

**CAPSTONE PROJECT FINAL REPORT ON**

**HEALTH CARE LIFE INSURANCE COST**

**BY**

**SUDEEP KUMAR DAS**

**PGPDSBA.O. SEP22.B**

**TABLE OF CONTENT**

**LIST OF FIGURES \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 03**

**LIST OF TABLES \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 06**

**Introduction \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 07**

**EDA and Business Implication \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 08**

**Data Cleaning and Pre-processing \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 34**

**Model building \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 37**

**Model validation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_50**

**Final interpretation / recommendation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 60**

**LIST OF FIGURES**

Figure 1: First few rows of the dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_08

Figure 2: Health and Habit related parameters \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_09

Figure 3: Descriptive details of the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_09

Figure 4: Info of the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11

Figure 5: Missing Value Details \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_12

Figure 6: Countplot of Occupation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13

Figure 7: Countplot of regular Checkups Last Year \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13

Figure 8: Countplot of Adventure Sports \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14

Figure 9: Countplot of Cholesterol Level \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 14

Figure 10: Countplot of Heart Disease History \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 15

Figure 11: Countplot of Gender \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_15

Figure 12: Countplot of Smoking Status \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_16

Figure 13: Countplot of Location of Customers \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_16

Figure 14: Countplot of Insurance from any other company \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 17

Figure 15: Countplot of Alcohol consumption \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_17

Figure 16: Countplot of Exercise \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_18

Figure 17: Distribution of Years of insurance with us \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_18

Figure 18: Distribution of regular checkups last year \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_19

Figure 19: Distribution of Adventure Sports \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 19

Figure 20: Distribution of Doctor visits in the last year \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_20

Figure 21: Distribution of Daily Average Steps \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20

Figure 22: Distribution of age \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 21

Figure 23: Heart disease history \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_21

Figure 24: Other Major Disease History \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_22

Figure 25: Average Glucose Level \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_22

Figure 26: Body Mass Index \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 23

Figure 27: Year last Admitted \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 23

Figure 28: Weight \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 24

Figure 29: Fat Percentage \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_24

Figure 30: Insurance Cost \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_25

Figure 31: Insurance Cost vs Years of insurance with us \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 25

Figure 32: Insurance Cost vs Doctor Visits in last 1 Year \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 26

Figure 33: Insurance Cost vs Regular Checkup Last Year \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_26

Figure 34: Insurance Cost vs Adventure Sports \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_26

Figure 35: Insurance Cost vs Other Major Disease History \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_26

Figure 36: Insurance Cost vs Heart Disease History \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 27

Figure 37: Insurance Cost vs Gender \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_27

Figure 38: Insurance Cost vs Cholesterol Level \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_27

Figure 39: Insurance Cost vs Occupation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 27

Figure 40: Insurance Cost vs Covered by any other Company \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 28

Figure 41: Insurance Cost vs Alcohol Consumption \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_28

Figure 42: Insurance Cost vs Smoking Status \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_28

Figure 43: Insurance Cost vs Exercise \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_28

Figure 44: Age Vs Insurance Cost \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_29

Figure 45: Insurance Cost vs Daily Average Steps \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 29

Figure 46: Insurance Cost vs Average Glucose Level \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 29

Figure 47: Insurance Cost vs Weight Change in last one Year \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 30

Figure 48: Insurance Cost vs Fat Percentage \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 30

Figure 49: Insurance Cost vs Weight \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_30

Figure 50: Insurance Cost vs Year Last Admitted in Hospital \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 31

Figure 51: Insurance Cost vs BMI \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 31

Figure 52: Correlation Plot Multivariate Analysis \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_32

Figure 53: Missing Values in the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 34

Figure 54: Dataset after Missing Value Treatment \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_35

Figure 55: Outliers in the Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_35

Figure 56: Dataset after outlier treatment \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_36

Figure 57: Equation of the OLS model 33 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 44

Figure 58: Decision Tree Output \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 45

Figure 59: Predicted vs Actual values on Train Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 46

Figure 60: Predicted vs Actual values on Test data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 46

Figure 61: Model Fitting & output Random Forest \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 47

Figure 62: Predicted vs Actual Values on Train Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 47

Figure 63: Predicted vs Actual on Test Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 48

Figure 64: XGBoost model fitting in data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 49

Figure 65: Predicted vs Actual on Train data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 49

Figure 66: Actual vs Predicted on Test Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 50

Figure 67: Predicted vs Actual on Train Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 52

Figure 68: Predicted vs Actual on Test Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 52

Figure 69: Predicted vs Actual on Train Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 53

Figure 70: Predicted vs Actual on Test Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 54

Figure 71: Predicted vs Actual on Train Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 55

Figure 72: Predicted vs Actual on Test Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 55

Figure 73: Predicted vs Actual on Train Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_57

Figure 74: Predicted vs Actual on Test Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_57

Figure 75: XGBoost model Fitting \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_58

Figure 76: Predicted vs Actual on Train Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_59

Figure 77: Predicted vs Actual on Test Data \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_59

**LIST OF TABLES**

Table 1: Data Dictionary \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_10

Table 2: Dataset after Encoding (Transformation) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_36

Table 3: Coefficients of the independent variables \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_39

Table 4: OLS Summary \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_40

Table 5: VIF of the Independent variables \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_42

Table 6: OLS model summary of model 33 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 43

Table 7: OLS Model 33 parameters \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 44

Table 8: Coefficients of the Regularised RIDGE model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_51

Table 9: Coefficients of the LASSO Model \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 53

Table 10: Evaluation metrics of all the models with tuned models \_\_\_\_\_\_\_\_\_\_\_\_\_60

1. **Introduction:**

* Brief introduction about the problem statement and the need of solving it.

**Solution:**

**Business Problem:**

We all know that Health care is very important domain in the market. It is directly linked with the life of the individual; hence we have to be always be proactive in this particular domain. Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance, then it will become a pretty tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced. As a result, over the past two years, the health insurance industry in India has undergone a significant shift. Customers' attitudes have changed considerably, as has their need for health insurance.

**Scope:**

The India health insurance market size was valued at USD 12.86 billion in 2022 and is expected to expand at a compound annual growth rate (CAGR) of 11.55% from 2023 to 2030. Health insurance coverage is rapidly increasing in India due to the rising costs of high-quality healthcare coupled with rising income levels, longer life expectancies, and an epidemiological change towards no communicable diseases. The outbreak of COVID-19 has made individuals more aware of life’s unpredictability and uncertainties and their lack of preparations in the occurrence of a medical emergency. As a result, over the past two years, the health insurance industry in India has undergone a significant shift. Customers' attitudes have changed considerably, as has their need for health insurance. total health insurance premium collected in India increased by a whopping around 25% in a year.

**Objective:**

The objective of this exercise is to build a model, using data that provide the optimum insurance cost for an individual. You have to use the health and habit related parameters for the estimated cost of insurance.

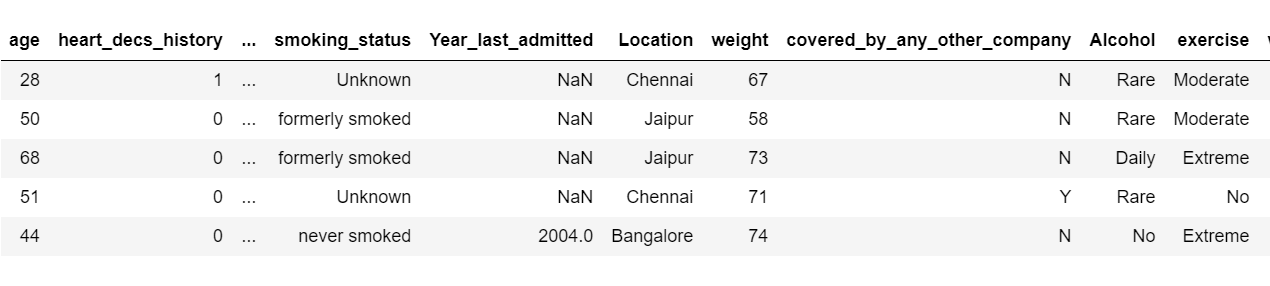
For this we will use the health and habit related parameters provided in the dataset to determine which of the parameters play major role in determining the cost of the insurance. Then we will use the machine learning algorithms and compare the performance results and finally predict the cost of health insurance accurately based on the best performing algorithms.

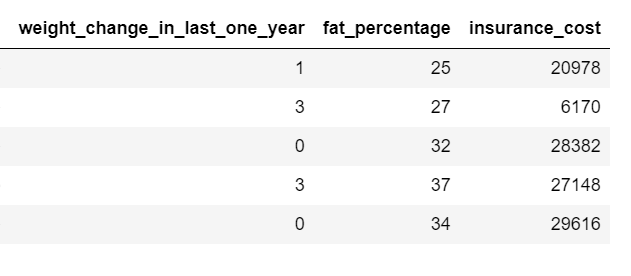
1. **EDA and Business Implication**

* Uni-variate / Bi-variate / multi-variate analysis to understand relationship b/w variables. How your analysis is impacting the business?
* Both visual and non-visual understanding of the data.

**Solution:**

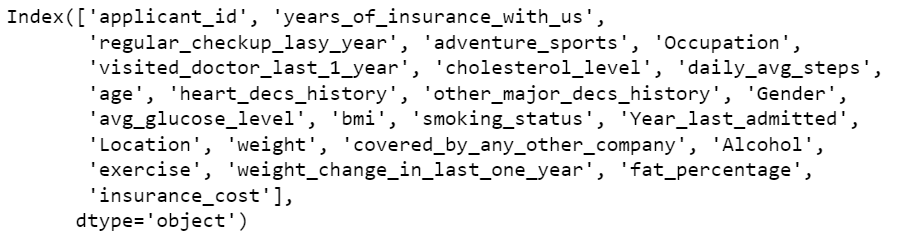
The below figure shows the first few rows of the dataset.****

****

****

***Figure 1: First few rows of the dataset***

Below are the health and habit related parameters in the dataset.

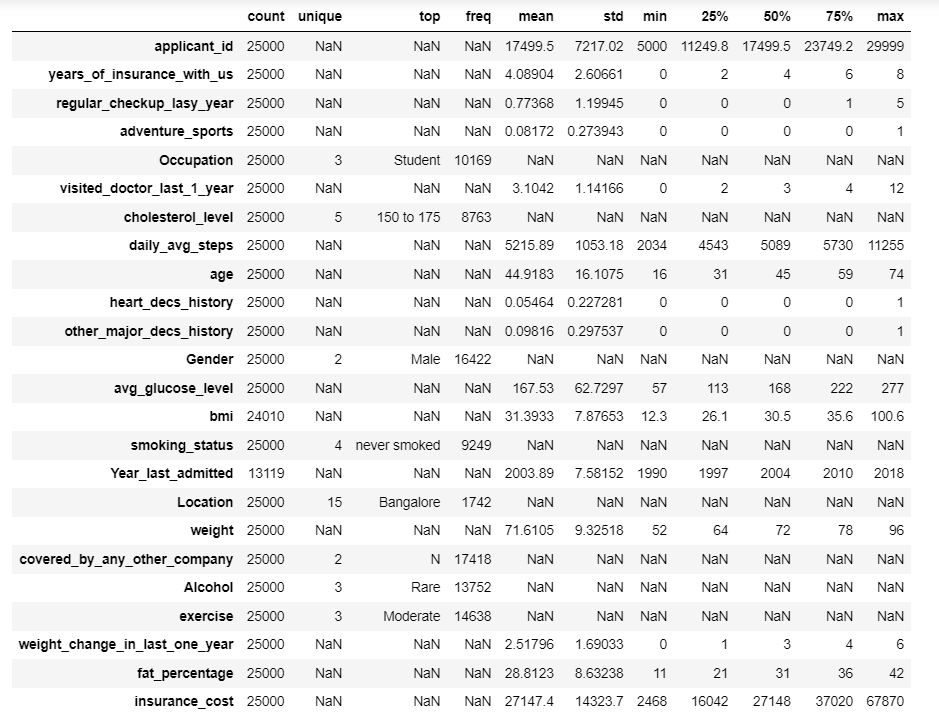


***Figure 2: Health and Habit related parameters***

The shape of the dataset is as follows:



The Descriptive details of the dataset are as follows:



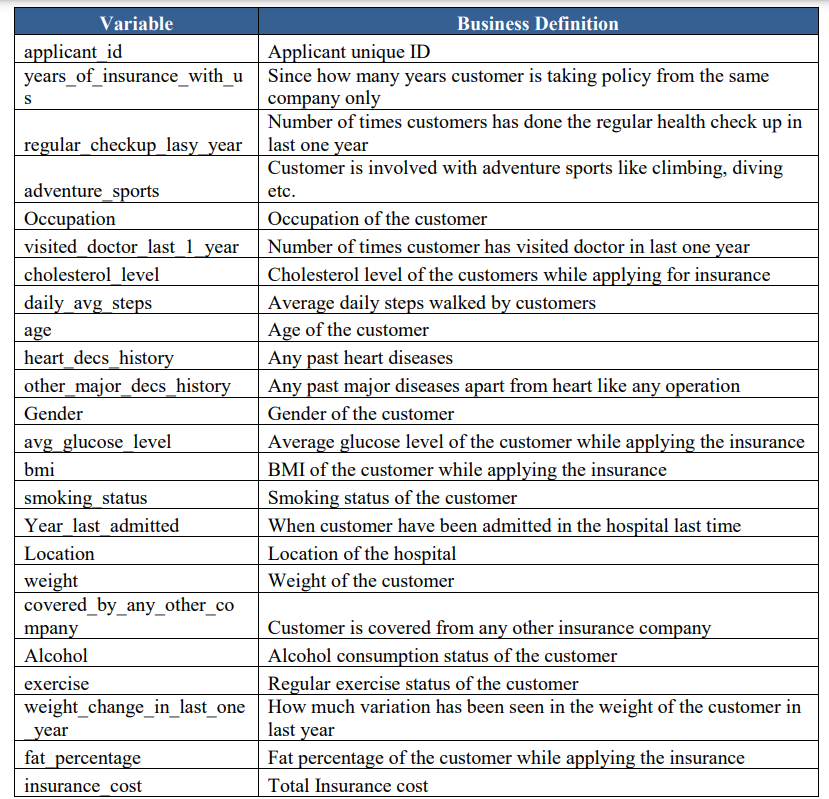
***Figure 3: Descriptive details of the Dataset***

Basic Inferences:

* The dataset contains of 25000 rows and 24 columns, out of which 22 columns contains health and habit related parameters.
* There appears to be some missing values which we will explore further.
* The applicants are insured with the company for an average period of 4 years.
* Around 40% of the applicants are students.
* The mean age of the applicants is approx. 45 years with the minimum age being 16 years.
* Approx. 65.7% of the applicants are male.
* The applicants are from 15 locations with the highest being from Bangalore.
* Most of the applicants do Moderate exercise.
* The average Insurance cost is 27147 Rs. With lowest insurance cost being 2468 Rs. And highest being 67870 Rs.

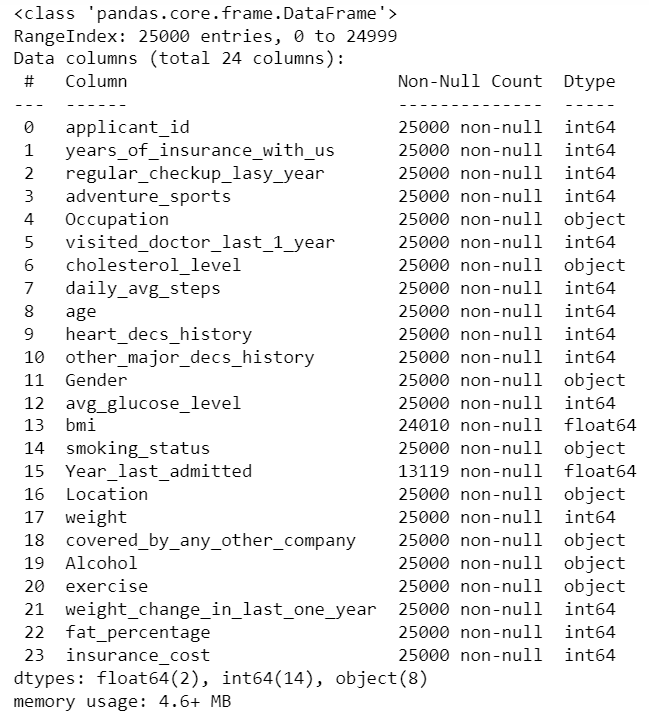
**Understanding of attributes:**

The dataset contains of 25000 rows and 24 columns, out of which 22 columns contains health and habit related parameters. The data dictionary and the information of the variables are as follows:

****

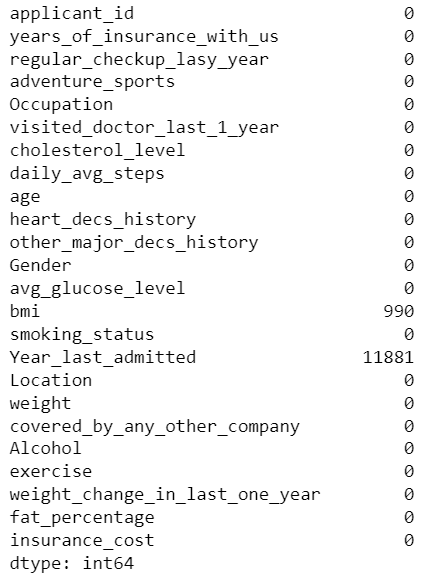
***Table 1: Data Dictionary***

The info of the variables are as follows:



***Figure 4: Info of the Dataset***

The Missing Value details of the dataset are as follows:



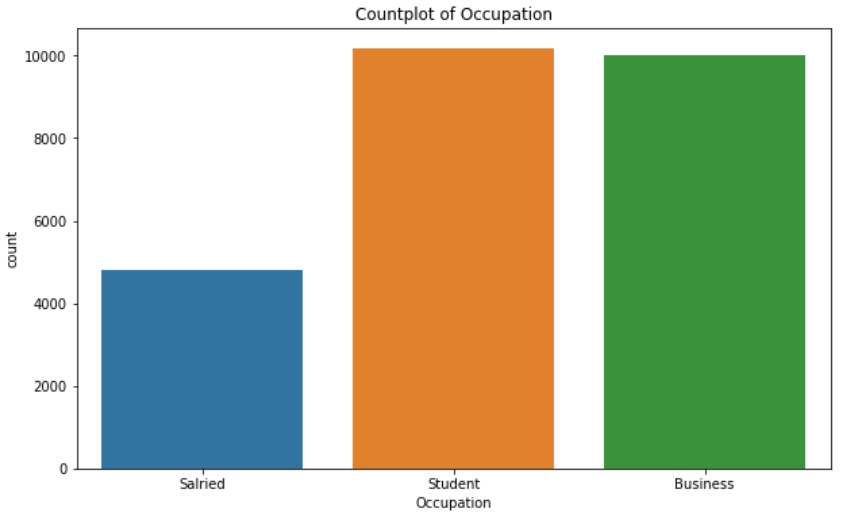
***Figure 5: Missing Value Details***

**Basic Inferences:**

* There are 24 variables in the dataset. Out of which insurance\_cost is the dependent variables and the column applicant\_id contains the information about the applicants.
* The rest of the 22 variables are the health and habit related information of the applicants.
* There are some missing values in the dataset. The variable BMI has 990 missing variables and year\_last\_admitted contains 11881 missing values.
* There are 16 continuous variables and 8 categorical variables. Out of these 17 variables are health related parameters and the rest are habit related parameters.
* Renaming is done for a variable. The variable “regular\_checkup\_lasy\_year” is renamed to “regular\_checkup\_last\_year”.
* **Exploratory data analysis**

1. **Univariate Analysis.**

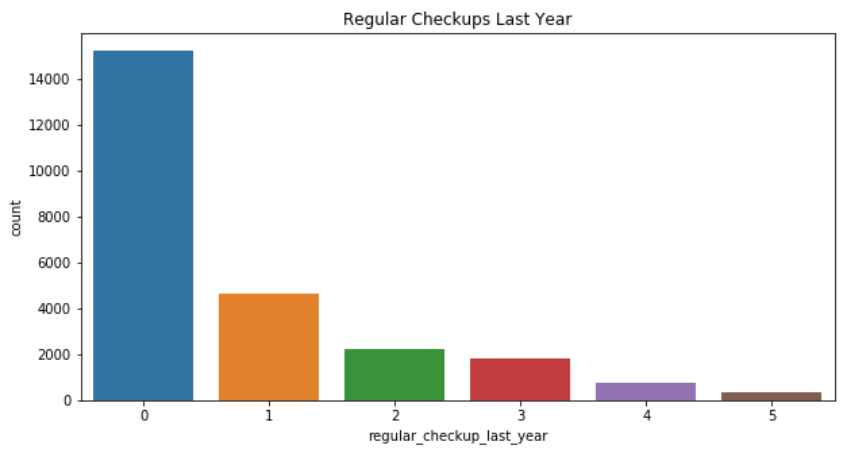
* **Occupation:**

****

***Figure 6: Countplot of Occupation***

From the above countplot we can observe that major customers are Students followed by Business and Salaried which is the lowest.

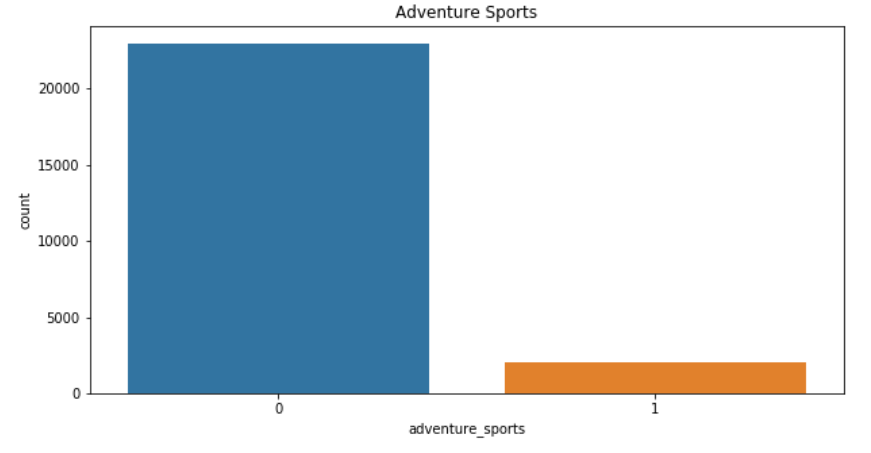
* **Regular Checkups last Year**



***Figure 7: Countplot of regular Checkups Last Year***

From the above plot we observe that most of the customers have not done any regular checkups last year.

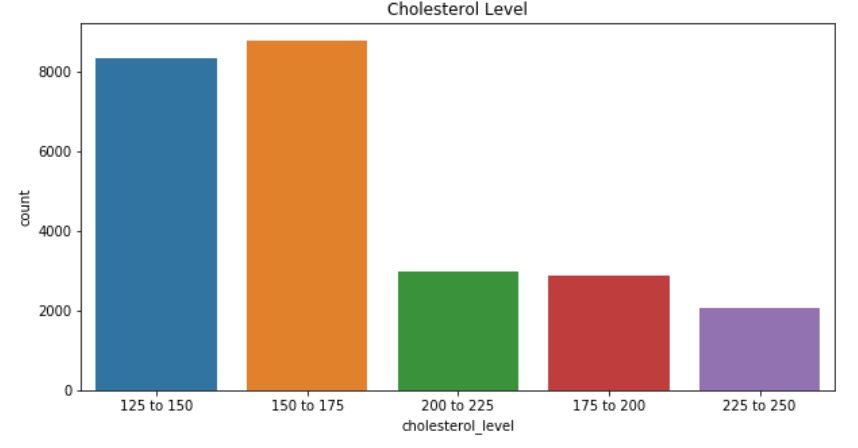
* **Adventure Sports:**



***Figure 8: Countplot of Adventure Sports***

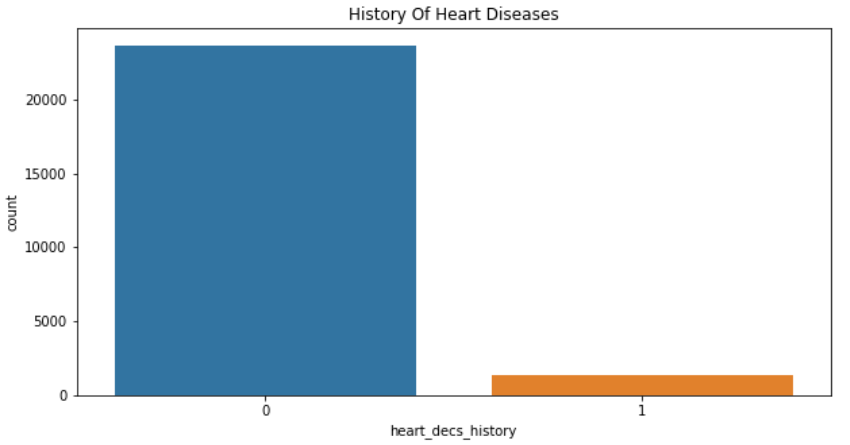
Most of the customers do not participate in any form of adventure sports.

* **Cholesterol Level:**

****

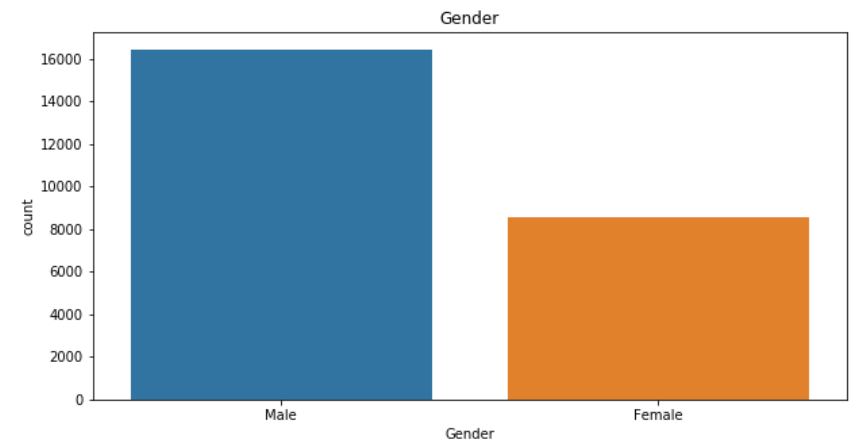
***Figure 9: Countplot of Cholesterol Level***

* **Heart Disease History:**

****

***Figure 10: Countplot of Heart Disease History***

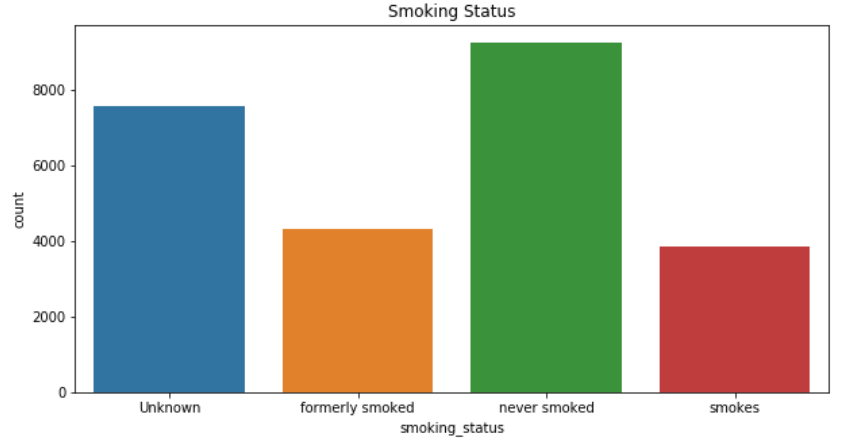
* **Gender:**

****

***Figure 11: Countplot of Gender***

We observe that maximum number of customers are male.

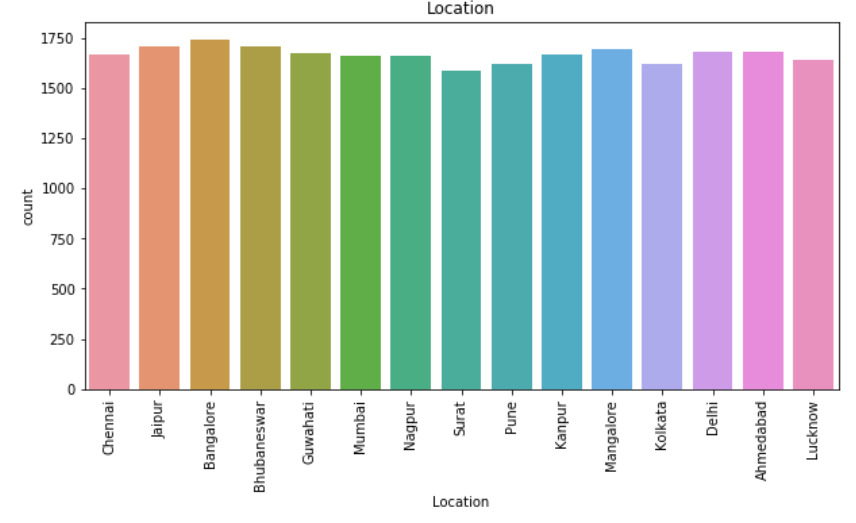
* **Smoking Status:**

****

***Figure 12: Countplot of Smoking Status***

We observe that most of the customers have never smoked followed by unknown. Formerly smoked and smoking customers are approximately similar.

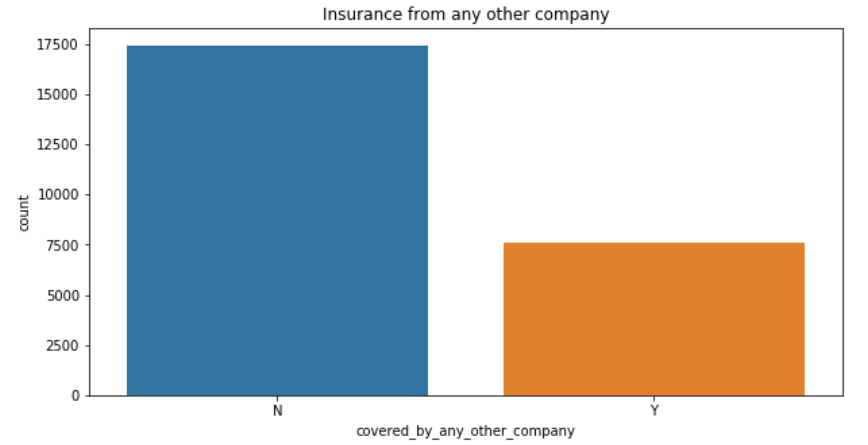
* **Location:**

****

***Figure 13: Countplot of Location of Customers***

We observe that the distribution is similar across.

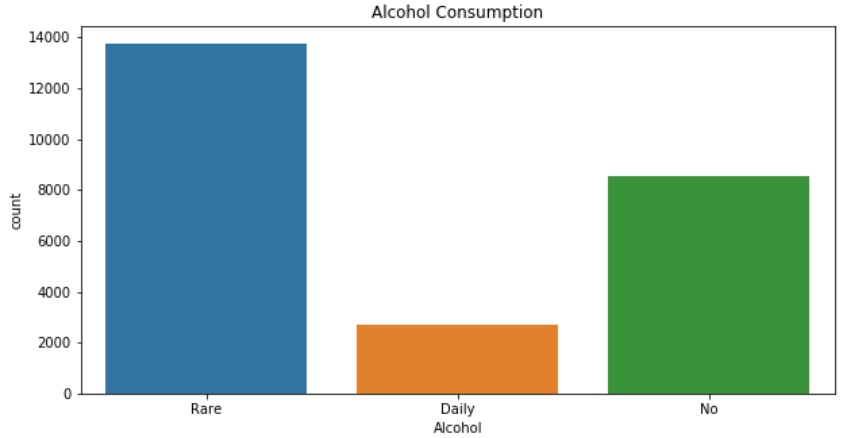
* **Covered By any other Insurance Company:**



***Figure 14: Countplot of Insurance from any other company***

We Observe that most of the customers do not have insurance coverage from any other company.

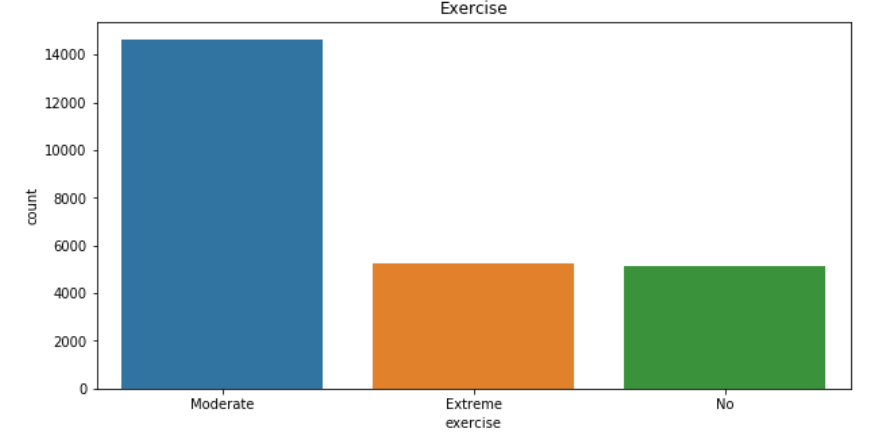
* **Alcohol Consumption:**

****

***Figure 15: Countplot of Alcohol consumption***

From The plot we can observe that most of the customers either rarely or don’t consume alcohol at all. Very few customers consume alcohol on a daily basis.

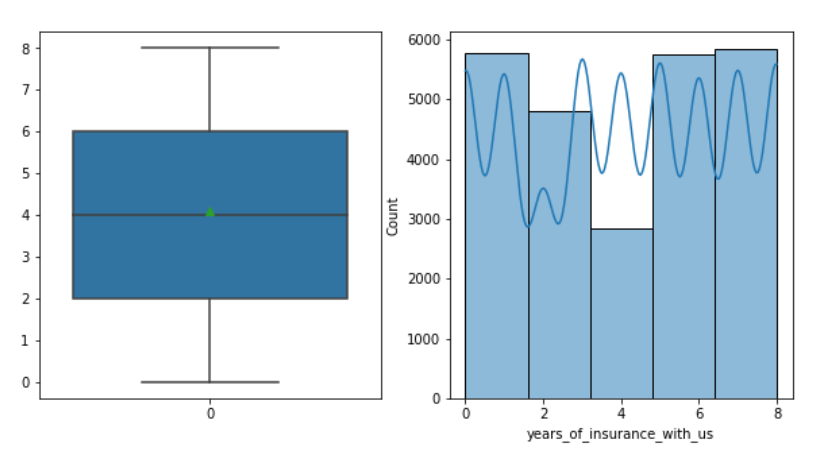
* **Exercise:**



***Figure 16: Countplot of Exercise***

We observe that most of the customers do moderate exercise while the number of customers doing extreme and no exercise at all are similar.

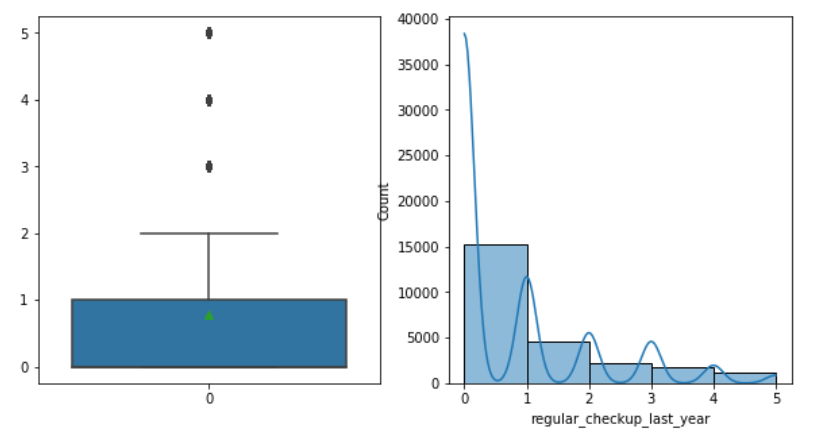
* **Years of insurance with us:**

****

***Figure 17: Distribution of Years of insurance with us***

We observe that the mean tenure of the customers’ insurance with us is around 4 years. The data is uniformly distributed. There are no outliers and the median tenure is approx. 4 years with highest tenure being 8 years.

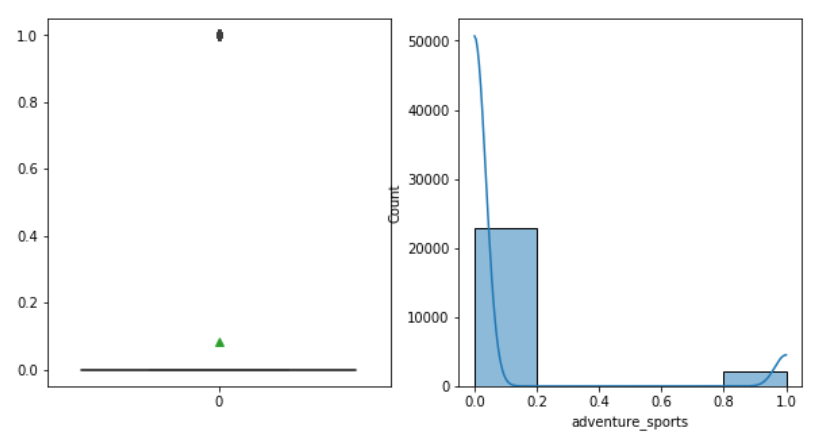
* **Regular Checkups last year:**

****

***Figure 18: Distribution of regular checkups last year***

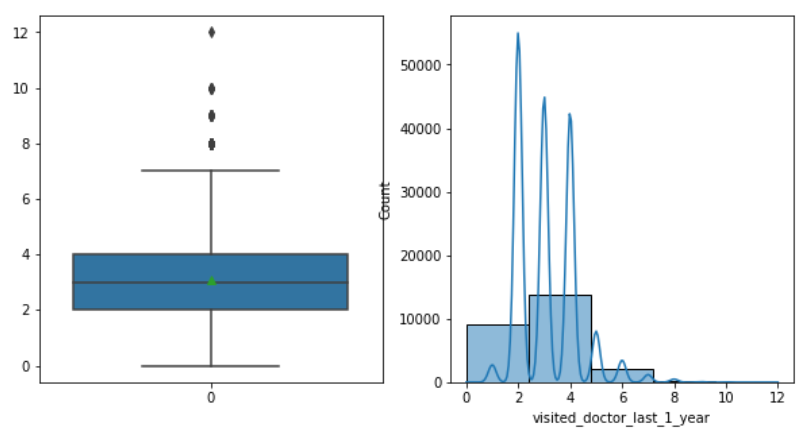
We observe that most of the customers have not done regular checkups in the last year. Average number of checkups is approx. 1. Also there are outliers in the data.

* **Adventure Sports:**



***Figure 19: Distribution of Adventure Sports***

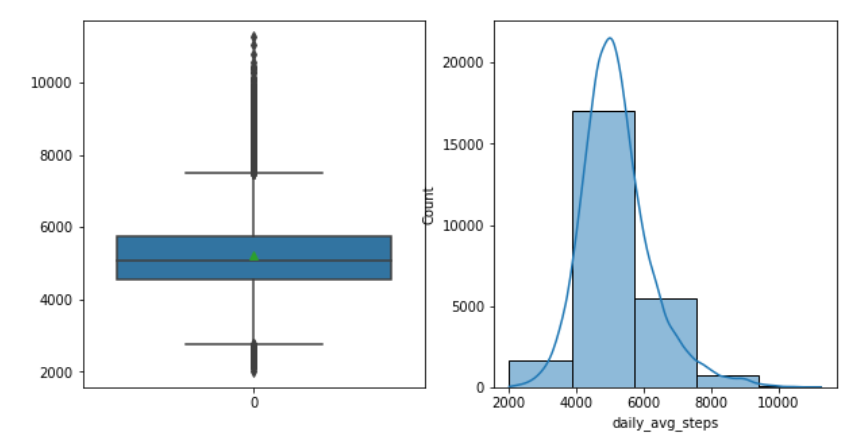
* **Doctor Visits in last 1 year:**



***Figure 20: Distribution of Doctor visits in the last year***

We observe that the customers visit doctors on an average 3 times a year. There are some outliers indicating that there are some cases where the customer had to visit 8-12 times.

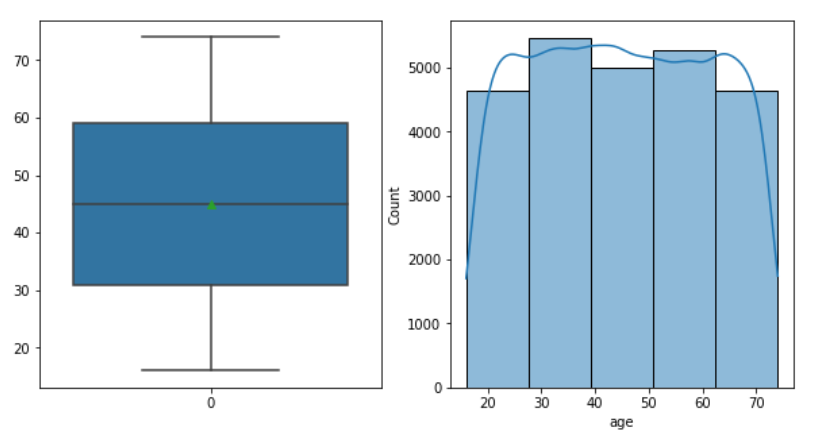
* **Daily Average Steps:**

****

***Figure 21: Distribution of Daily Average Steps***

We observe that on an average the customers take 5000 steps daily. There are outliers in the data indicating that some customers take more than 8000 steps.

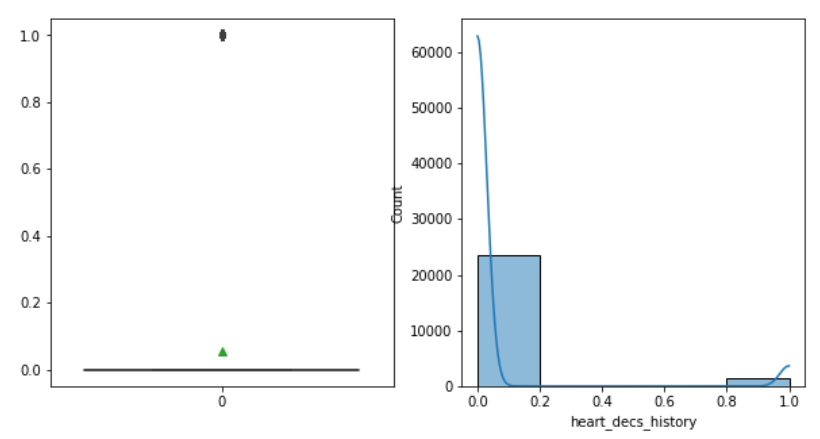
* **Age:**



***Figure 22: Distribution of age***

The average age of the customers is around 45 years. The lowest age is 16 years where as the highest age is 76 years.

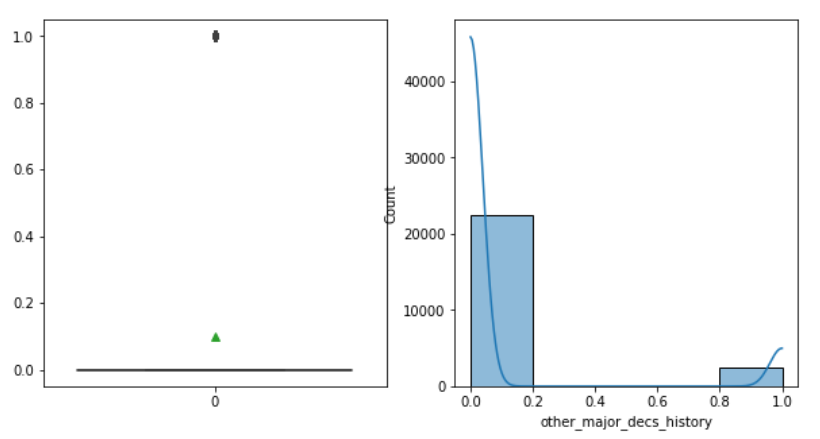
* **Heart Disease History:**



***Figure 23: Heart Disease history***

We observe that most of the consumers do not have a history of heart disease. Also there are outliers present in the data.

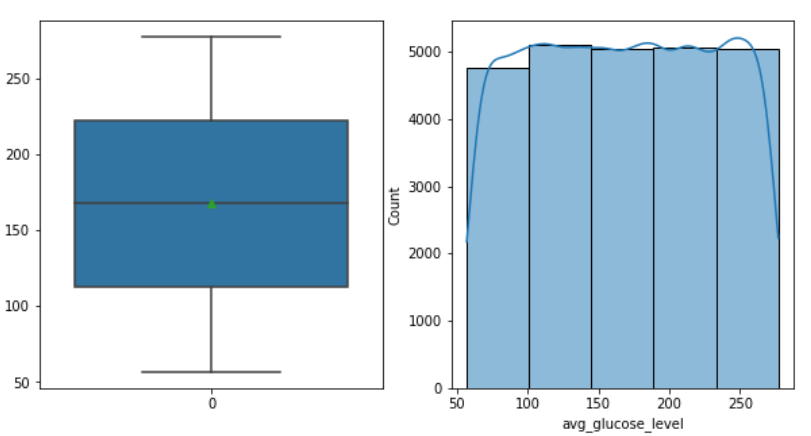
* **Other Major Disease History:**



***Figure 24: Other Major Disease History***

We observe that most of the customers do not have a history of any other major disease. Very few customers have other major disease. Also there are outliers present in the data.

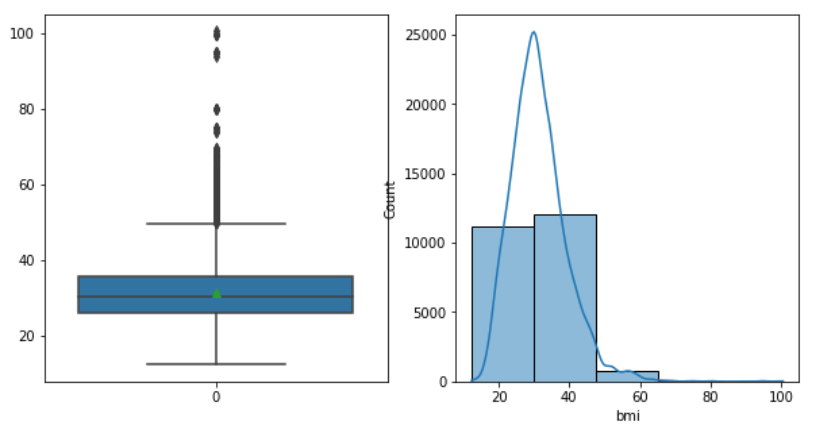
* **Average Glucose Level:**



***Figure 25: Average Glucose Level***

We observe that average glucose level of the customers is approx. 175. There are no outliers present in the dataset.

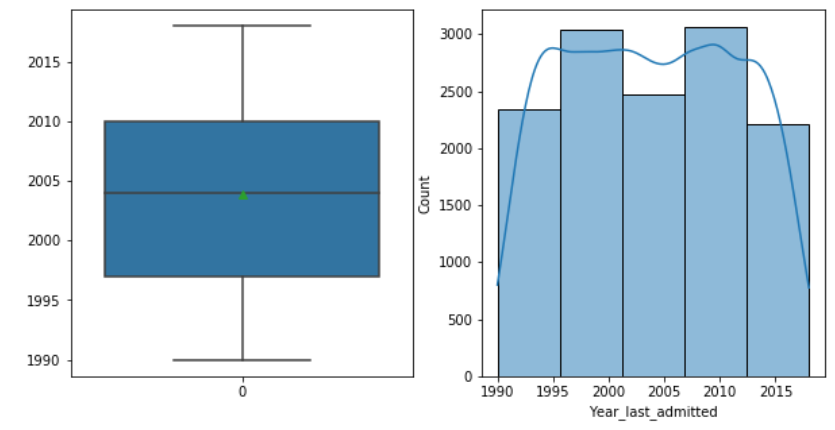
* **BMI:**



***Figure 26: Body Mass Index***

We observe that the average BMI is approx. 30. There are outliers present in the data which indicates that there are some customers who are overweight.

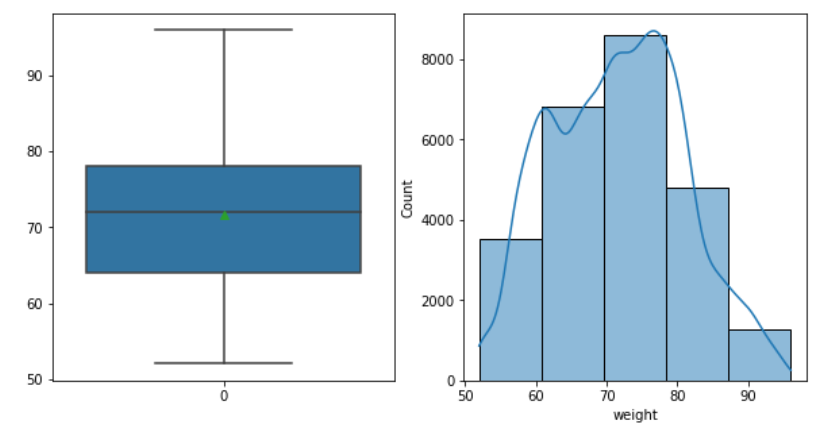
* **Year Last Admitted:**



***Figure 27: Year last Admitted***

The data is uniformly distributed. There are no outliers in the data.

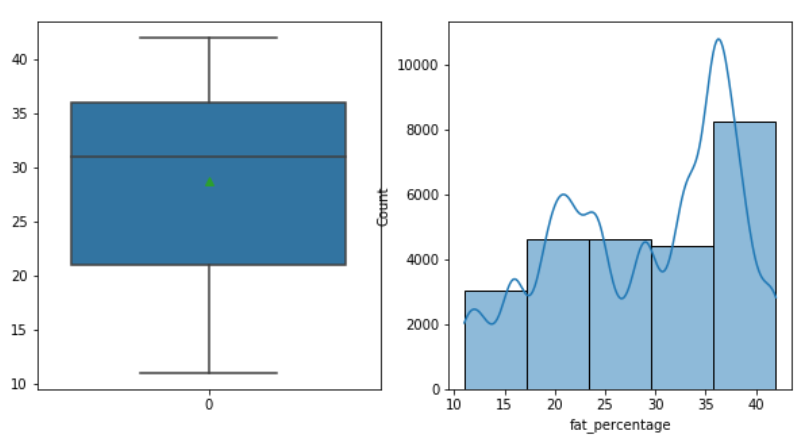
* **Weight:**



***Figure 28: Weight***

We observe that the data is uniformly distributed with average weight being approx. 72. There are no outliers in the dataset.

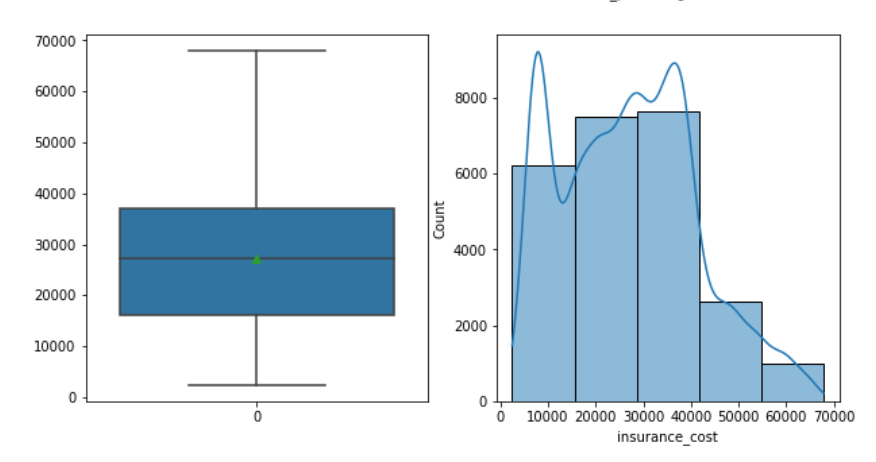
* **Fat Percentage:**



***Figure 29: Fat Percentage***

The data is right skewed and there are no outliers in the data. The average fat percentage is 28.

* **Insurance Cost:**

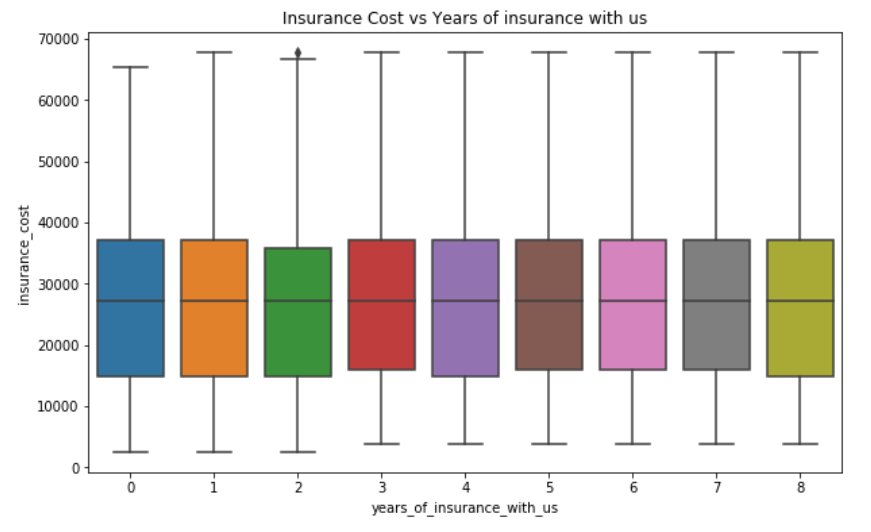


***Figure 30: Insurance Cost***

We observe that the average insurance cost is 27000 with highest being approx. 69000. There are no outliers in the data.

1. **Bivariate Analysis.**

* **Insurance Cost vs Years of insurance with us**

****

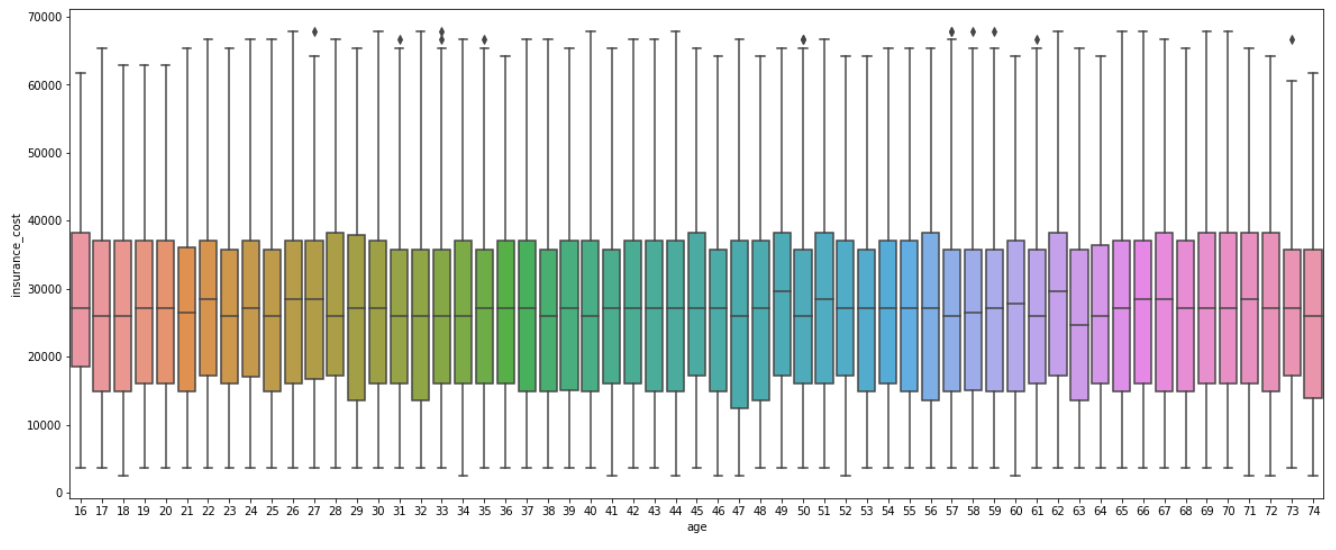
***Figure 31:* Insurance Cost vs Years of insurance with us**

|  |  |
| --- | --- |
| * **Insurance Cost vs Doctor Visits in last 1 Year**   ***Figure 32:* Insurance Cost vs Doctor Visits in last 1 Year** | * **Insurance Cost vs Regular Checkup Last Year**     ***Figure 33:* Insurance Cost vs Regular Checkup Last Year** |
| * **Insurance Cost vs Adventure Sports**     ***Figure 34:* Insurance Cost vs Adventure Sports** | * **Insurance Cost vs Other Major Disease History**     ***Figure 35:* Insurance Cost vs Other Major Disease History** |

|  |  |
| --- | --- |
| * **Insurance Cost vs Heart Disease History**     ***Figure 36:* Insurance Cost vs Heart Disease History** | * **Insurance Cost vs Gender**     ***Figure 37:* Insurance Cost vs Gender** |
| * **Insurance Cost vs Cholesterol Level**     ***Figure 38:* Insurance Cost vs Cholesterol Level** | * **Insurance Cost vs Occupation**     ***Figure 39:* Insurance Cost vs Occupation** |

|  |  |
| --- | --- |
| * **Insurance Cost vs Covered by any other Company**     ***Figure 40:* Insurance Cost vs Covered by any other Company** | * **Insurance Cost vs Alcohol Consumption**     ***Figure 41:* Insurance Cost vs Alcohol Consumption** |
| * **Insurance Cost vs Smoking Status**     ***Figure 42:* Insurance Cost vs Smoking Status** | * **Insurance Cost vs Exercise**     ***Figure 43:* Insurance Cost vs Exercise** |

* **Age vs Insurance Cost**

****

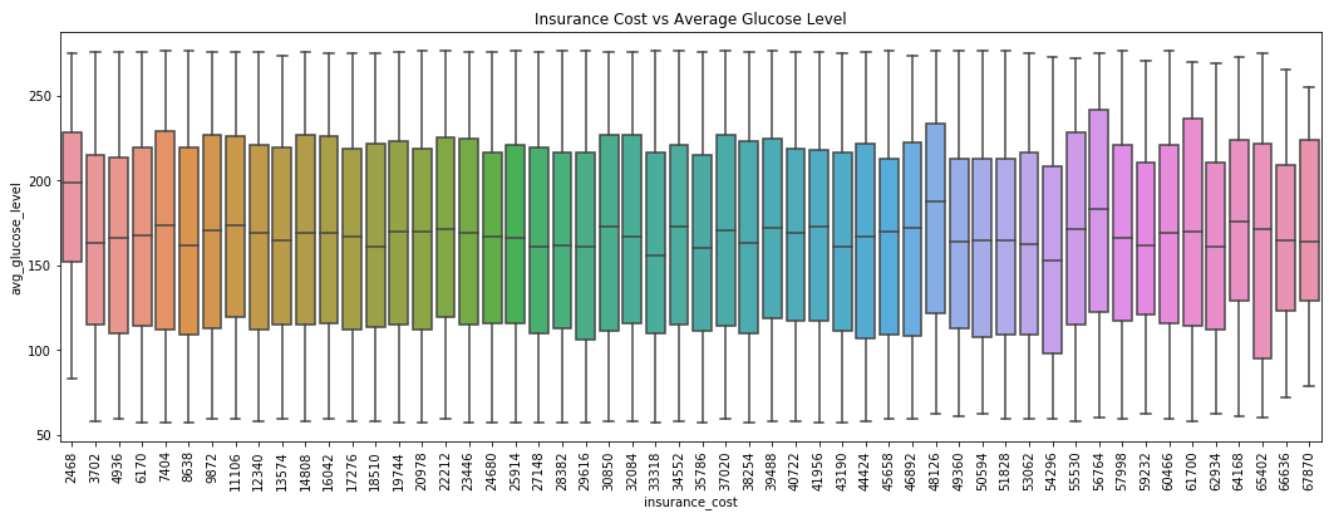
***Figure 44: Age Vs Insurance Cost***

* **Insurance Cost vs Daily Average Steps**

****

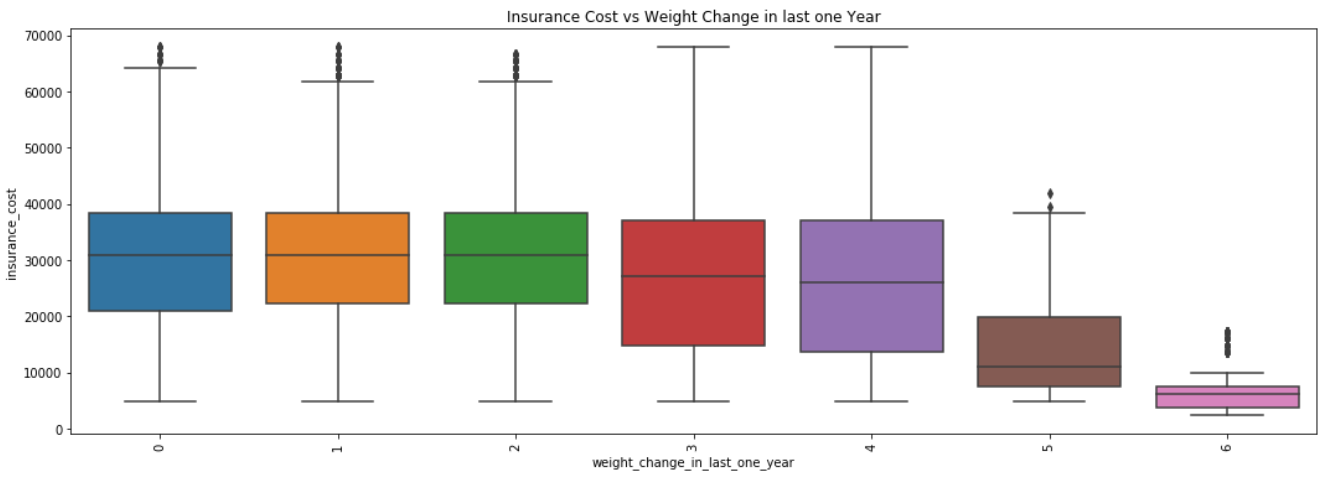
***Figure 45: Insurance Cost vs Daily Average Steps***

* **Insurance Cost vs Average Glucose Level**

****

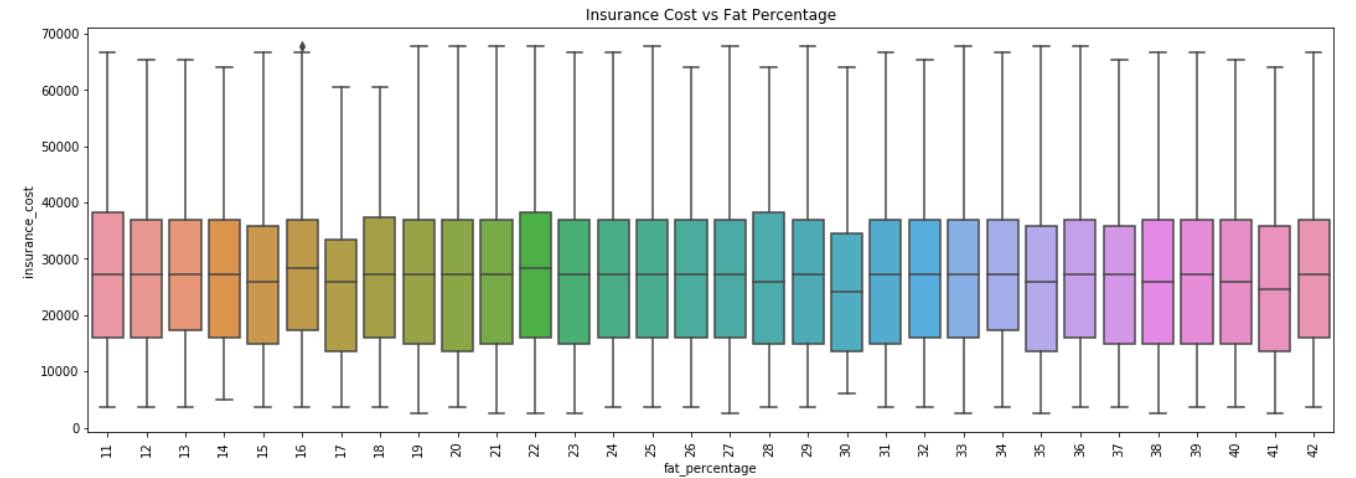
***Figure 46: Insurance Cost vs Average Glucose Level***

* **Insurance Cost vs Weight Change in last one Year**



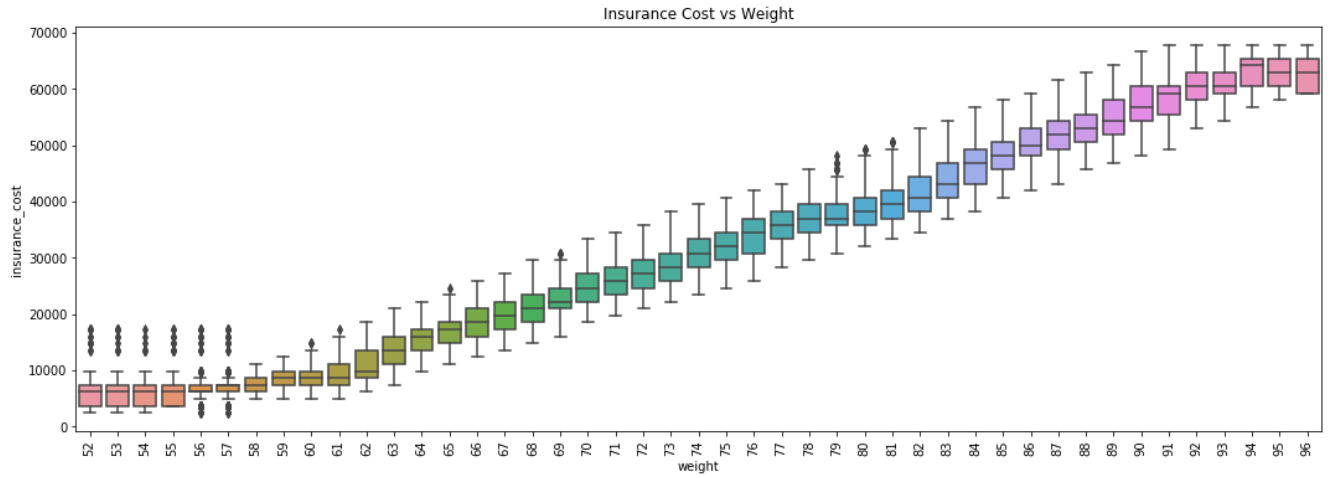
***Figure 47: Insurance Cost vs Weight Change in last one Year***

* **Insurance Cost vs Fat Percentage**



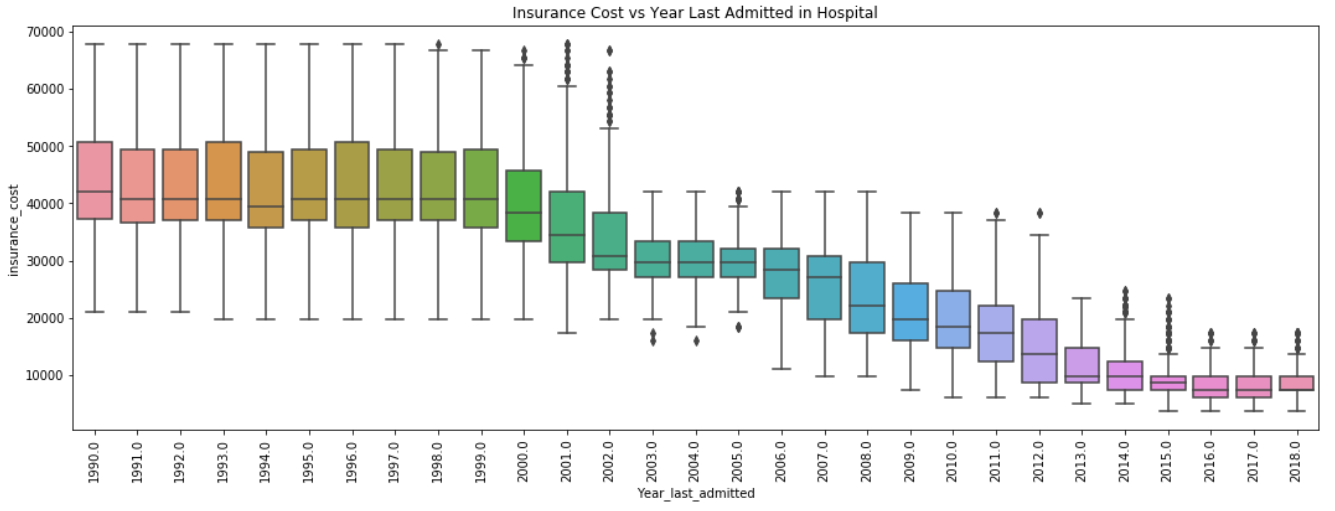
***Figure 48: Insurance Cost vs Fat Percentage***

* **Insurance Cost vs Weight**



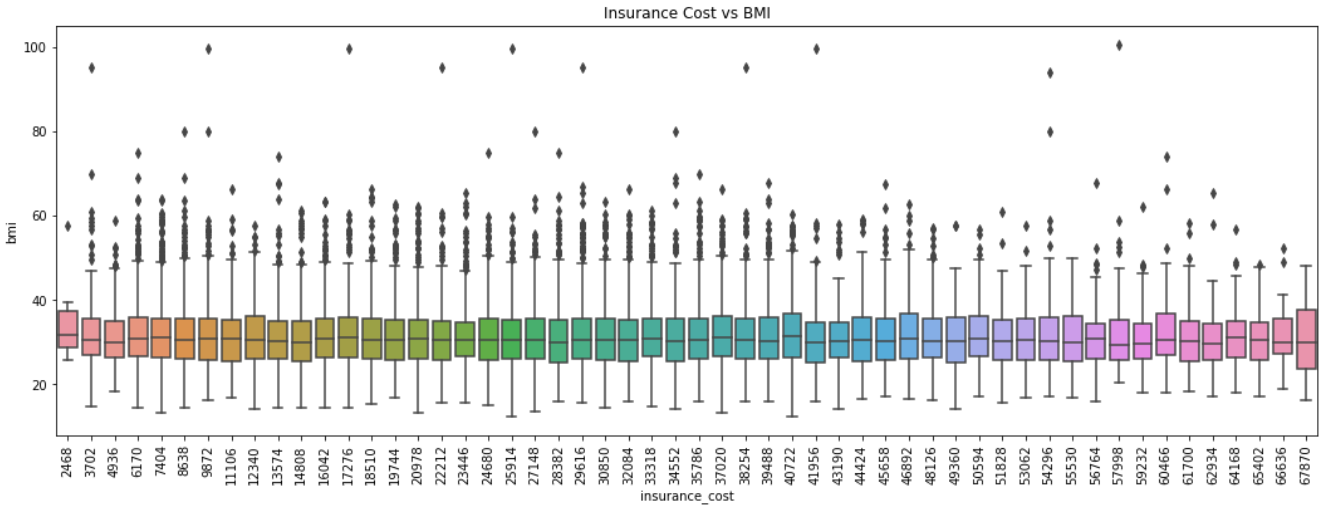
***Figure 49: Insurance Cost vs Weight***

* **Insurance Cost vs Year Last Admitted in Hospital**

****

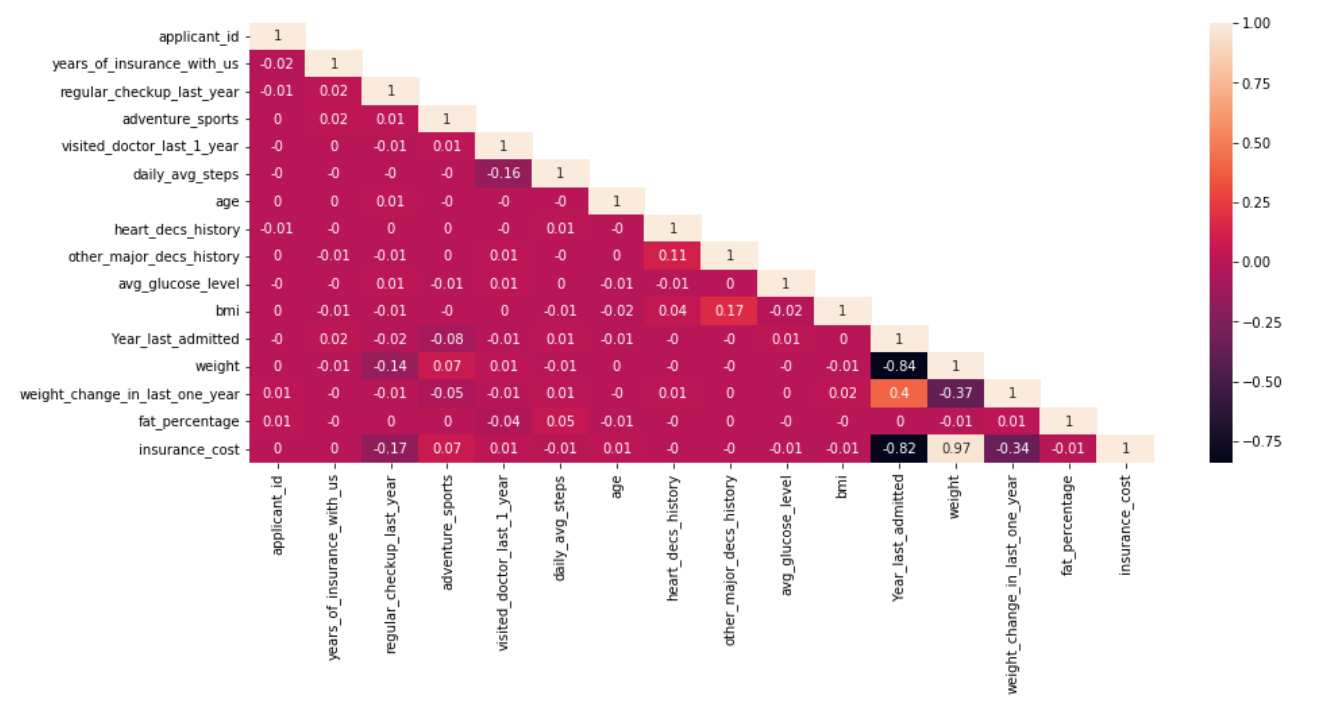
***Figure 50:* Insurance Cost vs Year Last Admitted in Hospital**

* **Insurance Cost vs BMI**

****

***Figure 51: Insurance Cost vs BMI***

**Multivariate Analysis:**

**  
 *Figure 52: Correlation Plot Multivariate Analysis***

**Inference:**

The correlation plot shows the correlation between the variables. Based on above graph, other variables don’t show much of a linear relationship with dependent variables except ‘weight’, they may have sinusoidal relationship with the dependent variable. The weight variable shows positive correlation with target variable.

**Summary Of EDA:**

* The major customers are Students followed by Business and Salaried which is the lowest.
* Most of the customers have not done any regular checkups last year.
* Most of the customers do not participate in any form of adventure sports.
* The maximum number of customers are male.
* The most of the customers have never smoked followed by unknown. Formerly smoked and smoking customers are approximately similar.
* Most of the customers do not have insurance coverage from any other company.
* Most of the customers either rarely or don’t consume alcohol at all. Very few customers consume alcohol on a daily basis.
* Most of the customers do moderate exercise while the number of customers doing extreme and no exercise at all are similar.
* The mean tenure of the customers’ insurance with us is around 4 years. The median tenure is approx. 4 years with highest tenure being 8 years.
* Most of the customers have not done regular checkups in the last year. Average number of checkups is approx. 1.
* The customers visit doctors on an average 3 times a year. There are some outliers indicating that there are some cases where the customer had to visit 8-12 times.
* On an average the customers take 5000 steps daily. There are outliers in the data indicating that some customers take more than 8000 steps.
* The average age of the customers is around 45 years. The lowest age is 16 years where as the highest age is 76 years.
* Most of the consumers do not have a history of heart disease.
* Most of the customers do not have a history of any other major disease. Very few customers have other major disease.
* The average glucose level of the customers is approx. 175.
* The average BMI is approx. 30. There are outliers present in the data which indicates that there are some customers who are overweight. Average weight being approx. 72.
* The average insurance cost is 27000 with highest being approx. 69000.
* For 'regular\_checkup\_last\_year' variable, the customers who had regular checkup 5 times within year they got less insurance cost. The customers who do not go for regular health checkup, the insurance cost is paid higher.
* For 'adventure\_sports', the customers who have participated in adventure sports, the insurance cost is paid little higher than who do not involved in adventure sports.
* For 'visited\_doctor\_last\_1\_year', the customers who have visited doctor 10 times with year, insurance paid higher.
* For variables ‘daily\_avg\_steps', 'age', 'heart\_decs\_history', 'other\_major\_decs\_history', 'avg\_glucose\_level', 'bmi', 'fat\_percentage', there is no much difference in insurance cost.
* For variable 'Year\_last\_admitted', customers who have admitted in hospital the year 1990 to 2000 the insurance paid higher when compared to who admitted in between 2015 to 2018.
* For ‘weight’, this variable is positively correlated with insurance cost. When age increases the insurance cost also increases.
* For 'weight\_change\_in\_last\_one\_year', the customers whose weight is changed 6 kg when compared to last year weight, the insurance paid less. Remaining customers who have changed weight from 0 to 5 kg, insurance paid to all same cost.

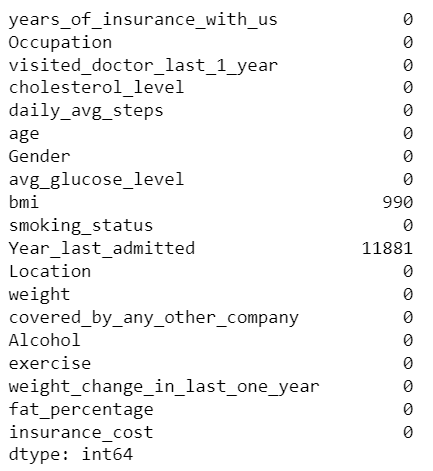
1. **Data Cleaning and Pre-processing**

* Approach used for identifying and treating missing values and outlier treatment.
* Need for variable transformation.
* Variables removed or added and why.

**Solution:**

**Missing Value Treatment:**

Following figure shows the missing values in the dataset.

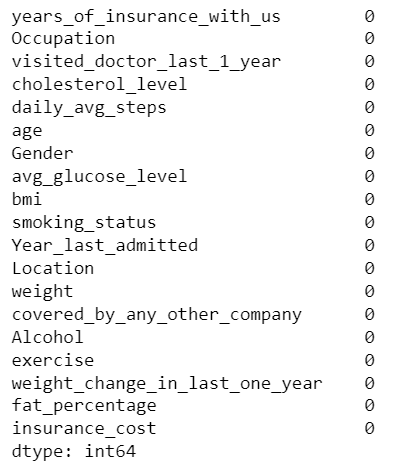


***Figure 53: Missing Values in the Dataset***

Based on above output the continuous variables **bmi** and **year\_ last \_admitted** have missing values.The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

* BMI contains 0.03% missing values and this variable has outliers. So, these missing values are imputed with median value.
* The Column “year\_ last \_admitted” has approx. 47% missing values from the total dataset. This variable has to be dropped as per the industry standards but while doing bivariate analysis we found negative correlation with target variable.

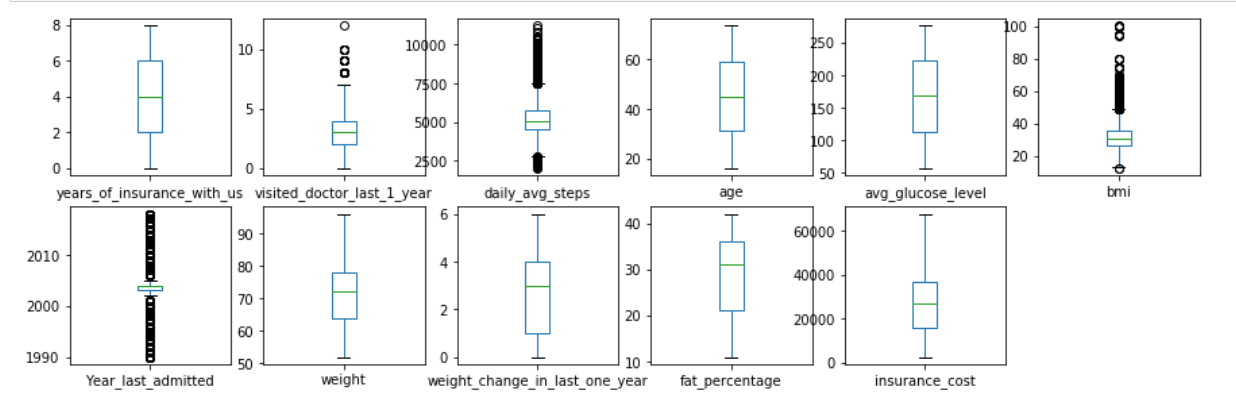
So, these missing values have been imputed with median value. The dataset after imputation of missing values is as follows.



***Figure 54: Dataset after Missing Value Treatment***

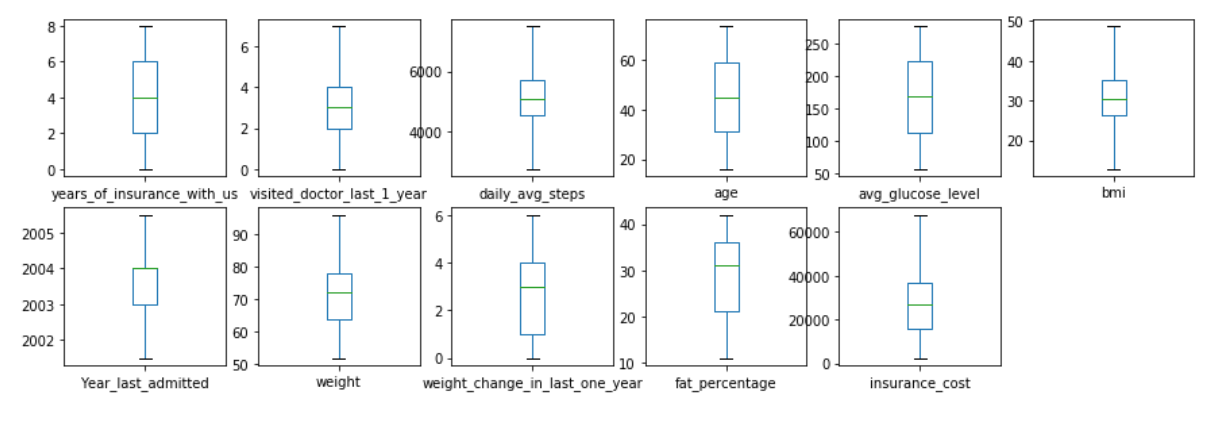
**Outlier Treatment:**

The Dataset has following Outliers:



***Figure 55: Outliers in the Dataset***

The columns “visited\_doctor\_last\_1\_year”, “daily\_avg\_steps”, “bmi”, “year\_last\_admitted” has outliers. These outliers have to be treated in order to build models. We used user defined function to get upper and lower bounds of numeric columns for outlier capping and flooring. After Treating the outliers the following is the result.

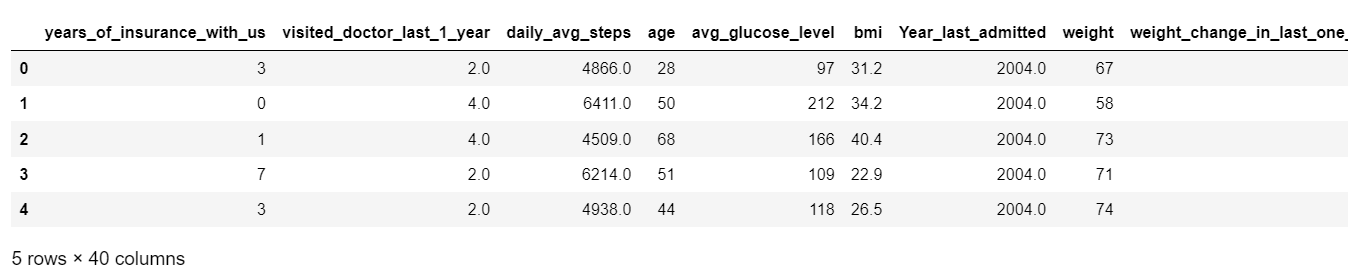
******

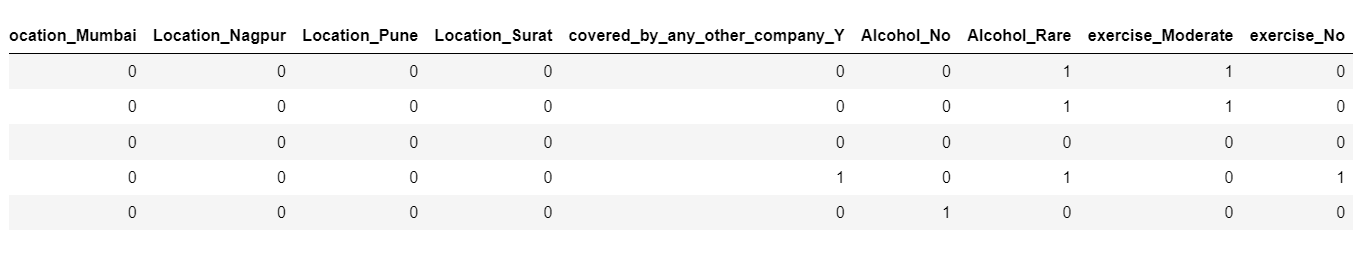
***Figure 56: Dataset after outlier treatment***

**Variable Transformation:**

In order to performing the machine learning algorithms. we need to convert the string values into numeric values. We can use two types of encoding one is label encoding and one hot encoding. For this, here I used one hot encoding method for object data type variables. Because, label encoding work in ranking wise. It will give high preference to the one subcategory this will create bias in our model. In our data all labels are equally important.

After Encoding the dataset looks like following.





***Table 2: Dataset after Encoding (Transformation)***

The shape of the dataset is



1. **MODEL BUILDING**

* Clear on why was a particular model(s) chosen.
* Effort to improve model performance.

**Solution:**

Model building is an essential part of data analytics and is used to extract insights and knowledge from the data to make business decisions and strategies.

For this dataset we will use regression models as the dependent variable (Insurance Cost) is a continuous variable.

Regression analysis is the statistical method used to determine the structure of a relationship between two variables (single linear regression) or three or more variables (multiple regression).

The regression analysis is a predictive method that explores the relationship between a dependent (target) and the independent variable (predictor). In this dataset we want to analyze the relationship between insurance Cost (target variable) and independent variables. We used different regression models to estimate health insurance costs on the basis of 18 independent variables, and by using this regression, we can forecast future health insurance fees based on current and past data.

The Models that are used in this dataset are as follows.

* Linear Regression
* Linear Regression using Stats Model
* Decision Tree Regression Model
* Random Forest Regression Model
* XGBoost Regression Model

Apart from the models used, we will also need to tune the models for better results. For this we will use the following methods.

* Ridge Regression Model
* Lasso Regression Model
* Decision Tree Regression Tuned Model
* Random Forest Regression Tuned Model
* XGBoost Regression Tuned Model

The above Models will tell us the relationship between the dependent and the independent variables and also help us in predicting the future insurance cost, but we need to measure the output. For this we will use the following metrics.

* R squared
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* Mean Absolute Error (MAE).

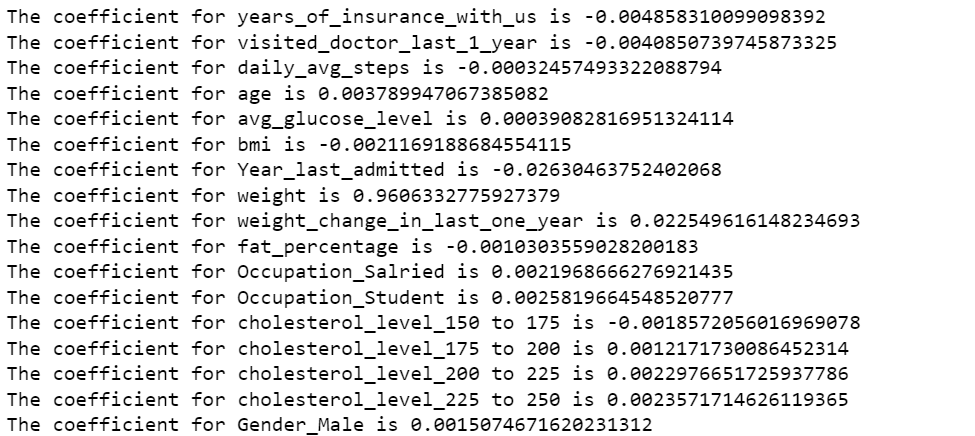
**Linear Regression Model**

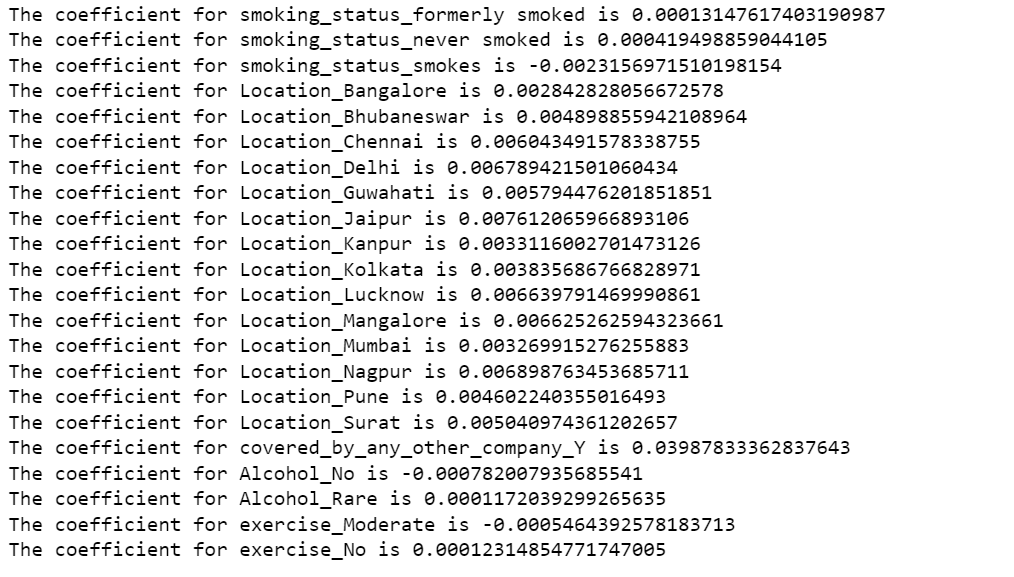
Linear regression analysis is used to predict the value of a variable based on the value of another variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions.

Linear regression not only tests for relationships but also quantifies their direction and strength. The regression coefficient describes the average (expected) change in the dependent variable for each 1-unit change in the independent variable for continuous independent variables or the expected difference versus a reference category for categorical independent variables. When including several independent variables, the regression model estimates the effect of each independent variable while holding the values of all other independent variables constant.

The coefficient of determination, commonly referred to as *R*2, describes the proportion of the variability in the outcome variable that can be explained by the independent variables.

The coefficients of the independent variables are as follows:



******

***Table 3: Coefficients of the independent variables***

**Intercept**: The intercept is the value of the linear predictor when all covariates are zero. Intercept is a point where the graph of the function crosses, or intercepts, the x-axis or y-axis. This function determines the value of the dependent variable when the independent variable is 0 (zero).

**The intercept for our model is -0.0007001713759112928**

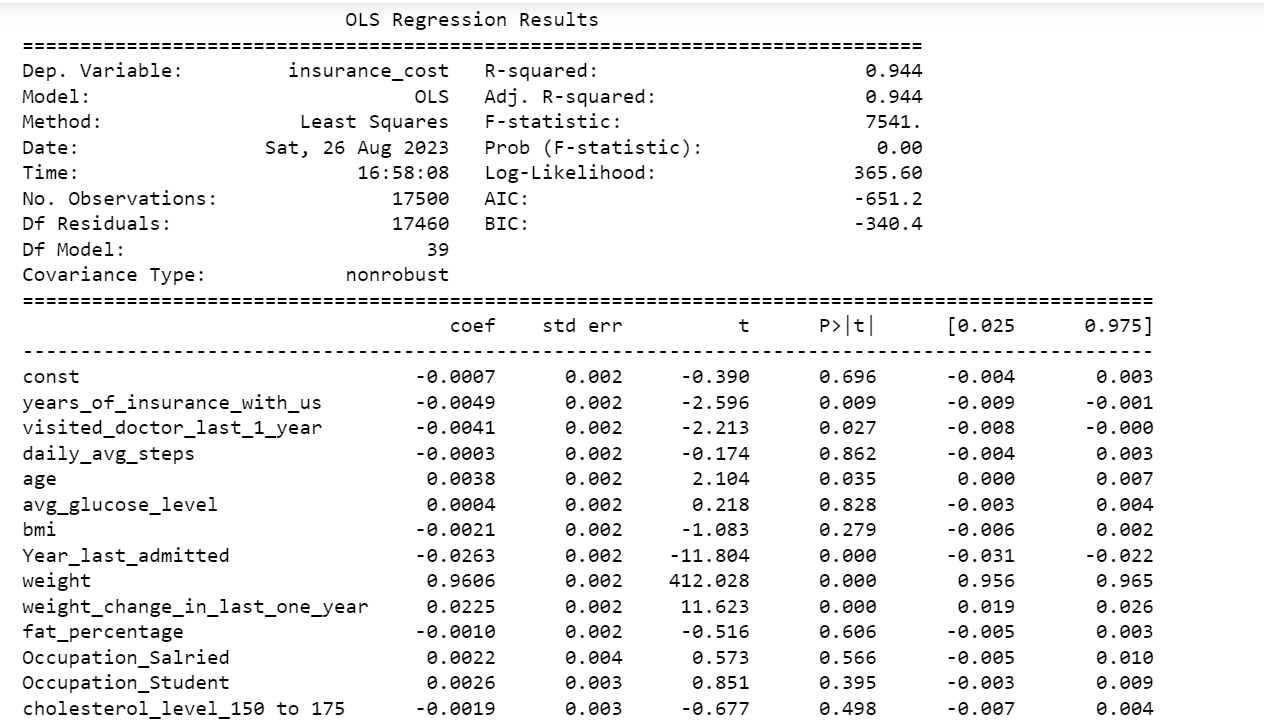
We have performed Linear Regression on our dataset and found out the coefficient of the independent variables. Now we will evaluate the model using our metrics. Following are the score of the metrics.

* R square score of train data = 0.94
* R square score of test data = 0.94
* MSE of the train data set is 0.05
* MSE of the test data set is 0.05
* RMSE of the train data set is 0.24
* RMSE of the test data set is 0.24
* MAE of the train data set is 0.19
* MAE of the test data set is 0.19

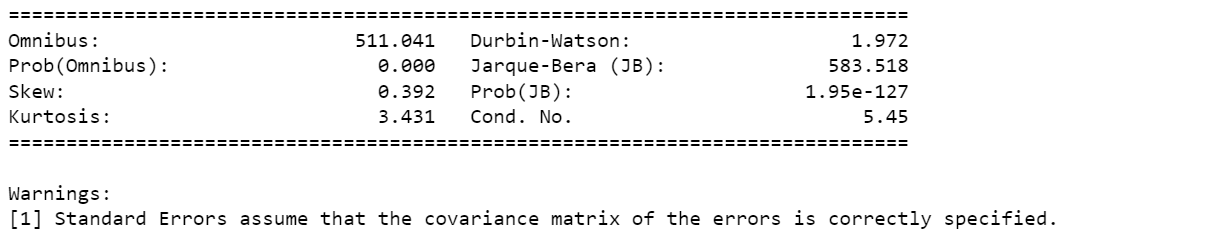
Looking at the scores we can say that the model worked very well. The R squared value for both the train and test data is 94%.

**Linear Regression using Stats Model:**

For this model we will first run a basic model using ordinary least squares method (OLS) where we will take into account all the independent variables. The following is the OLS summary.





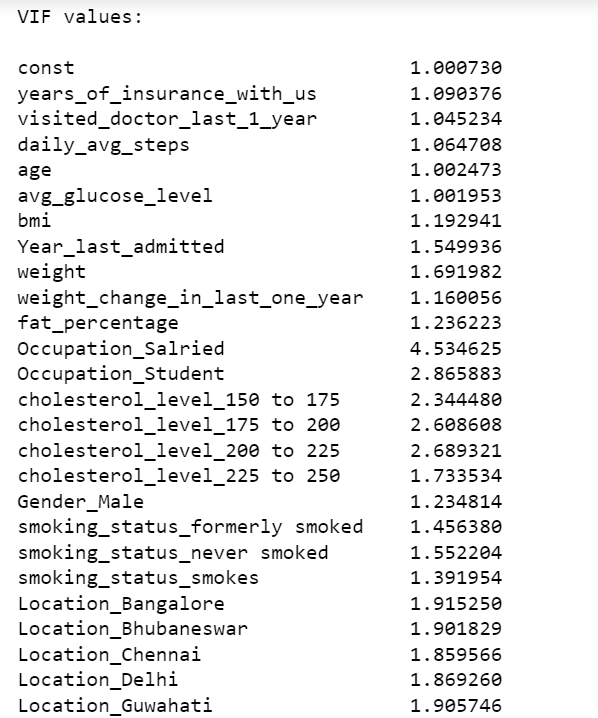


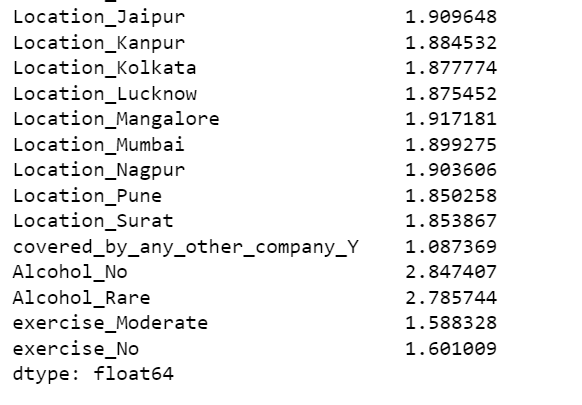
***Table 4: OLS Summary***

We will notice that in this model, R-squared and adjusted R-squared values are similar which holds a value of 0.944 and prob (F-statistic) is 0.00.

P-value of Alcohol Rare, smoking\_status\_formerly smoked and daily\_avg\_steps, avg\_glucose\_level, bmi, cholesterol\_level\_150 to 175, cholesterol\_level\_175 to 200, cholesterol\_level\_200 to 225, cholesterol\_level\_225 to 250, Gender\_Male, smoking\_status\_formerly smoked, smoking\_status\_never smoked, smoking\_status\_smokes, Location\_Bangalore, Location\_Bhubaneswar, Location\_Chennai, Location\_Delhi, Location\_Guwahati, Location\_Jaipur,Location\_Kanpur, Location\_Kolkata, Location\_Lucknow, Location\_Mangalore, Location\_Mumbai, Location\_Nagpur, Location\_Pune, Location\_Surat, Alcohol\_No, Alcohol\_Rare, exercise\_Moderate, exercise\_No, fat\_percentage, Occupation\_Salried, Occupation\_Student variable is quite high and this variables will be eliminated one by one in the next iterative model.

**VIF of the independent variables:**

****

****

***Table 5: VIF of the Independent variables***

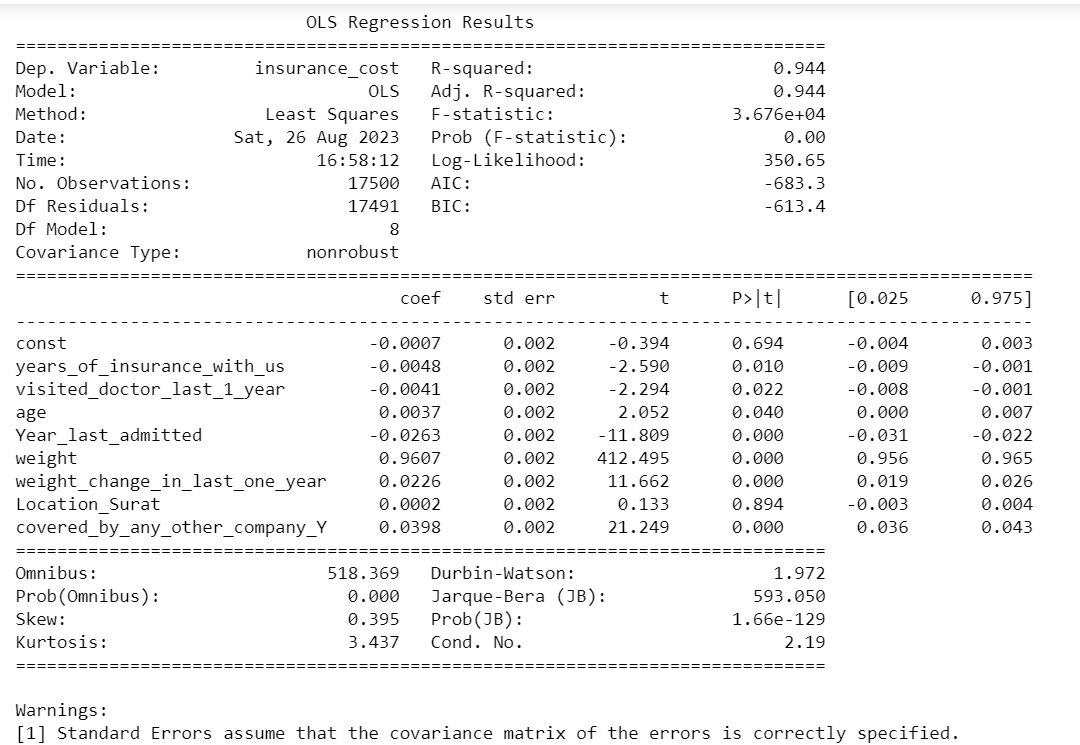
We observe that, there is no multicollinearity within all the variables subjected to base model. Since all the VIF values are within range (less than 5), its better to check p values obtained and make decisions for elimination of variables in the consecutive models.

We will remove the independent variables one by one. After removing the variables which have high P-value, we would be removing 33 variables. Based on above output, we can see clearly after removing 33 variables the R-squared score was still remains constant. So we can say these variables are insignificant for our model.

Now we have 7 variables for building model. But before that we need to check assumptions of linear regression.

**Final Model:**

The summary of the OLS of model 33 is as follows:



***Table 6: OLS model summary of model 33***

**Assumption of Linear Regression:**

These assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or use the model to make a prediction.

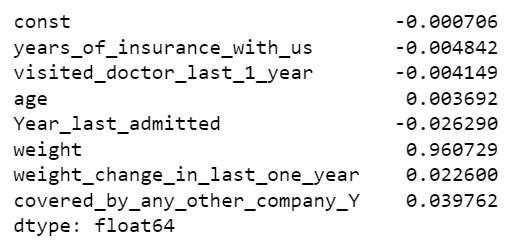
For Linear Regression, we need to check if the following assumptions hold:

* Linearity
* Independence
* Homoscedasticity
* No strong Multicollinearity
* Normality of error terms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

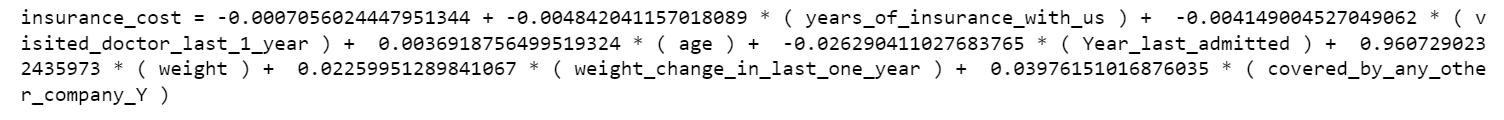
After confirming that all the assumptions are met, we can now make prediction on the dataset.

The Parameters of the model are as follows:



***Table 7: OLS Model 33 parameters***

The equation of the model is as follows:



***Figure 57: Equation of the OLS model 33***

We can now use the model for making predictions on the test data. The measurement of the output is as follows.

* R square score for train data = 0.94
* R square score for test data = 0.94
* MSE for the train data set is 0.05
* MSE for the test data set is 0.05
* RMSE for the train data set is 0.23
* RMSE for the test data set is 0.23
* MAE for the train data set is 0.19
* MAE for the test data set is 0.19

We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from over fitting. MAE indicates that our current model is able to predict insurance cost within a mean error of 0.19 units on the test data.

Hence, we can conclude the model "ols\_res33" is good for prediction as well as inference purposes.

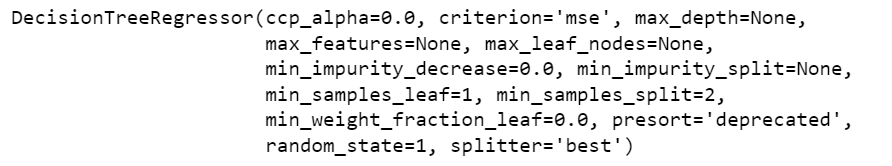
**Decision Tree Model:**

Decision Tree is a supervised learning techniquethat can be used for both classification and Regression problems. It is a tree-structured classifier, whereinternal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

Decision trees and ensemble methods do not require feature scaling to be performed as they are not sensitive to the variance in the data. so this model performed on normal data.

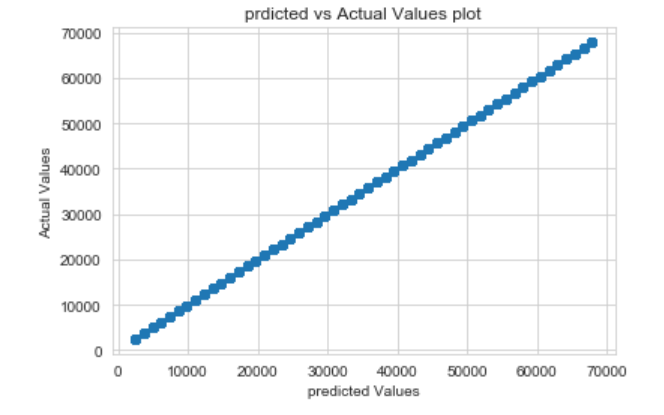
For this, all 40 variables which are obtained after treatments are subjected to decision tree regressor model without any hyperparameters tuning.



***Figure 58: Decision Tree Output***

Model Evaluation:

* R square score for train data = 1.0
* R square score for test data = 0.90
* MSE for the train data set is 0
* MSE for the test data set is 18640969.8496
* RMSE for the train data set is 0
* RMSE for the test data set is 4317.5189
* MAE for the train data set is 0
* MAE for the test data set is 3348.2533



***Figure 59: Predicted vs Actual values on Train Data***

******

***Figure 60: Predicted vs Actual values on Test data***

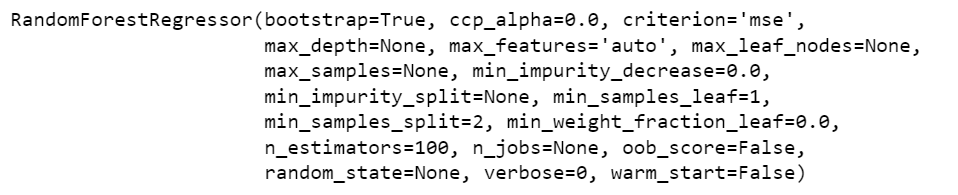
We can see from above graph this model is suffering from over fitting. We need to tune this model for better performance.

**Random Forest Model:**

Random forest regressor is an ensemble model which comes under non-linear category unlike linear regression which is highly depends on correlation of independent variables against target variable. However, this random forest is tree-based algorithm where group of decision trees have been ensembled till refining the results. In tree-based algorithm pruning of hyperparameters has to be made else there can be overgrown tree will result to overfitting of models.

For this, all 40 variables which are obtained after treatments are subjected to decision tree regressor model without any hyperparameters tuning.

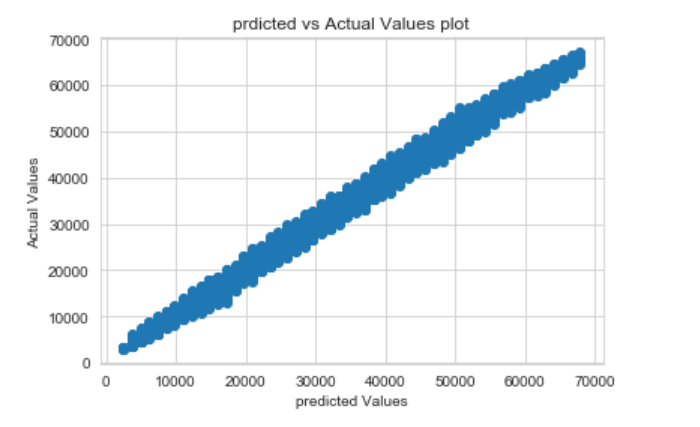
Fitting The model:



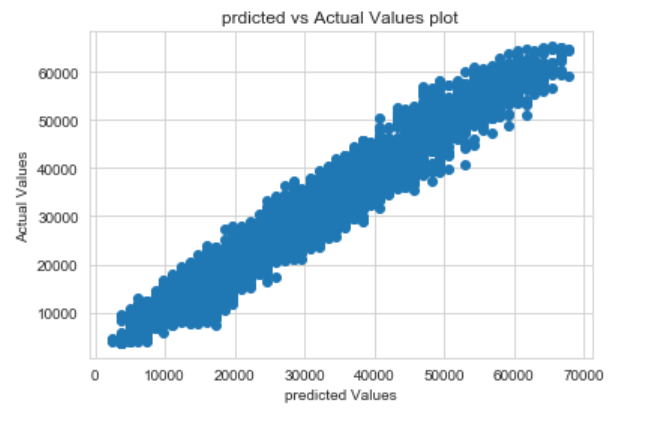
***Figure 61: Model Fitting & output Random Forest***

Model Evaluation:

* R square score for train data = 0.99
* R square score for test data = 0.95
* MSE for the train data set is 1374255.764
* MSE for the test data set is 9496462.5061
* RMSE for the train data set is 1172.2865
* RMSE for the test data set is 3081.5330
* MAE for the train data set is 927.700
* MAE for the test data set is 2450.2682



***Figure 62: Predicted vs Actual Values on Train Data***

******

***Figure 63: Predicted vs Actual on Test Data***

Observations:

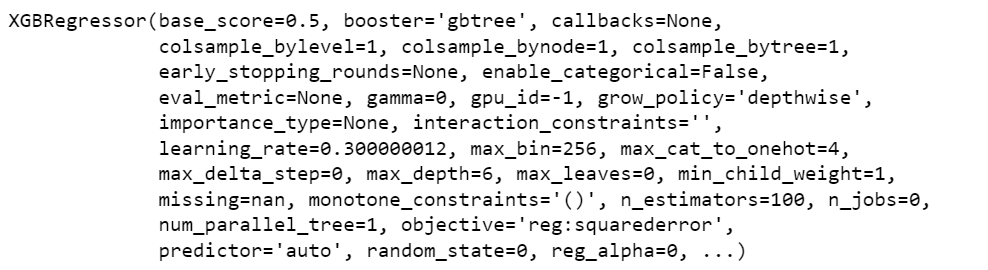
* As expected, an overgrown tree was been obtained with a train accuracy of 99.33% while the test data accuracy was 95.34%.
* We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
* MAE indicates that our current model is able to predict insurance cost within a mean error of 927.70 units on the test data.
* Hence, we can say this model is good for prediction as well as inference purposes.

**XGBoost Model:**

XGBoost (Extreme Gradient Boosting) is a popular supervised-learning algorithm used for regression and classification on large datasets. This gives high priority to weaker models in predicting target variable and gives priority to the weaker models in next iteration to make it stronger. Also, it does L1 and L2 regularization while reducing the complexity and suits high multicollinearity and thus gets a name extra gradient boosting.

For this a basic XGB model has been constructed with no hyperparameter tuning where it is highly possible to get an overgrown tree like other tree algorithms.

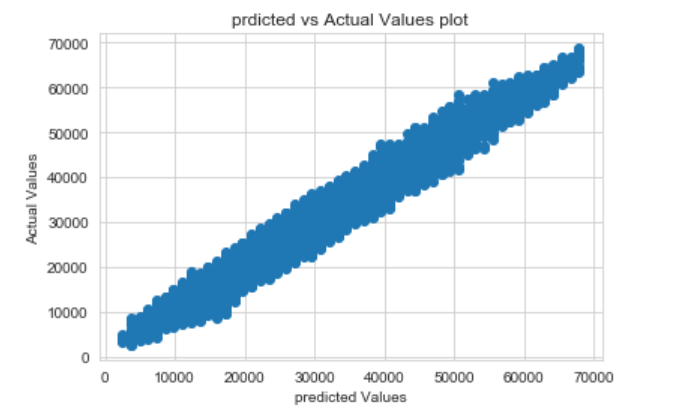
Fitting the model:



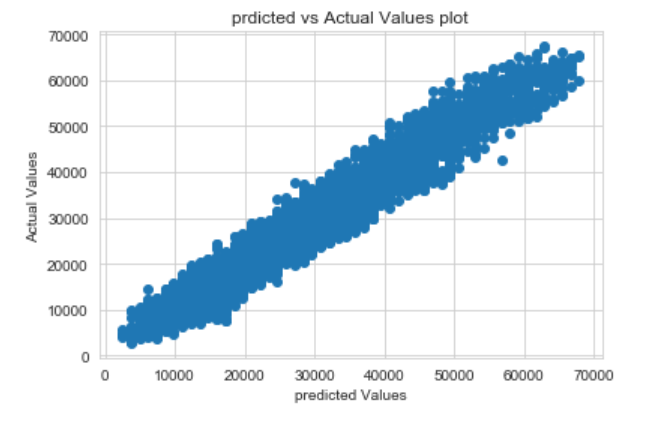
***Figure 64: XGBoost model fitting in data***

Model Evaluation:

* R square score for train data = 0.97
* R square score for test data = 0.95
* MSE for the train data set is 4766029.5521
* MSE for the test data set is 9750370.8370
* RMSE for the train data set is 2183.1238
* RMSE for the test data set is 3122.5583
* MAE for the train data set is 1714.7351
* MAE for the test data set is 2485.7351



***Figure 65: Predicted vs Actual on Train data***



***Figure 66: Actual vs Predicted on Test Data***

Observations:

We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting. MAE indicates that our current model is able to predict insurance cost within a mean error of 1714.73 units on the test data.

1. **Model Tuning:**

* How was the model validated? Just accuracy, or anything else too?

**Solution:**

**Linear regression model tuning:**

We tune the model to**maximize model performances without overfitting and reduce the variance error**in our model. We have to apply the appropriate Hyperparameter technique for our model.

**Regularization for liner regression model:**

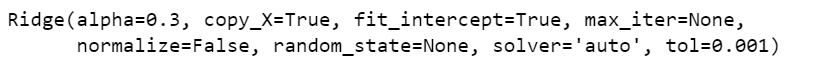
It reduces the overfitting nature of the model. Even if the model works well, this is done in order to prevent the problem from occurring in the future. This is done by introducing more errors and making the model learn more.

* Coefficient shrinks whenever we do regularization.
* We need to make sure that our model doesn’t get under-fitted by tuning too much in alpha as well.
* Alpha is a penalty factor. Error is introduced in the system by drawing a line that’s doesn’t touch the majority of the points. .
* Shrinkage in coefficient totally depends on the variables. Ifthe feature is significant then the shrinkage will be less but if the feature is not significant then shrinkage will more.
* If the feature is highly insignificant then the coefficient will become 0. The advantage of this regularizes models is that even if the assumptions are not checked the model will do all the work.

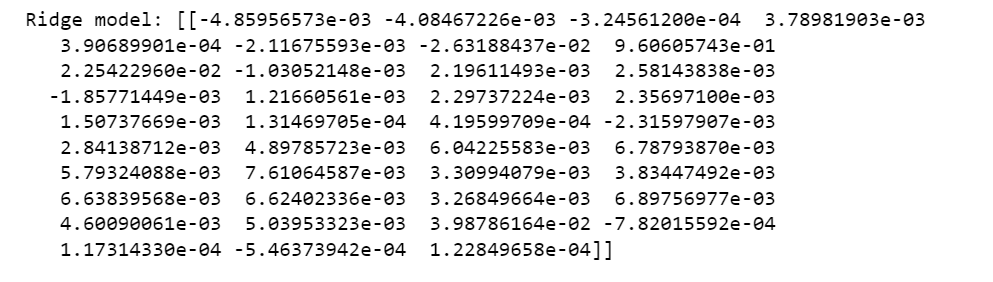
**Regularised RIDGE model:**

The regularized RIDGE model adds the “Squared magnitude” of coefficient as a penalty term to the loss function. It is called an L2 penalty.

Model Fitting:



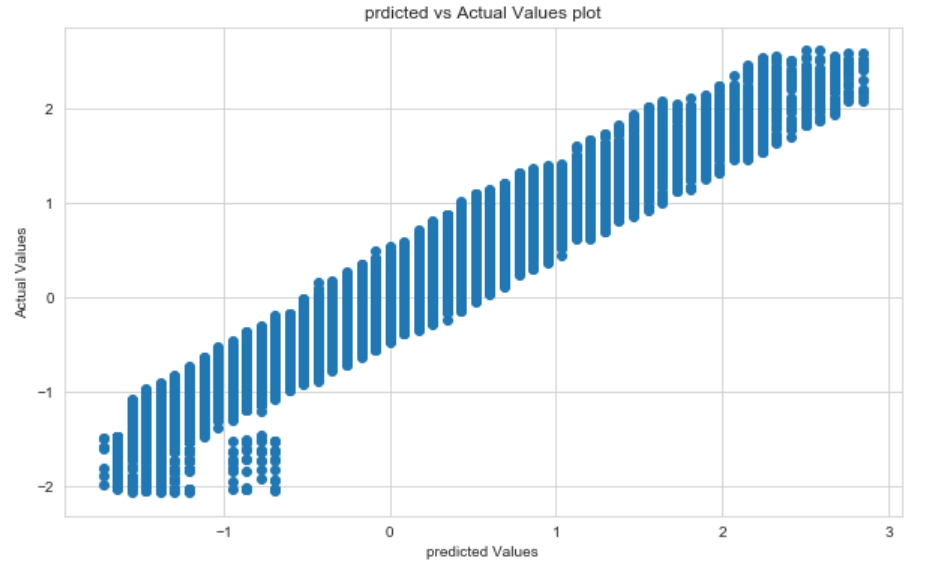
Coefficients:



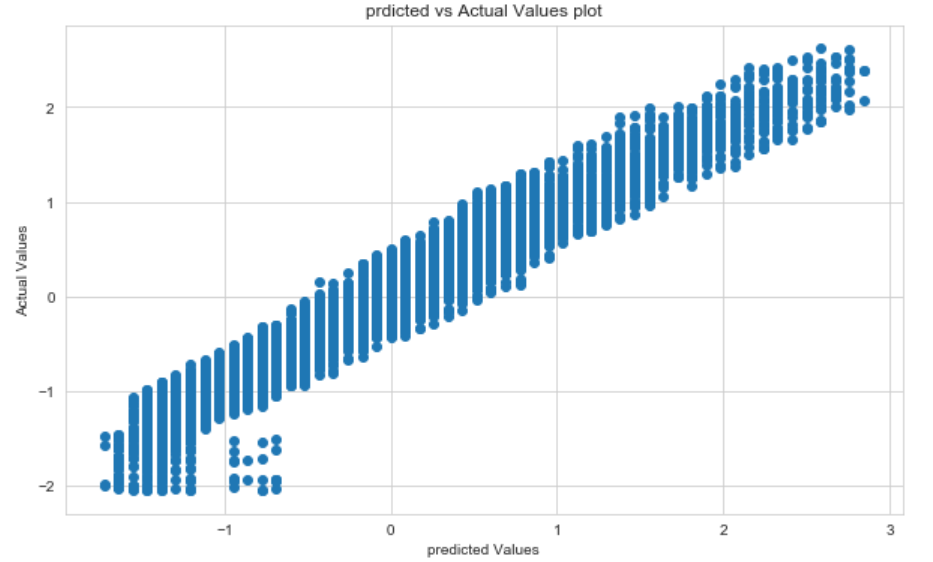
***Table 8: Coefficients of the Regularised RIDGE model***

Model Evaluation:

* R square score for train data = 0.94
* R square score for test data = 0.94
* MSE for the train data set is 0.05
* MSE for the test data set is 0.05
* RMSE for the train data set is 0.23
* RMSE for the test data set is 0.23
* MAE for the train data set is 0.19
* MAE for the test data set is 0.19



***Figure 67: Predicted vs Actual on Train Data***

******

***Figure 68: Predicted vs Actual on Test Data***

Observations:

A ridge regression result shows a similar behavior to linear regression on all variables where similar R score, MSE, RMSE and MAE scores has been visualized. Model prediction on dependent variable seems like it is predicting better both in train and test data.

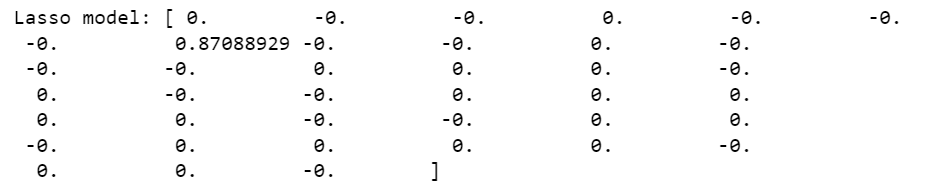
**Regularized LASSO Model:**

The (least absolute shrinkage and selection operator) adds the “**Absolute value of magnitude**” of coefficient as a penalty term to the loss function. It is called an L1 penalty.

Model Fitting:



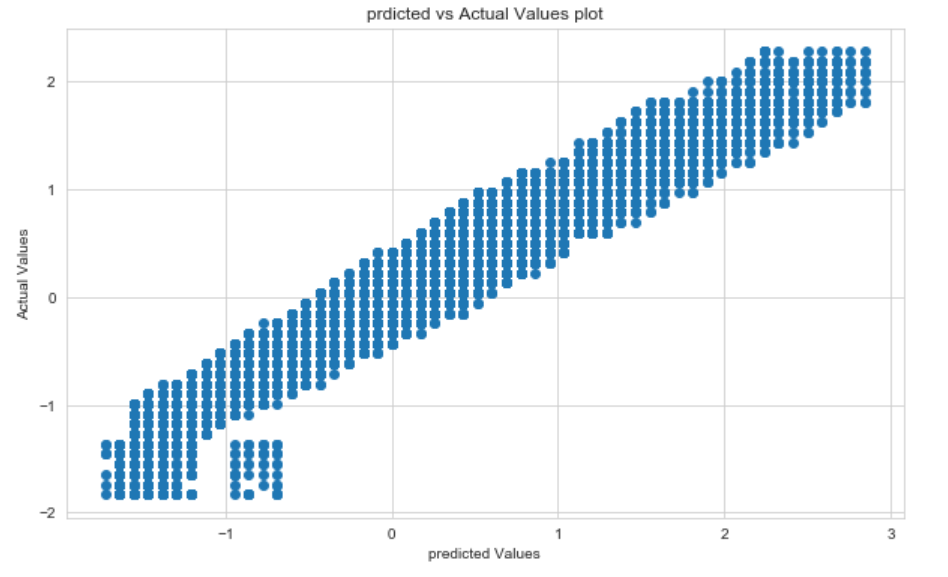
Coefficients:



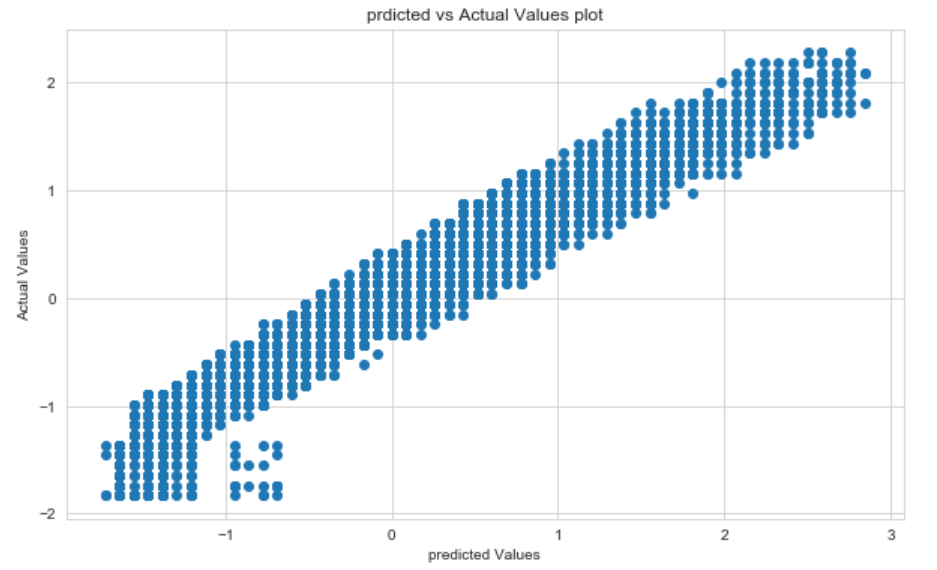
***Table 9: Coefficients of the LASSO Model***

Model Evaluation:

* R square score for train data = 0.93
* R square score for test data = 0.93
* MSE for the train data set is 0.06
* MSE for the test data set is 0.06
* RMSE for the train data set is 0.26
* RMSE for the test data set is 0.26
* MAE for the train data set is 0.20
* MAE for the test data set is 0.20



***Figure 69: Predicted vs Actual on Train Data***

******

***Figure 70: Predicted vs Actual on Test Data***

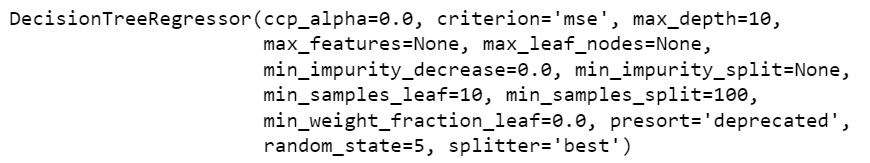
Observations:

* Based above coefficients this model explains the weight variable is more significant to predict insurance cost. The remaining variables are insignificant.
* From above output this model performance metrics like RMSE, MSE and MAE scores were slightly closer to previous models, however, there was seen slight reduction in R squared score in both train and test data.
* These methods were used within variables with high multicollinearity. But in this dataset, there is no multicollinearity issue with the clean data as the VIF scores seen were better.

**Decision Tree Model Tuning:**

Now we will tune the Decision Tree model to optimise our output. The parameters used in this model is max\_depth = 10, min\_samples\_leaf=10, min\_samples\_split=100, random\_state=5.

Model Fitting:

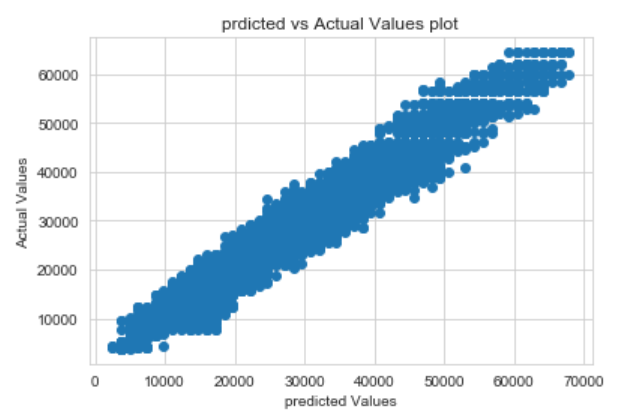


Model Evaluation:

* R square score for train data = 0.95
* R square score for test data = 0.95
* MSE for the train data set is 8448300.4801
* MSE for the test data set is 952442.5285
* RMSE for the train data set is 2906.5960
* RMSE for the test data set is 3086.1663
* MAE for the train data set is 2317.5676
* MAE for the test data set is 2450.6341



***Figure 71: Predicted vs Actual on Train Data***

******

***Figure 72: Predicted vs Actual on Test Data***

Observation:

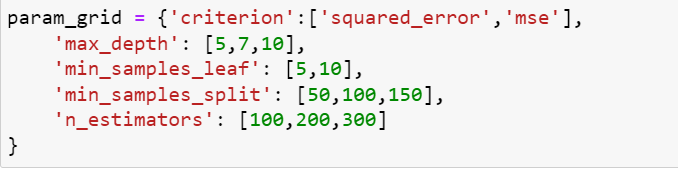
Upon running this regressor model with the above best variables an accuracy of about 95% and 95% has been obtained for train and test data and models seems to be more promising with lower difference in RMSE train and test scores now we have prevented our model from overfitting.

**Random Forest Model Tuning:**

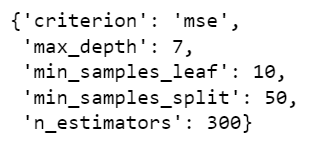
From the lessons of overgrown basic model, hyperparameters tuning has been performed using GridSearchCV function where a list of parameters was sent over a dictionary type and analyze best hyperparameters.

Grid Search for finding out the optimal values for the hyper parameters.

These are the parameters were passed in GridSearchCV with CV value of 3,



The Best parameters from Grid Search CV is:



Model Evaluation:

* R square score for train data = 0.95
* R square score for test data = 0.95
* MSE for the train data set is 8734539.4018
* MSE for the test data set is 9143981.3825
* RMSE for the train data set is 2955.4254
* RMSE for the test data set is 3023.9016
* MAE for the train data set is 2372.0435
* MAE for the test data set is 2416.7032



***Figure 73: Predicted vs Actual on Train Data***

******

***Figure 74: Predicted vs Actual on Test Data***

Observation:

Upon running a pruned random forest regressor with the above best variables an accuracy of about 95% and 95% has been obtained for train and test data while the difference is very less and models seems to be more promising with lower difference in other error metrics on train and test scores.

As discussed, both test and train accuracies have been reflected clearly on the plots where both plots give a similar representation which shows how the difference is very less between train and test data.

**XGBoost Model Tuning:**

XGBoost hyperparameter tuning involves adjusting the various hyperparameters of the algorithm to optimize the model's performance.

As a part of tuning of hyperparameters, variables like n\_estimators and max\_depth has been passed through a dictionary with these following values.

n\_estimators = 100,150,200

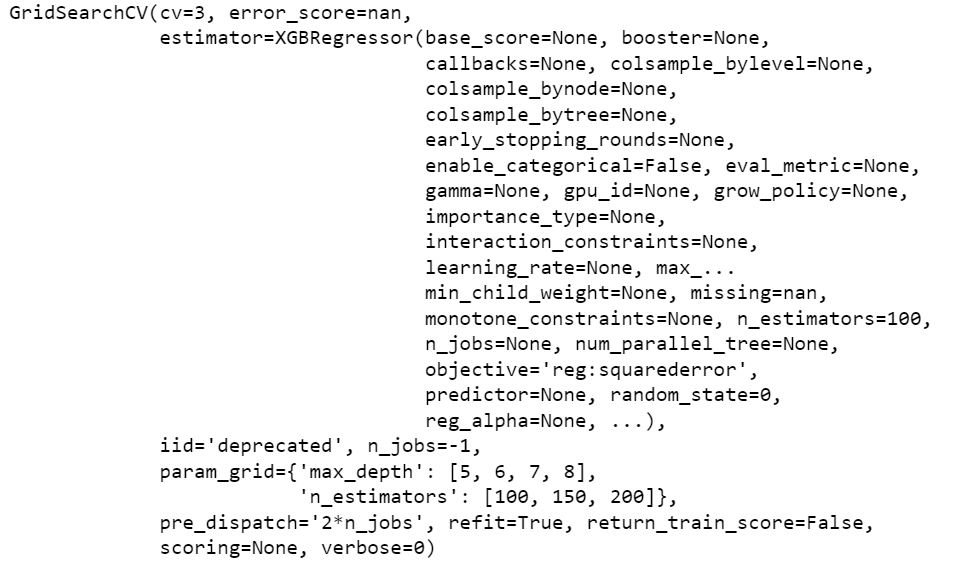
max\_depth = 5,6,7, 8

Upon running through GridSearchCV with CV value of 3 following results have been obtained,

n\_estimators = 100

max\_depth =5

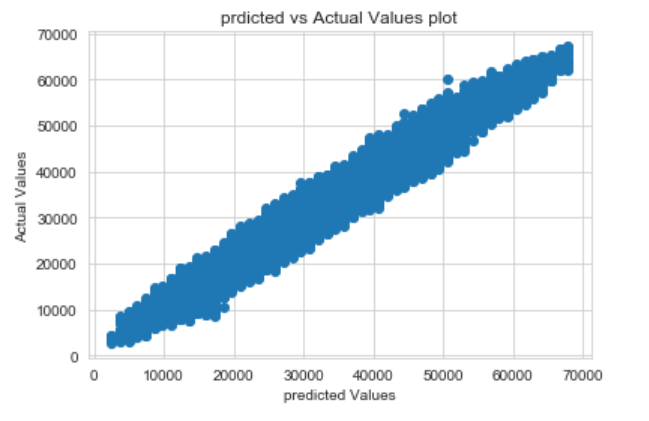
Model Fitting:



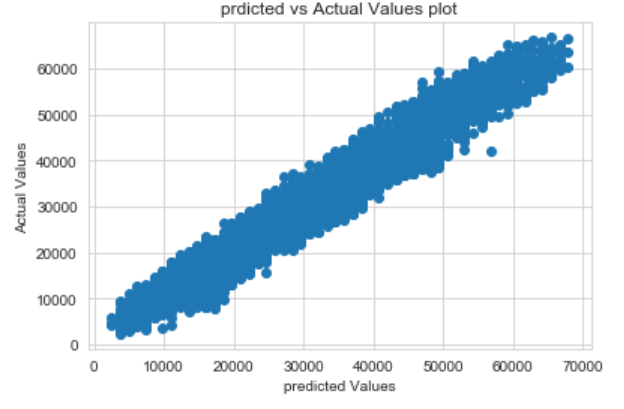
***Figure 75: XGBoost model Fitting***

Model Evaluation:

* R square score for train data = 0.96
* R square score for test data = 0.95
* MSE for the train data set is 6390348.6132
* MSE for the test data set is 9443093.7915
* RMSE for the train data set is 25279138
* RMSE for the test data set is 3072.9617
* MAE for the train data set is 2015.3899
* MAE for the test data set is 2458.9762



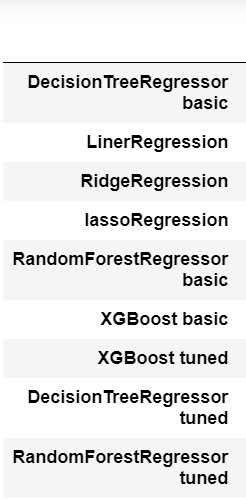
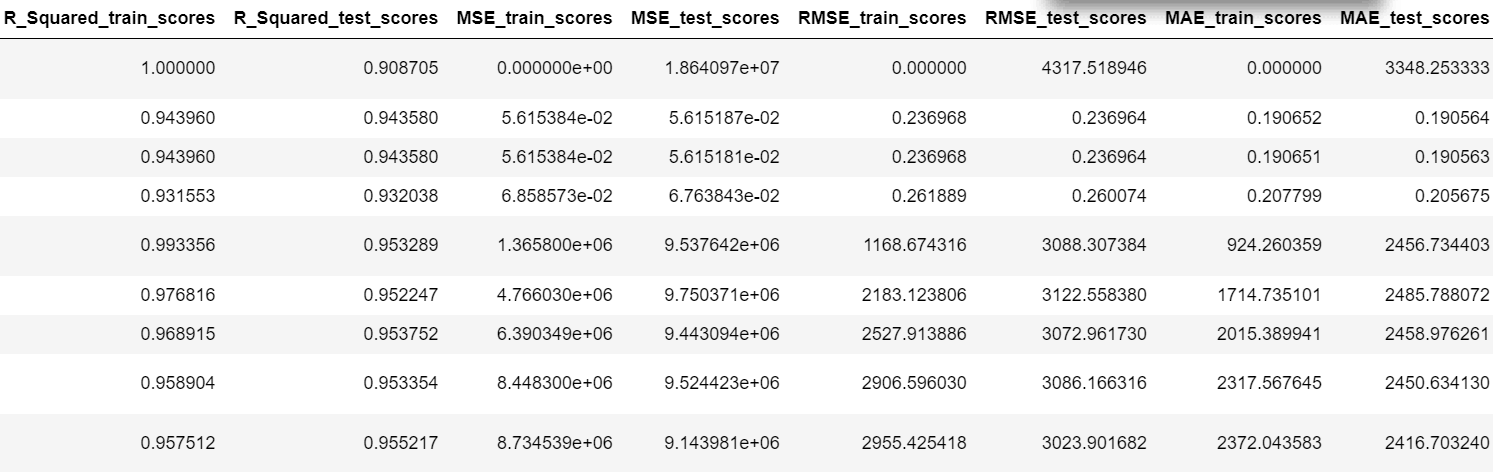
***Figure 76: Predicted vs Actual on Train Data***

******

***Figure 77: Predicted vs Actual on Test Data***

**Interpretation:**

After performing the models and tuning them to get the optimum, the following is the overview of all the models’ performance:

***Table 10: Evaluation metrics of all the models with tuned models***

Observation:

* Based on the results obtained, this model performed well on both train and test data. The other error metric scores are low.
* As discussed, both test and train accuracies have been reflected clearly on the plots where both plots give a similar representation

1. **Final interpretation / recommendation**

* Detailed recommendations for the management/client based on the analysis done.

**Solution:**

**Insights:**

* It is observed that the variables years of insurance with us, visited\_doctor\_last\_1\_year, age, Year\_last\_admitted, weight, weight\_change\_in\_last\_one\_year and covered\_by\_any\_other\_company\_Y variables are most influential independent variables for predicting insurance cost.
* The liner regression, ridge and lasso models got less RMSE values as well as 94% of the variation in the insurance cost is explained by the predictors in the model for train set.
* Random forest regressor basic model performs well among all the models. This model explains 99% of the variation in the insurance cost is explained by the predictors in the model for train set with less error.
* To conclude finally, Random Forest regressor basic model holds best for predicting product insurance cost.
* We have found variables years of insurance with us, visited\_doctor\_last\_1\_year, age, Year\_last\_admitted, weight, weight\_change\_in\_last\_one\_year and covered\_by\_any\_other\_company\_Y variables are most influential independent variables for predicting insurance cost.

So, we can fit a random forest model in to the training set using only these 7 variables and predict the optimum insurance cost.

**Recommendations:**

* While doing liner regression stasmodels we have found which variables are most significant and how these attributes increase or decrease insurance cost those are when **age** increases by 1 unit, insurance\_cost increases by 0.0036 units, keeping all other predictors constant.
* Similarly, when **weight** increases by 1 unit, insurance\_cost increases by 0.96 units, keeping all other predictors constant.
* When **weight\_change\_in\_last\_one\_year** increases by 1 unit, insurance\_cost increases by 0.02 units, keeping all other predictors constant.
* When **covered\_by\_any\_other\_company\_Y** increases by 1 unit, insurance\_cost increases by 0.03 units, keeping all other predictors constant.
* There are also some negative co-efficient values, for instance, **years\_of\_insurance\_with\_us** has its corresponding co-efficient as -0. 0048.This implies, when the customer has insurance with this company more than year, the insurance cost decreases by -0.0048 units, keeping all other predictors constant.
* Similarly,the co-efficient of **visited\_doctor\_last\_1\_year** has -0.004, so customer who went for health checkup last year , the insurance cost reduced by -0.004 units, keeping all other predictors constant.
* Similarly,the co-efficient of **Year\_last\_admitted** has -0.026, so customer who have admitted in hospital last year, the insurance cost reduced by - -0.026 units, keeping all other predictors constant.
* So, by using these variables we can predict optimum insurance cost.
* While performing Univariate analysis we found it most of the customers are students followed by business people and very few members are salaried.
* So, the insurance company should focus on job holders to enroll for an insurance policy. The company must provide optimal insurance cost for these people.
* The locations Surat, Kolkata and Lucknow are the slightly lowest number of customers. So, insurance companies should focus on marketing to uninsured population with low health costs to maximize success and increase potential profit.
* While observing insurance costs with regular check-ups last year variable we found whoever has gone for regular checkups got paid less insurance cost. So, the company will conduct workshops and encourage customers to go for regular health checkups.
* Variable weight is positively correlated with insurance cost this means if the person's weight increases the insurance cost also increased.
* So, we need to provide some weight loss programs and encourage them to follow these programs. And creating awareness of healthy weight will prevent health risks.
* To get a better idea of the predicted health costs for a potential customer, insert attributes into our predictive model.