HIGH LEVEL DESIGN

Insurance Premium Prediction

Created by Vikram Singh

GitHub Profile : VkasRajpurohit (github.com)

Document Version Control :

Date issued	Version	Description	Author
Jan 15th, 2022	1	Initial HLD V1.0	VIKRAM SINGH

TABLE OF CONTENTS

ABSTRACT

The purpose of this HLD (High Level Design) document is to add necessary details to the current project to represent a suitable model for coding. This document will help to detect the contradictions prior to coding and it can also be used as a reference manual for how the modules interact at high level.

HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1. 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.

1.1 Overview:

Title	:	Insurance Premium Prediction
Domain	:	Insurance

2. Domain Knowledge:

The **Insurance** word means to **the protection from financial loss**, basically it is a form of risk management against the risk of a contingent or uncertain loss. Insurance have many different-different types, some of the examples are health and life insurance (an individual must have these two).

The **Premium** word means to **an amount paid periodically** to the insurer by the insured **for covering risk**. In an insurance contract, the risk is transferred from the insured to the insurer. For taking this risk, the insurer charges an amount called the premium.

- Regarding the health/life insurance, during the **COVID**, people became more aware about the importance of health/life insurances, however still there is some kind of lag to understand the premium of the health/life insurances for an individuals.

2.1 Business Problem:

When we see insurance industries in business prospective - If an individual willing to take insurance plan, so they have to call the insurance company's helpline number or an insurance agent, and have to consult as per the need and different-different parameters, this is hectic process, also in this process, there are many of the negative points mentioned below.

1) Sometimes they have to wait for long hold.
2) May be they did not find the enough educated insurance rep, might lead to wrong info.
3) Delay in taking final decision and many more.

- Hence trying to resolve the above problems using Machine Learning Algorithms.
- Trying to build a user friendly ML model which can save time and effort for an individuals to reach at the insurance premium estimate with accuracy.

3. Product Understanding:

While taking insurance policy, an individual have to contact an insurance agent or insurance representative and provide required information to reach at the premium for the insurance policy.

However, it is hectic and time-consuming process. To fill-up this gap, build the application which allows an individual to insert the required information and get the **Insurance Premium** as result.

4. Data Requirements:

For the Problem statement data collected via Kaggle platform.

The dataset contains 1338 observations (rows) and 7 features (columns).

- There are **4 numerical features** (age, bmi, children and expenses) and **3 categorical features** (sex, smoker and region).
- Unique values for categorical features- sex: 2, smoker: 2, region: 4.

Features:

- **Age:** Age is the domain feature, as the age increases, insurance premium is more.
- **BMI:** it is Body mass index. It is a value derived from the mass and height of a person. The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2, resulting from mass in kilograms and height in meters.
- **Children:** Number of children an individual have.
- **Expenses:** It is the target/dependent feature. It is the overall medical expense/premium yearly.
- **Sex:** Sex of an individual i.e. male or female.
- **Smoker:** It is also a domain feature, an smoker individual have to pay high premium compared to a non-smoker.
- **Region:** It is representing region of an individual belongs. It have four unique values i.e. southeast, southwest, northwest, northeast.

5. Expected Solution:

- Build a solution that should be able to predict the premium of the an individual health insurance based on given features in dataset.

The purposes of this case study is to look into different-different **features** (related to the insurance domain) and observe their relationship between/among them. Based on several features of an individual such as **age**, **physical/family condition** and **location** against their existing medical expense, to be used for predicting future medical expenses of individuals that help medical insurance or individuals to make decision on charging the premium.

6. ML formulation of the business problem:

First Cut Approach

As per the problem statement, we need to predict **Expenses.**

- As checked dataset, it is labeled data i.e. we can go with the supervised machine learning techniques.
- **It is a Regression problem**.

- For Regression problem, we can go with Linear Regression, SVM, Decision Tree Regressor and Ensemble techniques (RF, GB), also the **StackingRegressor**.
- With Linear Regression, have to keep in mind basic 4 assumptions

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

- ✧ QQ plot to check if normally distributed.
- ✧ If not normally distributed, convert to normal distribution using Log transform, box-cox transform, exponential transform, power-low transform.

- If data is non linear, we can go with SVM, Decision Tree Regressor and Ensemble techniques, however time complexity is more for these compared to Linear Regression.

- First we will perform EDA.
- We can check, dataset is linear or non linear by correlation.
- We will try to build different-different ML models and try to improve performance by hyper-parameter tuning, the model which will give the best **cross_validation_score**, we will select that one.

7. Business Constraints:

1) **Time :** Latency is really not a major issue, even a few seconds of the latency is considerable.
2) **Accuracy:** Accuracy is a vital constraint, high accuracy in predicted premium with actual premium will build-up a trust, which will lead to referral i.e. more user, more business.
3) **Interpretability:** Model should be easy to interpret and user friendly.

8. Tools & Technology Requirements:

Tools & Technology	:	Python | Data-Preprocessing | EDA | Feature Engineering | Machine Learning | Flask-API | HTML | GitHub | Heroku | AWS
IDE	:	PyCharm | Google Colab

9. Conclusion:

Insurance Premium Prediction is a Machine Learning Algorithms based model. For the Problem statement data collected via Kaggle platform and built an end-to-end deployment ML model.

User will enter the required values & hit Get Premium, and it will show the **premium** as result.

---------- **End of HLD** ----------