MEDICAL INSURANCE PREMIUM PREDICTION

Predicting yearly Medical cover cost ₹

Introduction:

The Health Insurance Premium is the amount of money you need to pay periodically to an insurer in order to avail the medical coverage as well as to ensure that the policy remains in force. Health insurance premium calculator facilitates you to calculate your mediclaim premium, based on your insurance needs.

Problem definition:

To create a model that predicts the yearly medical cover cost.

Content:

The Dataset contains Health Related Parameters of the customers such as whether they have some medical conditions or not and if they have gone through some surgeries or not and also their age, height and weight. We have nearly 1000 records. The premium Price is in INR(₹) currency and showcases prices for a whole year.

Practical Implementation:

There are various insurance agencies with different kinds of premium plans. This will be helpful for predicting the premium cost based on several factors.

Data Source:

Dataset is taken from the kaggle. Medical insurance dataset

Data:

Our dataset has 986 no. of rows and 11 columns, which are already in encoded format.

Features:

- Age
- Diabetes
- BloodPressureProblems
- AnyTransplants
- AnyChronicDiseases
- Height
- Weight
- KnownAllergies
- HistoryOfCancerFamily
- NumberOfMajorSurgeries

Target:

- PremiumPrice

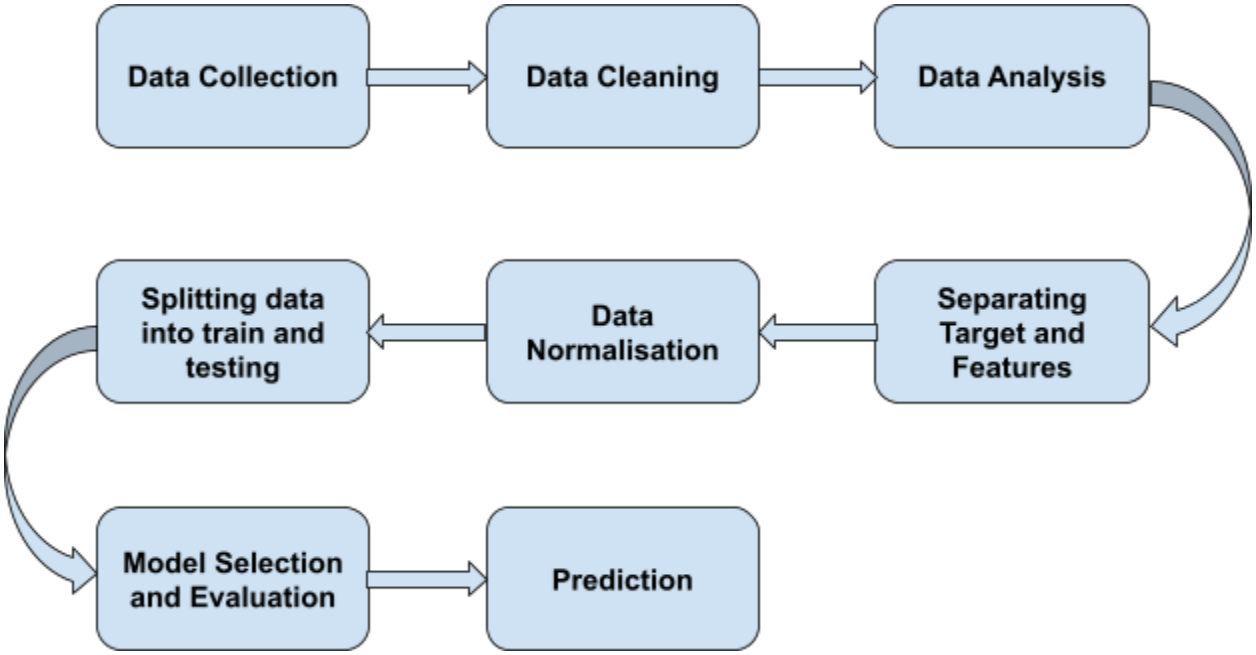
  (continuous variable)

Connection between the Features and the target:

As per our dataset we need to use supervised ml algorithms. Since the target features a continuous variable, we have to use regression algorithms to predict our target features.

Methodology:

In this section we are going to walk through the life cycle of a data science project.

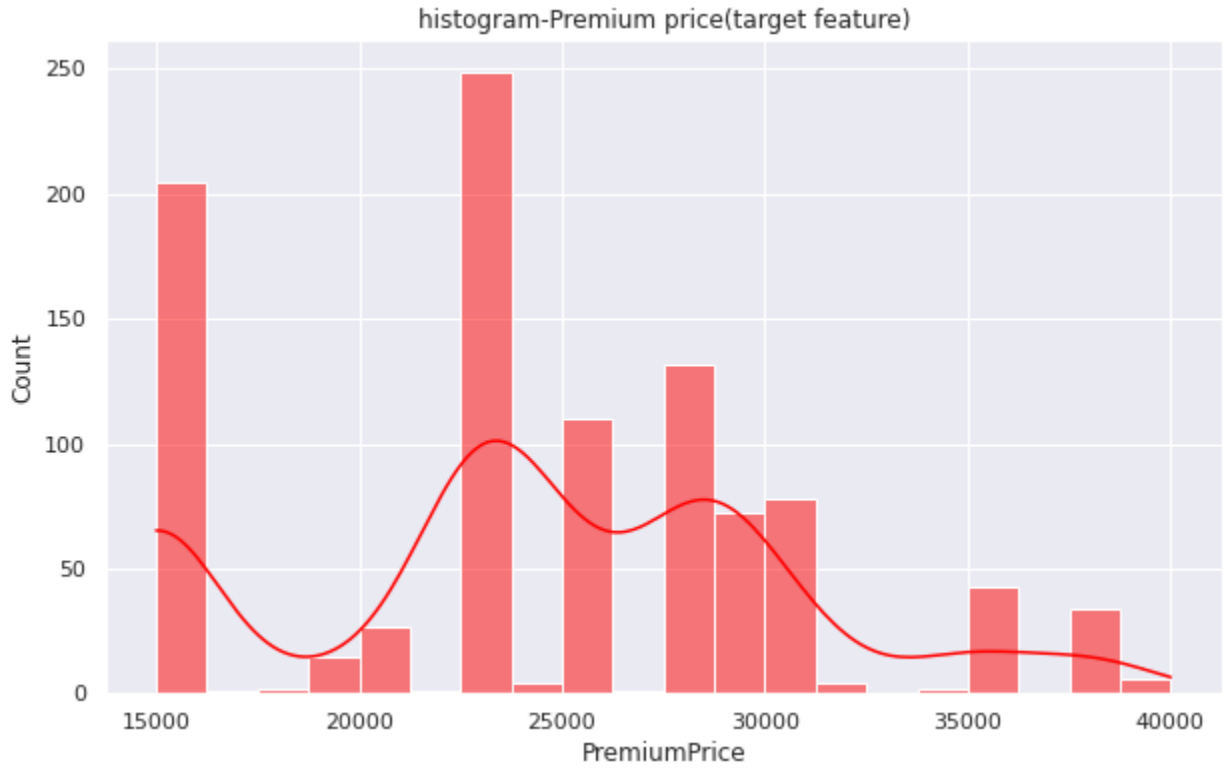


Our data is already in encoded format, so we don't have to do feature engineering. We have already talked about data collection .In this section we are going to explain our data cleaning methodology, Findings of our analysis, data transformation, splitting of data into training and testing, model selection and evaluation.

DATA CLEANING & DATA ANALYSIS:

As a first step of data cleaning, I searched for the null values. But there wasn't any. There are categorical values in the dataset such as diabetes, blood pressure problems, any transplants and so on but they are already presented as label encoded format. Then I look for the consistency of all the features. There are no outliers found in any of the features.

By looking at the histogram of the target values in the training set, I found that it is not following a normal distribution.

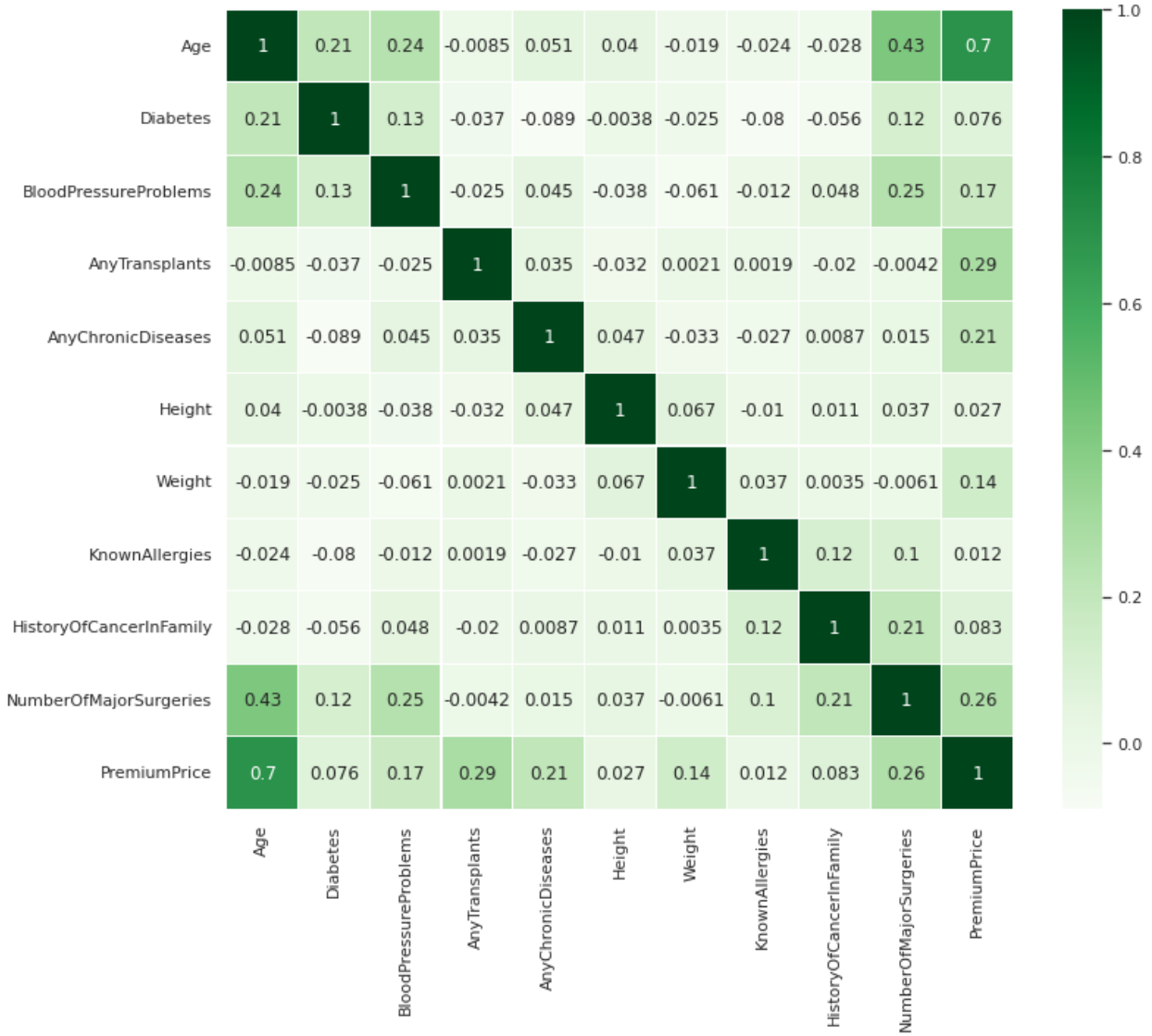


Our target feature is a continuous variable so we need to use regression models.

Now going to further data visualisation.

How do all the features and target values correlate?

To get an idea of this question I plot a heatmap with the correlation values. As a result I got the figure below.



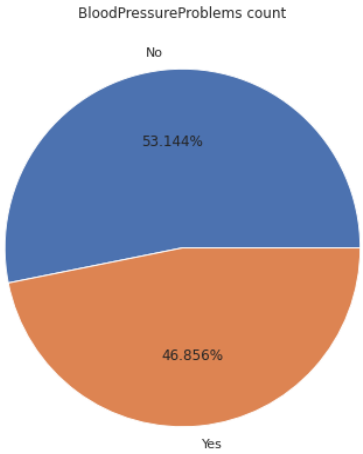
From the above heatmap we can see that there is no feature having a correlation value >.8. But from the heatmap we can see that age is having comparatively high correlation with the target feature.

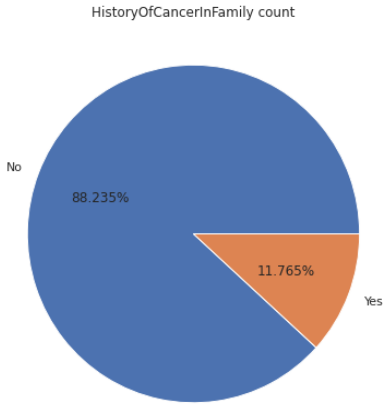
The other features having a good correlation with the target feature are any transplants, any chronic diseases and no. of major surgeries.

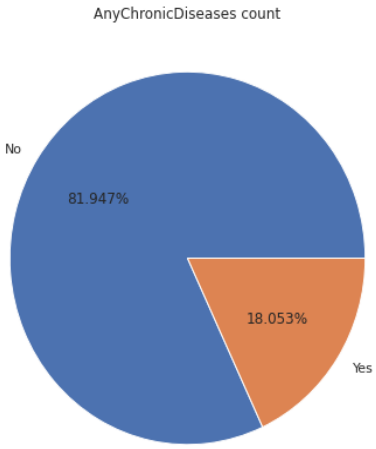
We can also see that there is a correlation between age and features such as diabetes, blood pressure problem and no. of major surgeries.

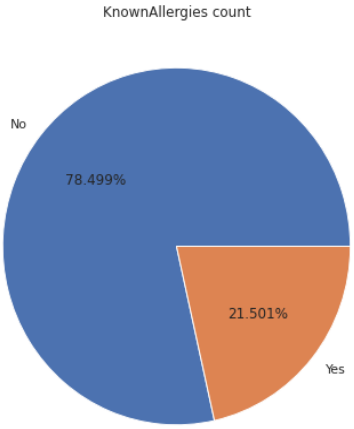
Based on these correlation we are analysing our data,

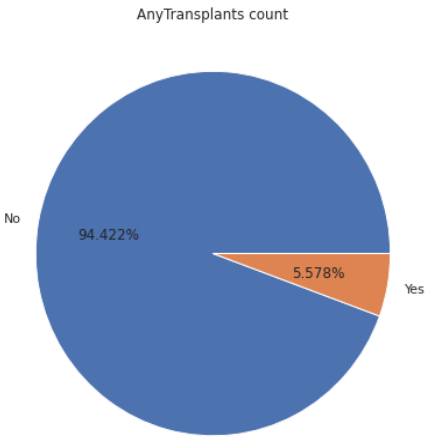
What percent of the data represent each category of feature ?

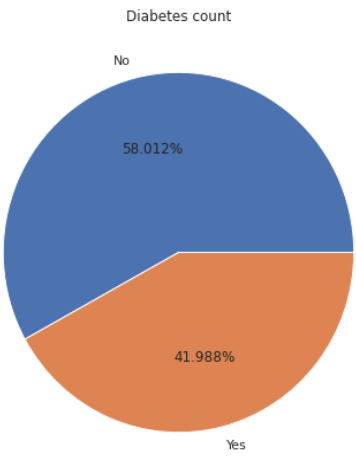


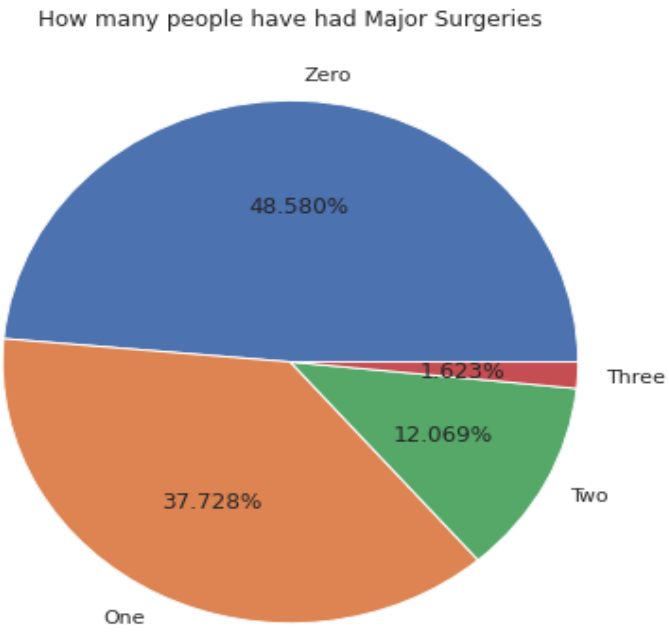












From these pie charts, we can see that around 53.14% of the total records, we have individuals that do not have the blood pressure problem and around 46.85% with blood pressure problems.

Around 88.23% of the total records are of individuals without cancer history in the family.11.67% of records with cancer history in the family.

Around 81.94% do not have any chronic disease and around 18.05% with some chronic diseases.

Around 79.49% of records do not have any known allergies, whereas, 21.50% are having some known allergies.

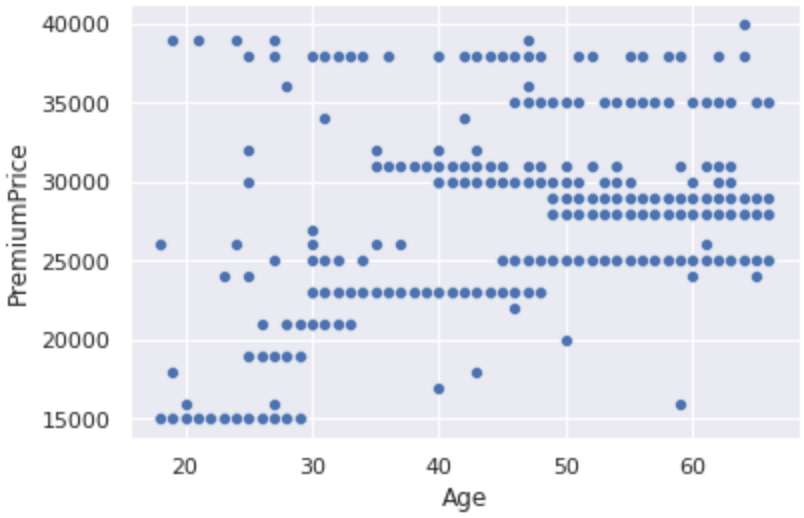
Around 94.42% of records are of those people who don't undergo any transplants and only 5.78% of records are showing any transplant history.

Around 58.012% are people without diabetes and the remaining 41.98% with diabetes.

Considering the case of no. of major surgeries. The majority do not undergo any surgery i.e., around 48.58% of zero surgeries, 37.72% with one surgery, 12.069% with two surgeries and only 1.623% have undergone three surgeries.

How are features related to the target feature?

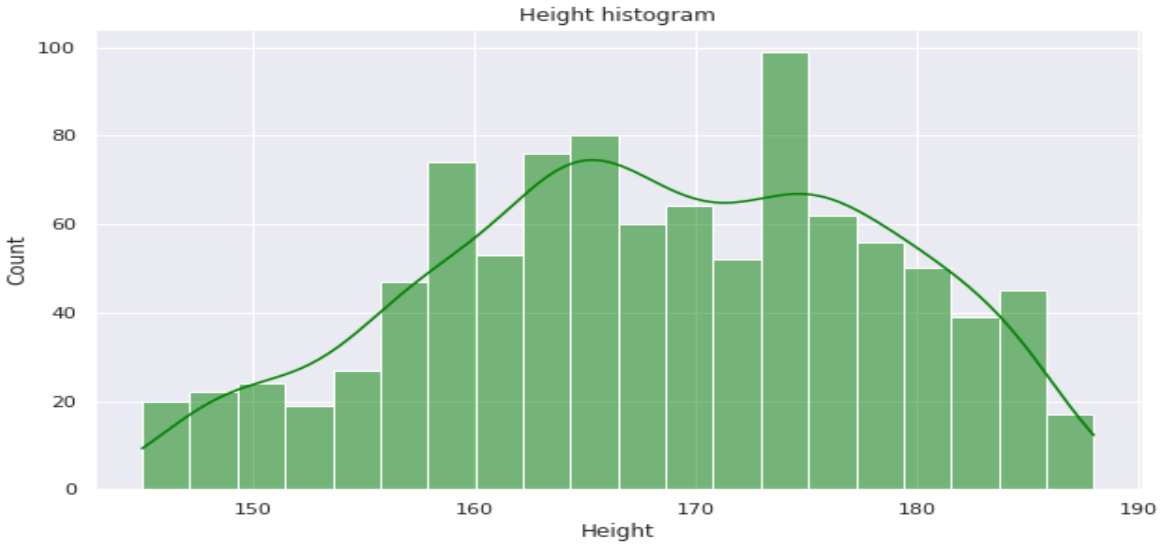
Age is the feature having higher correlation with the target feature. So first we are analysing how age is related to premium price.

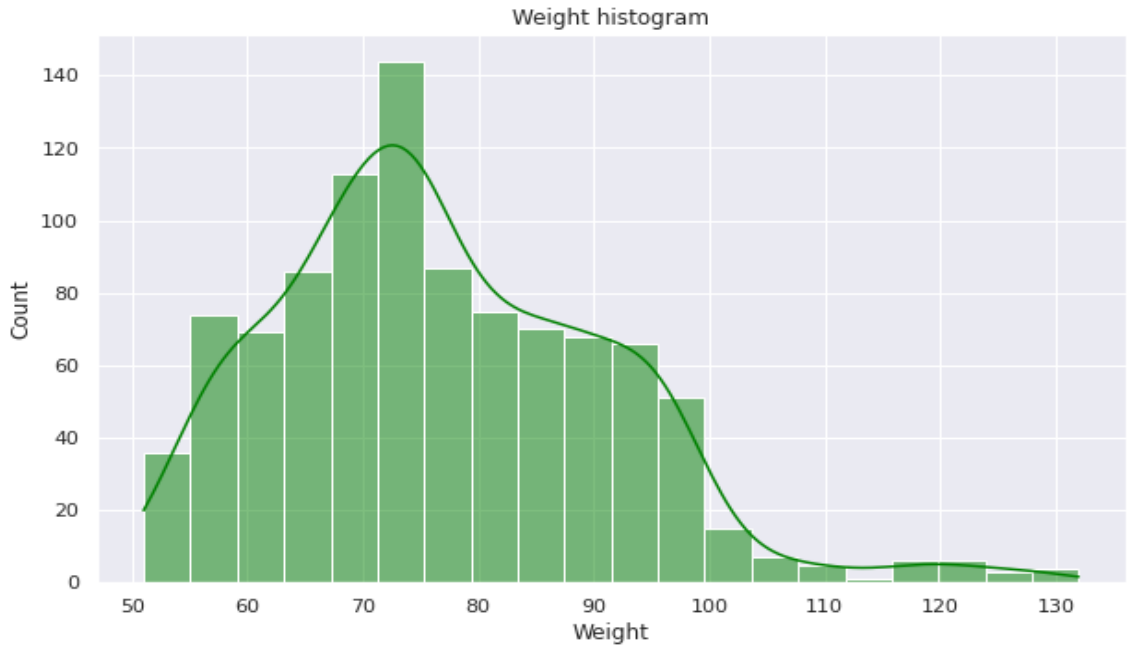


From both these figures we can see that as age increases the premium price also increases. That may be due to some health conditions or some other factors. Most of the people between 18 to 30 years pay a premium of 15000. Few people between age 25 to 30 years a premium of around 19000.

Histogram of Height and weight:

From the below hist graphs we can see that the most of the height is greater than 160 and the weight is less than 100. We are getting a skewed distribution for both the height and the weight.



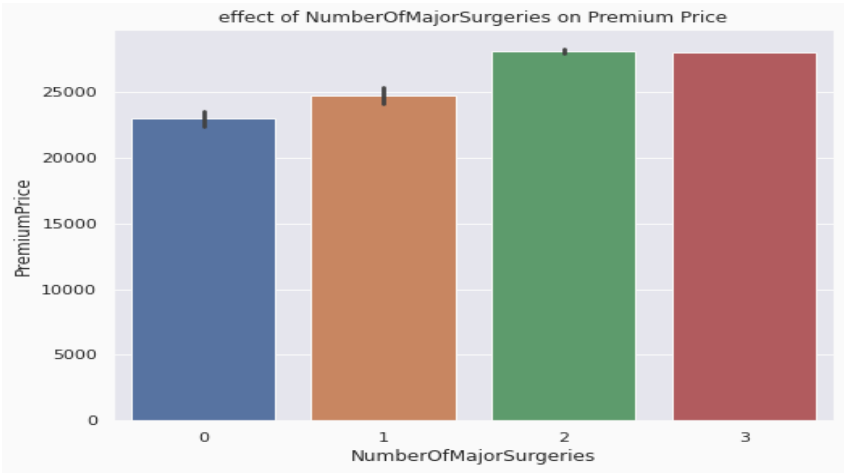


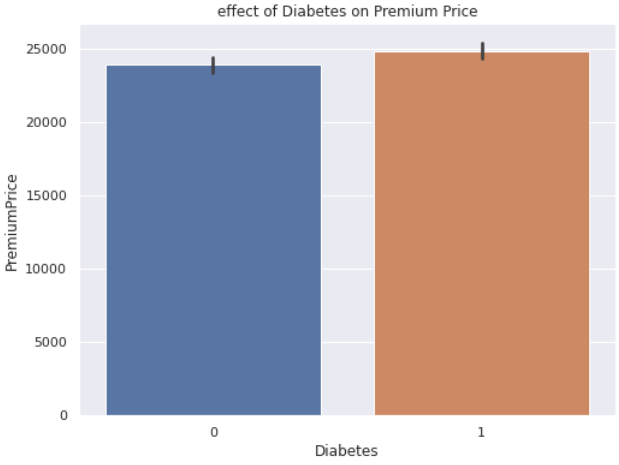
Now, let us analyse what is the relation between other features and the target feature?

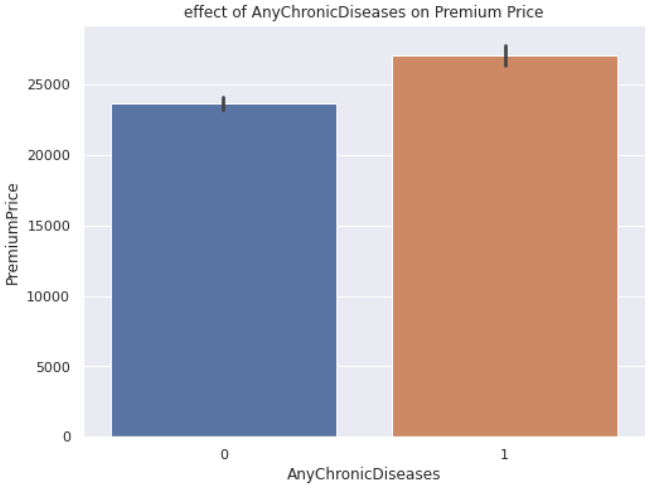
I used barplot for visualising this. For creating all the barplot I created a function called mybarplot() with some arguments. So that it will be helpful for me to create a barchart multiple times.


effect of BloodPressureProblems on Premium Price
PremiumPrice
BloodPressureProblems
0
5000
10000
15000
20000
25000
0
1


effect of AnyTransplants on Premium Price
PremiumPrice
AnyTransplants
0
5000
10000
15000
20000
25000
30000
35000
0
1


effect of NumberOfMajorSurgeries on Premium Price
PremiumPrice
NumberOfMajorSurgeries
0
5000
10000
15000
20000
25000
0
1
2
3


effect of Diabetes on Premium Price
PremiumPrice
Diabetes
0
5000
10000
15000
20000
25000
0
1


effect of AnyChronicDiseases on Premium Price
PremiumPrice
AnyChronicDiseases
0
5000
10000
15000
20000
25000
0
1

Interpreting the above barcharts.

The first barchart depicts the effect of blood pressure problem in premium price prediction. We can see that the individuals having blood pressure problems are paying a higher premium price than those without any blood pressure problems. Even though there is no drastic chance in the premium price we can see this in the bar chart.

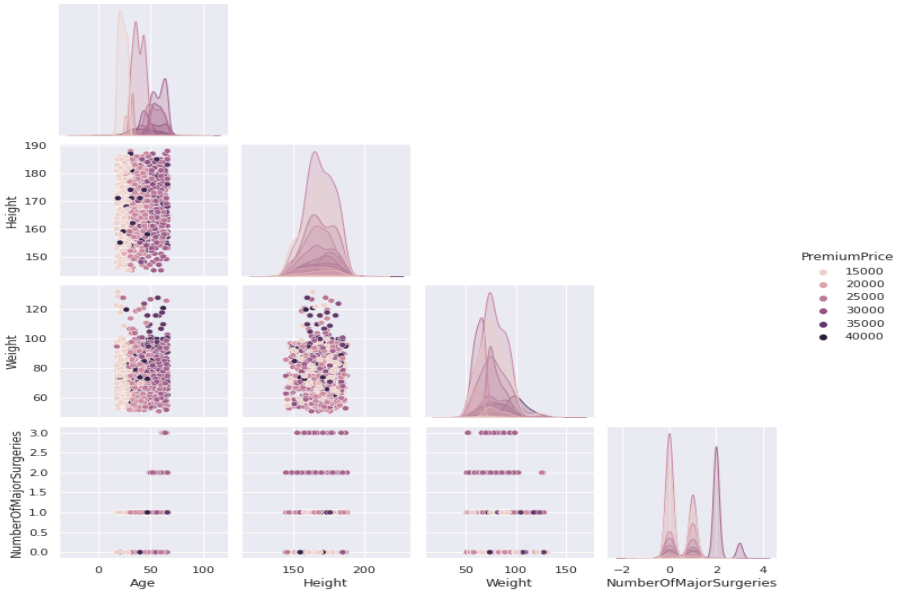
In the graph depicting the effect of any transplants on premium price also, we can see that the records with any transplant history are paying a higher amount than the others.

In the bar chart depicting the no. of surgeries and the premium amount, we can see that those who do not undergo any surgery are paying less premium than the others. We cannot see a huge difference in the premium price for the records for 2 and 3 no. of surgeries. But they are paying higher than those with 1 surgery.

By analysing the effect of diabetes and the premium price, we can see that those having diabetes are paying a little more than the average premium of those without diabetes.

From the final bar chart we can see that the one with some chronic disease are paying a higher premium than those without having any chronic disease.

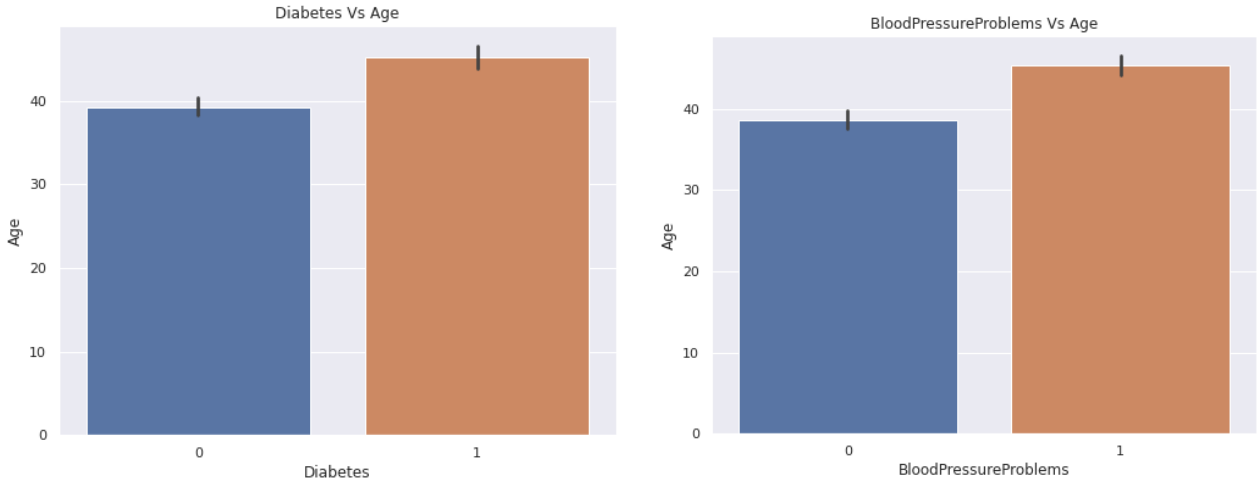
Pairplot representing all the numerical values based on target feature



Now, from the correlation matrix we have seen that some features are having a good correlation with the age, let us analyse those features as well.

Analysing the relation between age and those features having high correlation with age:

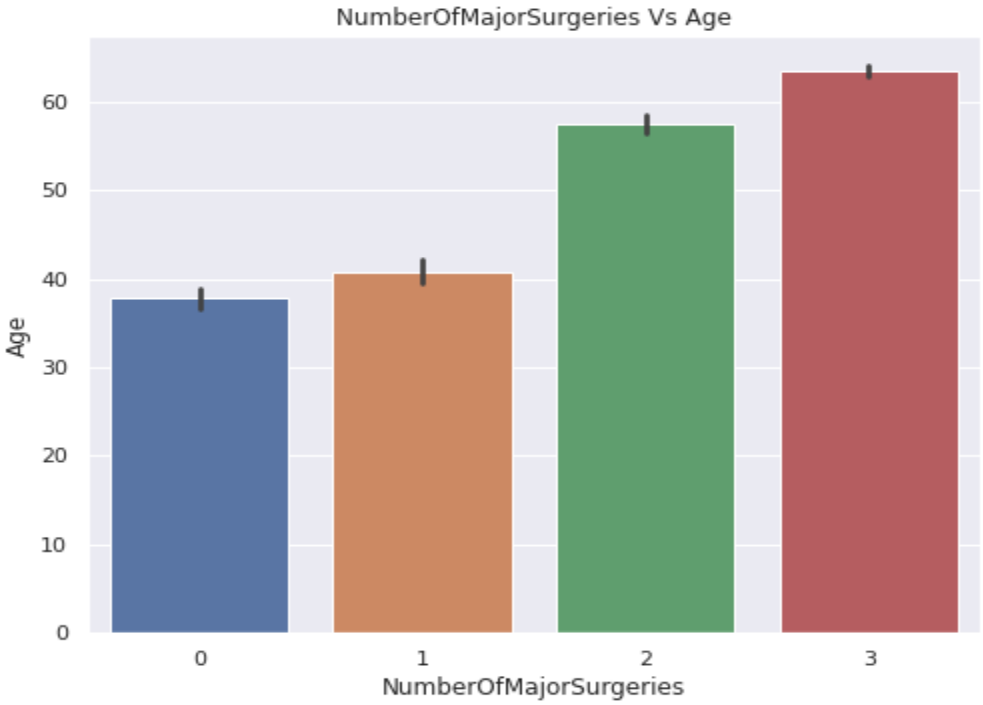
The features like Diabetes, blood pressure problem and no. of surgery are having good correlation with age. Here is the analysis regarding this.



From the above two bar charts we can see that the chances of diabetes is high for higher age groups. And also we can see that the average age of the person with diabetes is greater than 40 and the average age of the person without diabetes is less than 40.

In case of blood pressure also, we can see that the average age of the person having blood pressure is higher than that of the average age of the individuals without blood pressure.

Now, considering the age and the no. of surgeries. From the bar graph depicted below indicate that the no. of surgeries is high for the individuals with high age. That may be due to the health issues we face when we become old.



Here we can see that the average age of the persons undergoing 3 surgeries is greater than 60. And the average age of the person without any surgery is less than 40.

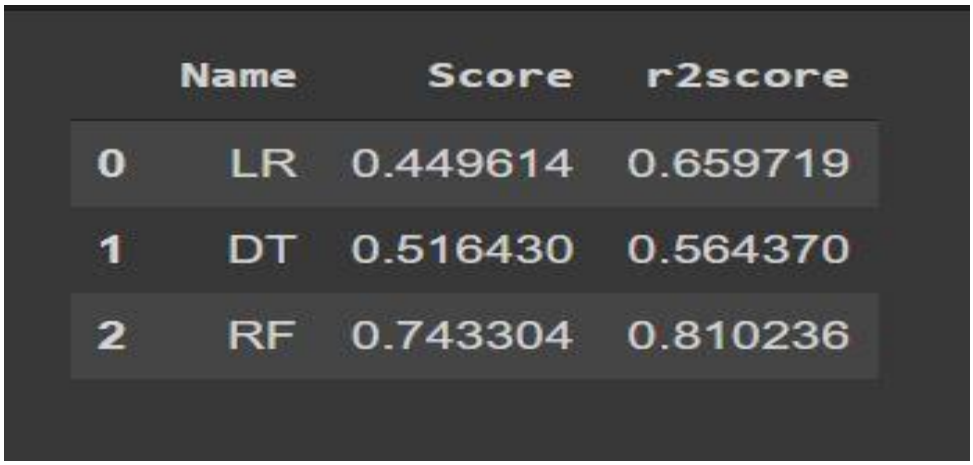
Splitting the dataset to feature set and the target:

In this step I separate the target feature that is the premium price from the dataset and make it separate as y. And all the other features together as x.

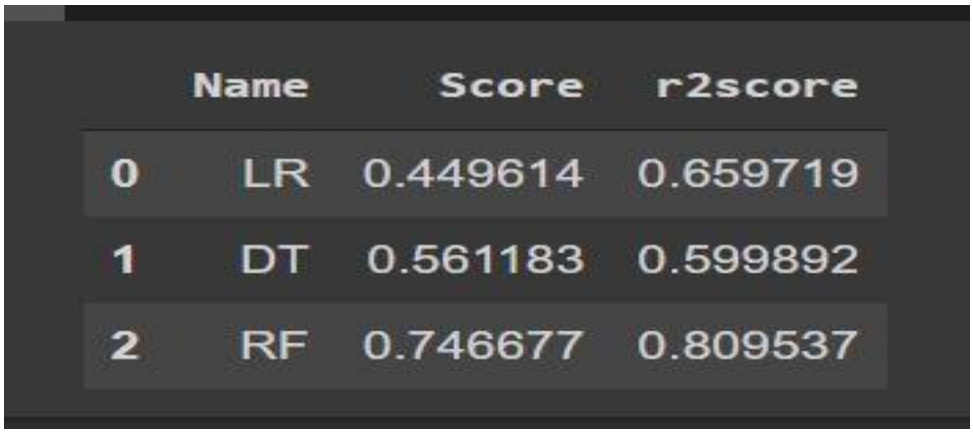
Feature Scaling:

In the dataset we have some numerical features such as age, height and weight. So for improving the performance of the model we need to scale these numerical values. Otherwise some variables may get higher importance in the model. Since our target feature is not following normal distribution, I am applying normalisation for feature scaling.

(note: I have also applied standardisation technique for scaling but the normalisation technique is giving higher prediction score so I opted normalisation)



	Name	Score	r2score
0	LR	0.449614	0.659719
1	DT	0.516430	0.564370
2	RF	0.743304	0.810236



	Name	Score	r2score
0	LR	0.449614	0.659719
1	DT	0.561183	0.599892
2	RF	0.746677	0.809537

The First figure represents the r2 score and the explained variance score for different models after applying normalisation and the second figure represents the r2 score and the explained variance score after applying standardisation for scaling.

Splitting into Train and Test data:

After applying the normalisation technique I split the dataset to train and test data. I divide the train and the test dataset in a 3 : 1 ratio. That is, of the total records 75% is my training data and 25% is test data .

MODEL SELECTION:

For selecting a good model I tried modeling with three algorithms which can be used in regression type problems like our. I predicted the value using those three algorithms and checked the prediction score for all these three algorithms. I have also done the same

process with the both scaling techniques and conclude that normalisation is best for my model.

A complete process and performance of these models is given below.

Algorithms that I used:

DecisionTreeRegressor

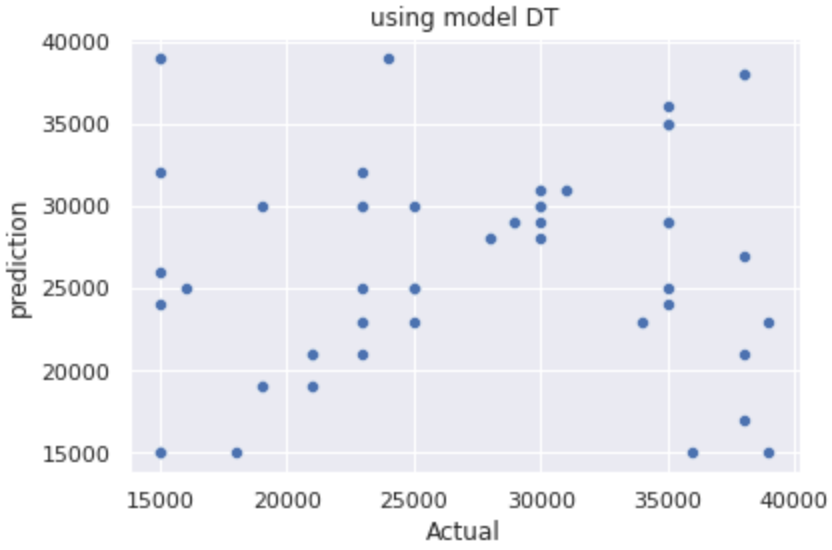
RandomforestRegressor

Linearregression

DecisionTreeRegressor

Decision tree regression models also belong to this pool of regression models. The predictive model will either classify or predict a numeric value that makes use of binary rules to determine the output or target value. The decision tree model, as the name suggests, is a tree-like model that has leaves, branches, and nodes.

By using the decisiontreeregressor I got an r2_score of .557654 and explained variance score as .510287 This values are very less so I reject this model.

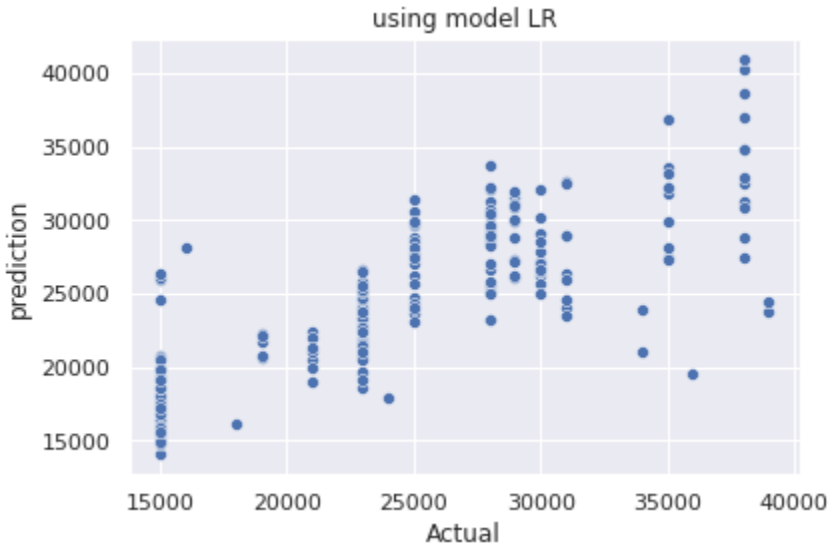


This is the graph plotting the actual and predicted values using the decisiontreeregressor. The values are scattered very much so we cannot take this model.

Linear Regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

When I used this technique I got the r2_score as .659719 and the explained variance score as .449614 .This is also very less so we rejected this model also.

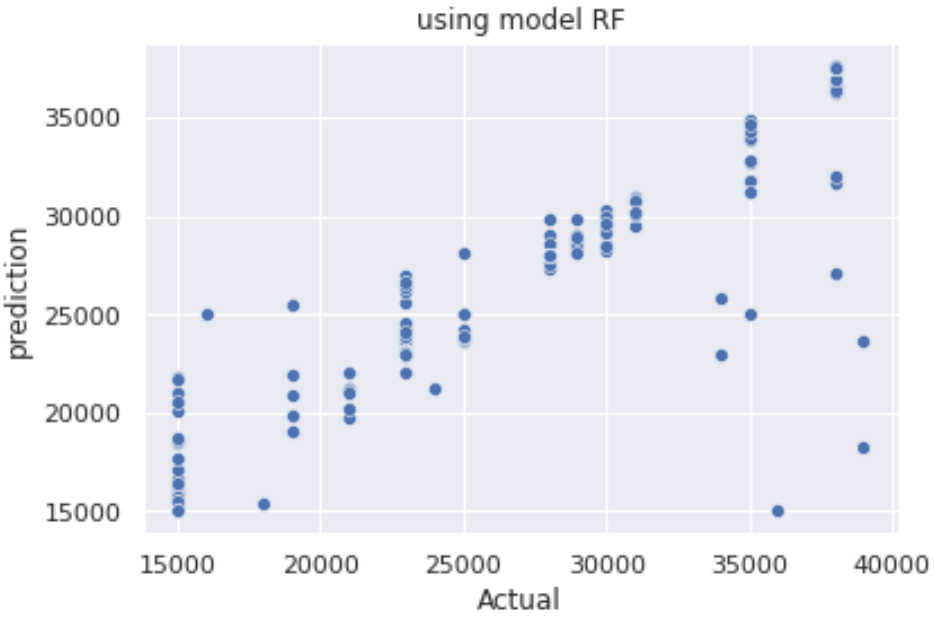


This is the scatter diagram plotting the predicted and the actual value using linear regression. This pattern is better than the decision tree but we are going for the next model.

RandomforestRegressor:

A random forest regressor. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

When I used the RandomForestRegressor I got a r2_score of .812104 and the explained variance score as .748176. Which is better than all the other models so I selected this model for prediction.



This is the scatter diagram plotting the actual and predicted values using RandomForestRegressor.

Conclusion:

After all the analysis and applying different machine learning algorithms we can say that by using RandomForestRegressor with all the engineered features we can predict the premium price with .81 r2 score and .74 explained variance score.