Database	Collection and storing of all information
IDE	Integrated development Environment
API	Application programming Interface
EDA	Exploratory Data Analysis
AWS	Amazon Web Services

2. General Description

2.1 Product Perspective

The health insurance premium system is a machine learning prediction model helping users to get individually catered estimated cost of their health insurances.

2.2 Problem Statement

To develop a web application which curates individual health insurance premium costs, using a dataset containing information of 1338 individuals according to the following features displayed by the individual pertaining to :
Whether the person is an avid smoker or not, will impact the cost of premium.
Age wise segregation to see the categories preferring health insurance premiums.
The body mass index of a person as well as the number of children they have will impact the cost of insurance premium.

2.3 Proposed Solution

The solution is to help people estimate their insurance premiums, taking into consideration their individual attributes of age, gender, BMI, smoker to non-smoker, region they belong to, by taking the most suitable machine learning algorithm which is the Random Forest Regressor in this case which will help predict the cost of their insurance premiums.

2.4 Technical requirements

The solution can be hosted on a cloud plateform like Heroku, AWS, Azure and definitely on your local machine.
Requirements for running machine learning algorithms are:
Pycharm/ Jupyter Notebook
Operation System : Windows, Mac, Linux
Stable internet connection

2.5 Data requirements

The data requirement varies according to the different projects. Here we had the data in a csv file format. There were six variables out of which sex, smoker and region were categorical variable which had to be encoded to transform them into numeric variables at large.

2.6 Tools used

Python programming language and frameworks listed below are used to build the whole model.

- PyCharm is used as IDE.
- Matplotlib , Seaborn are used for visualization of plots.
- GitHub is used as version control system.
- Cassandra used for insert, retrieve and update the database.
- Pandas used for data analysis, NumPy for scientific computation.

- Flask is used to build API.
- Scikit-learn used for machine learning.
- Front end development done using HTML/CSS.
- AWS is used for deployment of the model.

jupyter
PC
scikit learn
Cassandra
matplotlib
HEROKU
Flask
git
NumPy
pandas
seaborn

2.7 Constraints

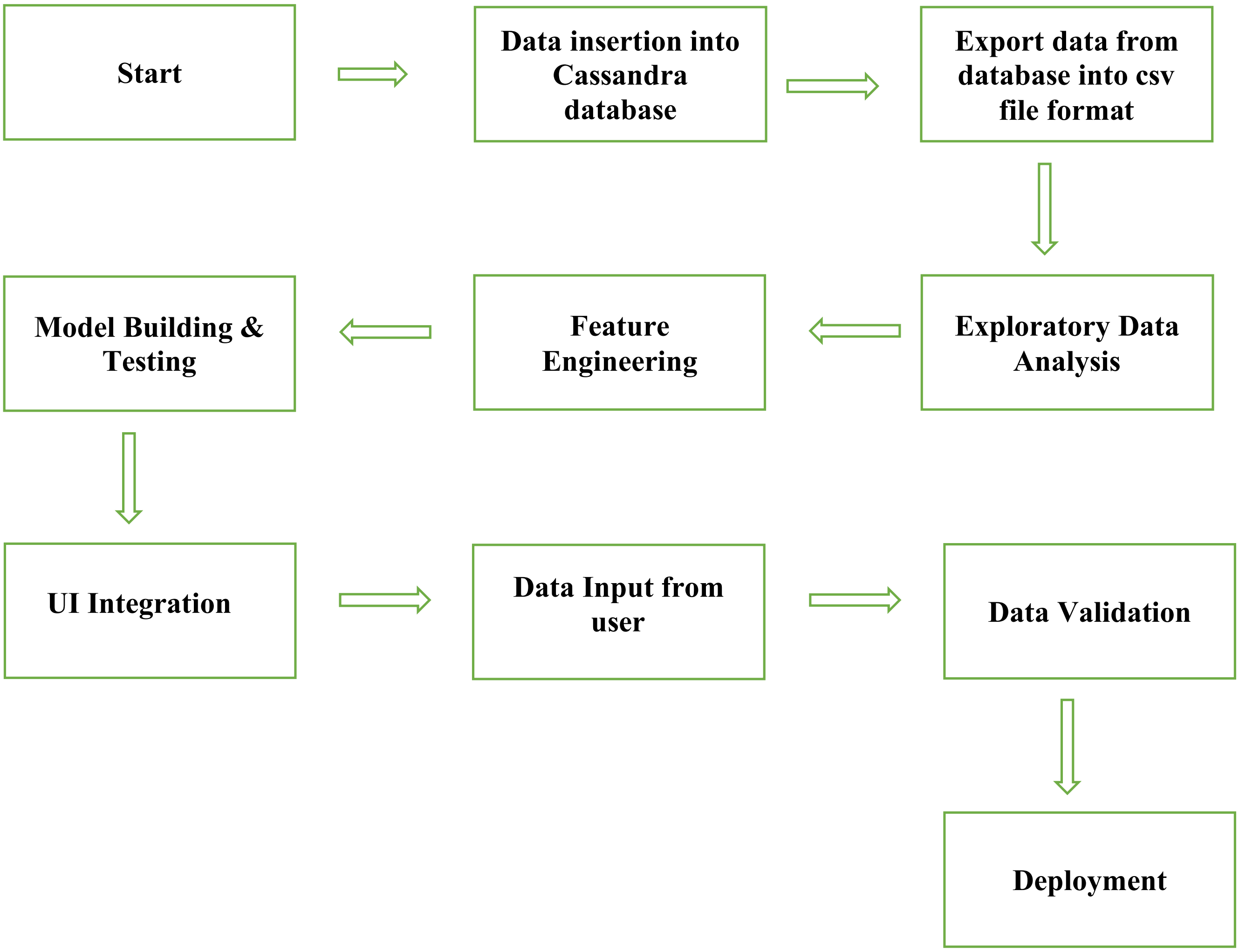
The health insurance premium app should be user friendly, automatic, smoothly running without the user having to understand any of the operations gone into making it.

2.8 Assumptions

The main objective is to create a web application which will be able to take in data from individual users and showcase the flexibility to figure out their cost for insurance premium according to their characteristic health situations.

3 Design Details

3.1 Process Flow


Start
Data insertion into Cassandra database
Export data from database into csv file format
Exploratory Data Analysis
Feature Engineering
Model Building & Testing
UI Integration
Data Input from user
Data Validation
Deployment

3.2 Event Log

The system logs every event with date and time so that the user can know how the process is running internally.

The system should be able to log each and every system flow.

System should not hang after using continuous loggings.

3.3 Error handling

Using the exception handling method, one can display the error encountered which can further help in handling and rectifying the error.

4. Performance

4.1 Reusability

The entire system is API oriented and made in such a modular fashion and can be reused and modified smoothly without any problem as such.

4.2 Application Compatibility

Python acts as an interface which will facilitate smooth transition of the tasks to be performed and transfer of data between the various parts of this project.

4.3 Deployment

Here we have used AWS as a cloud based plateform for deployment of the model.

5. Dashboards

Dashboards help to visualize and display Key performance indicators, the relation exhibited between variables, and other important business metrics.

5.1 Key Performance Indicators

Key indicators show us the relationship between expenses which is our dependent variable and other independent variables as listed below:

- Visualisation the ratio of males to females.
- Visualisation the ratio of smokers to non-smokers.
- Age of the individual has been put in age bin category for clearer visuals, and figuratively the age group of 20-30 and 40-50 each, make about 30 per cent of the current health insurers.
- Body mass index (BMI) which have been put into categorical bins, show that the obese people are the highest in number to have current medical insurance expenses which makes sense as high BMI is typically associated with higher risk of chronic disease.
- Another scatterplot suggests that body mass index (BMI) and expenses are positively correlated, where customers with higher BMI typically also tend to pay more in insurance premium.
- A boxplot indicates that individuals having 2 or 3 children consisted of having higher medical expenses throughout as compared to having more or less children than that.

6. Conclusion

The health insurance premium system will help the individual to get a predicted value of their health insurance premium. Majorly it can be seen that being a smoker, having a higher body mass index, gender based expenses do seem to impact/raise the cost of insurance premiums. Accuracy plays a major role in prediction henceforth the Random forest Regressor is choosen and used for predicting the cost of health insurance premium for which the individuals can save accordingly owing to their health situation.