Prediction of Health Insurance Premium

Amey Shankar Basangoudar

Gokul Narain Natarajan

Abstract:

The goal of this project is to build a predictive data science model using python programing. The motivation behind the project is make people aware of the costs of Insurance premium and plan their healthcare expenses efficiently. The major problem today is that people lack the resources to comprehensively come up with the right insurance premium based on factors like age, weight and other demographics. To solve this, we have come up with three best models – one with normal algorithm, 2nd with the bagging ensemble model and the 3rd with the boosting ensemble model. Any data given to this model will be fed into all three models and the best will be taken as the final suggested premium price.

1. INTRODUCTION:

1.1. Objective:

Health Insurance is a must in this country for all people. Without health insurance the medical bills tend to be exorbitant. A person may have an insurance plan from either a private company or the government, so it is imperative for all insurance companies to have a model that predicts the right premium amount for each person. Any general applicant can also use this model to know their own approximate premium amount. The premium amount needs to be decided based on various parameters of a person.

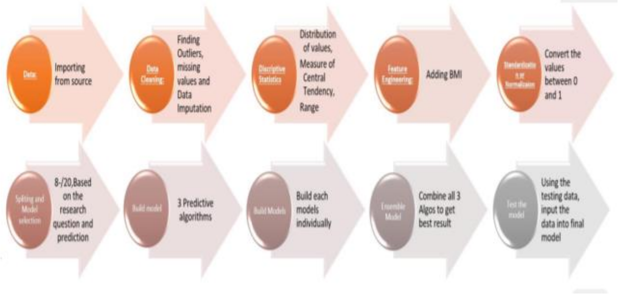
1.2. Importance of problem:

For the financial year 2020, 8.6% of US population did not have any insurance throughout the year to support their medical expenses. 66.5% of US population had enrolled for private health insurance (both self and employed) [1]. In the above scenario, for the former case, this model would aid a person in predicting their health insurance. In the latter case, this model will help the person to pick a better insurance plan. This model can also aid companies and organizations in predicting the insurance amount of any applicant.

1.3. Our understanding of problem and solution:

The cost of healthcare insurance premium is whimsical and if a consumer knows the basic cost of the premium, the consumer can plan his financials for buying a health insurance accordingly. To help in predicting the cost of the premium, we have designed an ensemble predictive model of 3 machine learning algorithms and come up with the best algorithm for each observation for accurate prediction. I also help insurance companies to better access the customers background and come up with the consistent premium pricing across the industry.

The below figure shows us the analytical framework of our project that we build:


Data
Importing from source
Data Cleaning
Finding Outliers, missing values and Data Imputation
Discriptive Statistics
Distribution of values, Measure of Central Tendency, Range
Feature Engineering
Adding BMI
Standardizatio n or Normalizaion
Convert the values between 0 and 1
Spliting and Model selection
8-/20,Based on the research question and prediction
Build model
3 Predictive algorithms
Build Models
Build each models individually
Ensemble Model
Combine all 3 Algos to get best result
Test the model
Using the testing data, input the data into final model

2. DATA ANALYSIS and MODELING

2.1. Data Analysis

2.1.1. Our analysis anticipation

The dataset that we used consists of 12 different parameters based on which a predictive model was designed. The target variable is the health insurance premium amount in dollars. This model can be deployed and made available to any insurance company as well as for the general public.

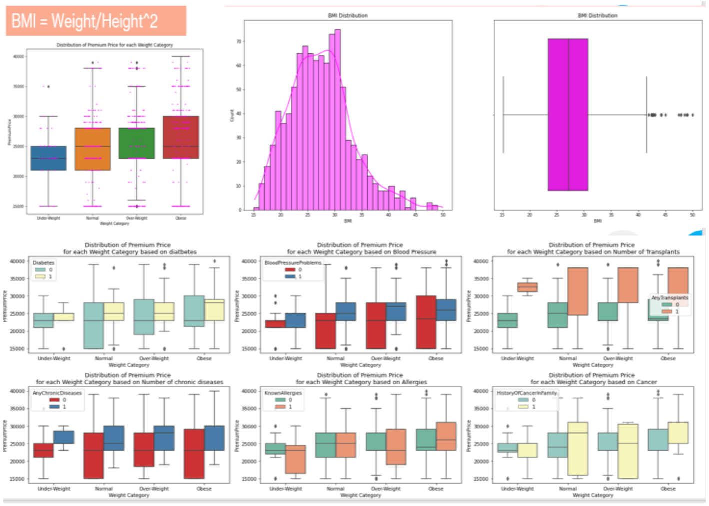
The figure below shows the categorical and numerical variable that are present in the dataset.

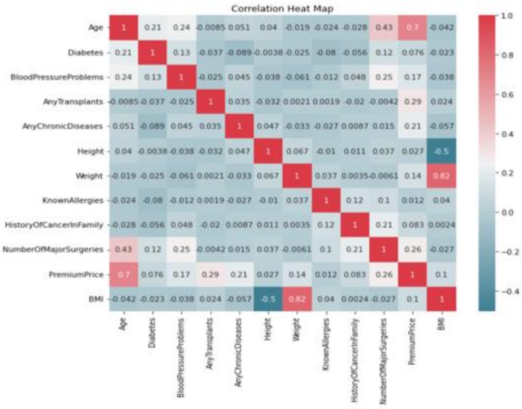
```
Categorical variables:
Diabetes
BloodPressureProblems
AnyTransplants
AnyChronicDiseases
KnownAllergies
HistoryOfCancerInFamily

Numeric variables:
Age
Height
Weight
NumberOfMajorSurgeries
PremiumPrice
```

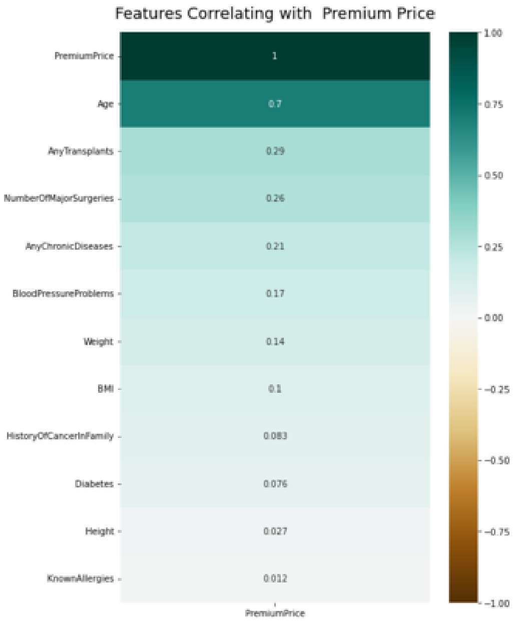
We anticipated that all the variables will have some good correlation with the premium price of atleast 0.5. However, Couple of things were surprising to us in the Pearson correlation analysis result– All of the correlation values were less than 0.5 other than Age and also the correlation of cancer with the premium price was very less.

Additionally, we conducted literature review and found that BMI is one of the major factors for computing the health insurance premium price in the real world. Hence we feature engineered a new column called "BMI" and added it to our dataset. The correlation between weight and BMI is strong (i.e 0.8), this is due to the fact that BMI is directly proportional to weight. Next, we converted the continuous numeric variable, BMI, into 4 different categories namely Under-Weight, Normal, Over-Weight and Obese. Surprisingly, the BMI had a weak correlation of only 0.1 with premium price.





- Pearson analysis



2.2. Methods Used:

Import and analyze data – **Numpy and Pandas**

Data Cleaning – **Numpy and Pandas**

Statistical analysis – **Scipy and Numpy**

Build visualization to see the distribution of data – **Matplotlib and Seaborn**

Normalizing & standardizing the data – **Scikit-Learn**

Correlation analysis – **Numpy, Scipy (Pearson) and corrplot**

Gradient Boosting (XGBoost), Multiple Linear regression and Random Forest— **Scikit-Learn**

2.3. Model:

We have used the ensemble model of 3 Machine learning algorithm – Multilinear regression, Random forest and XGBoost. We shall look into the each model one by one:

Multilinear regression – It is used to predict the dependent value for multiple independent features. Here we have 11 features (including the feature engineered BMI column) to predict the premium price (dependent variable). For the given input, the model will draw a best fit between dependent and independent variable and predict the most applicable value.

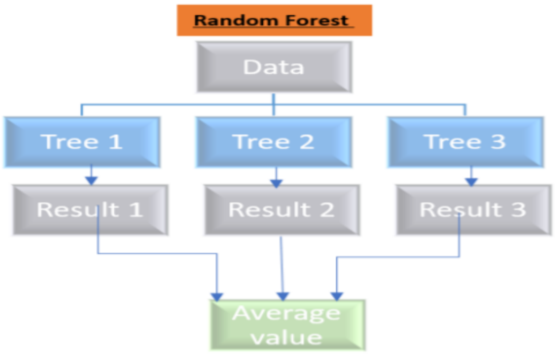
$$Y=\beta_0+\sum \quad \beta_i\check{X}_i$$

Where Y is the dependent variable, beta(s) is the constant and the X(s) is the independent variable.

For the given dataset, we have got 64.84% accuracy. The advantage of Multilinear regression is easy to find the association of independent and dependent variable and so we have selected it as one of our algorithms.

2.3.1. Random Forest:

Random forest is the bagging of multiple Decision Tress and select the final value based on ranking/voting. Lets say we have 100 trees where 51 says premium price as '$25000' and 49 says premium price as '$20000', the model will end up in selecting '$22500', the average value. Since we are using it for regression sometime equal voting or multiple values will occur, in such cases, average of result of various trees. Python allows the user to select the number of trees to be used as one of the parameters (n_estimatorsint) – the default value is 100 and we used the same. The advantage of Random forest is that it given more accuracy with the bootstrap aggregation of multiple trees and avoid the problem of overfitting. The accuracy of this model for our dataset is 80.24%.



$ni_{j} = w_j c_j - w_{left(j)} * c_{left(j}) - w_{right(j)} * c_{right(j)}$

ni $_{j}$ = The importance of node j

W_j = Weight no of samples reaching node j

c_j = The impurity value of node j

$left_{(j)}$ = child node from left split on node j

$right_{(j)}$ = child node from right split on node j [2]

2.3.2. XGBoost:

XGBoost is the boosting of multiple decision trees which computes the accuracy of each tree and selects the result with the highest accuracy that has less error than the previous tree. Like Random forest, we have an option to give the number of trees and depth of the trees as the parameter and we leave it as a default value for our model (100 and 6). The advantage of XGBoost is less bias and variance due to the sequential learning concept. The accuracy of this model for our dataset is 73.54%.

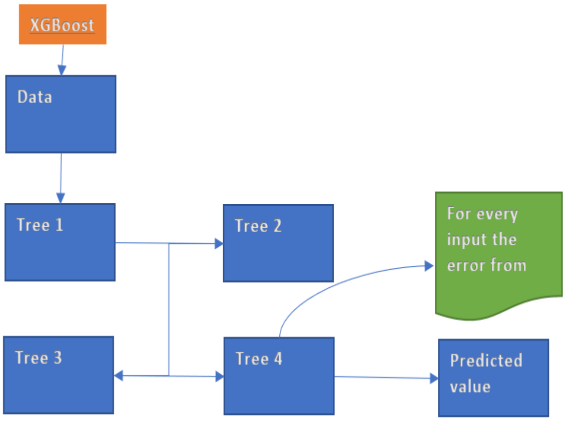
$F_m(X)=F_{m-1} + h_m$

F_m is the final model

F_{m-1} is the initial model

h_m is the next model to Fn-1

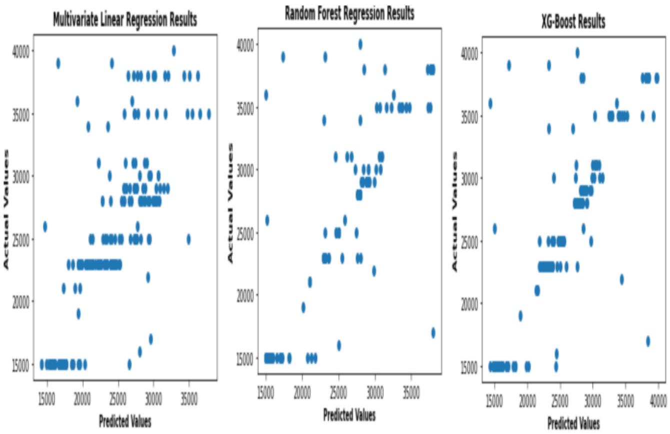
X is the given data [3]

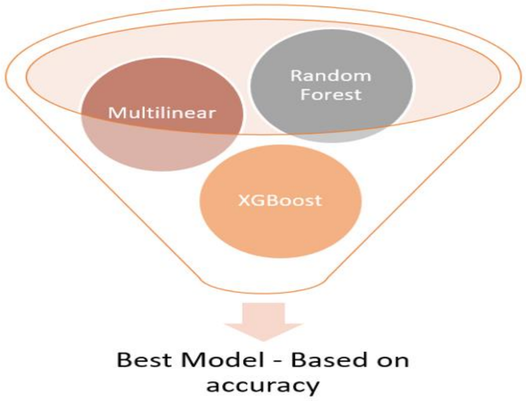


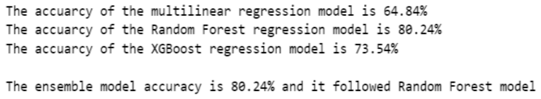
2.3.3. Ensemble Model:

The data is fed into the ensemble model of all 3 model and every time when a new data comes, the model will automatically tune and choose the best out of 3 models and gives the predicted value. And we used R square to find the accuracy of our model.

The figure below shows the Actual Vs Predicted for each model:







```
The accuarcy of the multilinear regression model is 64.84%
The accuarcy of the Random Forest regression model is 80.24%
The accuarcy of the XGBoost regression model is 73.54%

The ensemble model accuracy is 80.24% and it followed Random Forest model
```

3. **Conclusion:**

- The ensemble model chose the Random Forest Algorithm as the best model for our dataset. But this may vary again based on the dataset that the customer feeds into the ensemble model
- Furthermore, the ensemble model can be improved by incorporating more Neural Networks algorithms. Other future scopes of the project involve’s development of mobile or web application for deployment
- We found that age and number of surgeries plays a vital role in determining the premium cost
- The need for good healthcare is growing in demand owing to the pandemic. This project can aid a lot of people to make better decisions in selecting the best payer Health Insurance plan.

Our Python code:

Basangoudar_Nataraj
an_DS_5010_Final_Pr

GitHub link - https://github.com/gokulnarain/DS5010

Readme – Please change the file path name to your local file/folder path

References:

[1]: “Health Insurance Coverage in the United States: 2020” dated Oct 18th, 2021. Retrieved from https://www.census.gov/library/publications/2021/demo/p60-274.html

[2]: “The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark” dated May 11th, 2018. Retrieved from https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3

[3]: “An End-to-End Guide to Understand the Math behind XGBoost” dated Sept 6th, 2018. Retrieved from https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/