HEALTHCARE PROJECT

CAPSTONE

PROJECT2

DIKSHA SHUKLA

Table of Contents

Contents

[Introduction: 2](#_Toc91454783)

[Objective: 2](#_Toc91454785)

[Data Description 3](#_Toc91454786)

[Sample of the dataset: 3](#_Toc91454787)

[Exploratory Data Analysis 3](#_Toc91454788)

[Let us check the types of variables in the data frame. 3](#_Toc91454789)

[Check for missing values in the dataset: 4](#_Toc91454790)

Univariate Analysis 5

Bivariate Analysis 18

Multivariate Analysis 19

Outlier Treatment 19

Model Building and interpretation 20

Business Insights 31

Recommendations 31

**Table of Figures:**

[Figure 1: years\_of\_insurance\_with\_us 7](#_Toc95049659)

[Figure 2 regular\_checkup\_lasy\_year 7](#_Toc95049660)

[Figure 3 adventure\_sports 8](#_Toc95049661)

[Figure 4 visited\_doctor\_last\_1\_year 8](#_Toc95049662)

[Figure 5 daily\_avg\_steps 9](#_Toc95049663)

[Figure 6 age 9](#_Toc95049664)

[Figure 7 heart\_decs\_history 10](#_Toc95049665)

[Figure 8 other\_major\_decs\_history 10](#_Toc95049666)

Figure 9 avg\_glucose\_level………………………………………………………………………………………… 10

[Figure 10 bmi 11](#_Toc95049667)

[Figure 11 Year\_last\_admitted 11](#_Toc95049668)

[Figure 12 weight 12](#_Toc95049669)

[Figure 13 fat\_percentage 12](#_Toc95049670)

[Figure 14 insurance\_cost 13](#_Toc95049671)

[Figure 15 Occupation 13](#_Toc95049672)

[Figure 16 cholesterol\_level 14](#_Toc95049673)

[Figure 17 Gender 14](#_Toc95049674)

[Figure 18 smoking\_status 15](#_Toc95049675)

[Figure 19 Location 15](#_Toc95049676)

[Figure 20 covered\_by\_any\_other\_company 16](#_Toc95049677)

[Figure 21 Alcohol 16](#_Toc95049678)

[Figure 22 exercise 17](#_Toc95049679)

[Figure 23 Bivariate Analysis 17](#_Toc95049680)

[Figure 24 Multivariate Analysis 18](#_Toc95049681)

[Figure 25: Scatterplot 24](#_Toc95049683)

# Introduction:

# 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 individuals eat healthy and do proper exercise the chance of getting ill is drastically reduced.

# Objective:

The Goal & 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

**Data** **Dictionary** **for** **Market** **Segmentation:**

**Variable Business Definition**

applicant\_id unique ID for every applicant

years\_of\_insurance\_with\_us Since how many years customer is taking policy from the same company only

regular\_checkup\_last\_year Number of times customers has done the regular health check-up in last one year

adventure\_sports Customer is involved with adventure sports like climbing, diving etc.

Occupation Occupation of the customer

visited\_doctor\_last\_1\_year Number of times customer has visited a doctor in last one year

cholesterol\_level Cholesterol level of the customers while applying for insurance

daily\_avg\_steps Average daily steps walked by customers

age Age of the customer

heart\_decs\_history Does customers have any past heart diseases

other\_major\_decs\_history Does customers have any past major diseases apart from heart like any operation

Gender Gender of the customer

avg\_glucose\_level Average glucose level of the customer while applying the insurance

bmi BMI of the customer while applying the insurance

smoking\_status Smoking status of the customer

Year\_last\_admitted When customer has been admitted to the hospital last time

Location Location of the hospital

weight Weight of the customer

covered\_by\_any\_other\_company Customer is covered from any other insurance company

Alcohol Alcohol consumption status of the customer

exercise Regular exercise status of the customer

weight\_change\_in\_last\_one \_year How much variation has been seen in the weight of the customer in last year

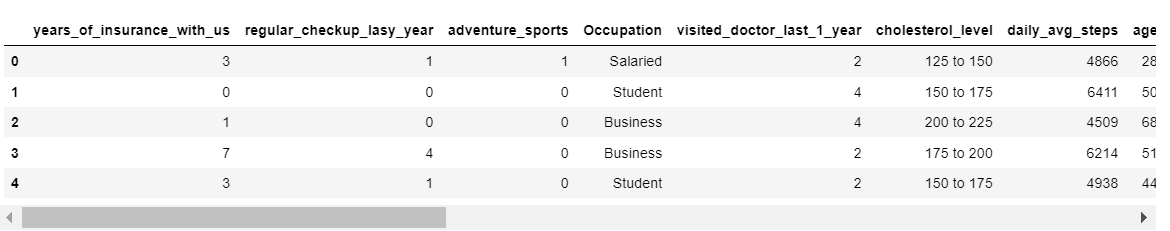
fat\_percentage Fat percentage of the customer while applying the insurance

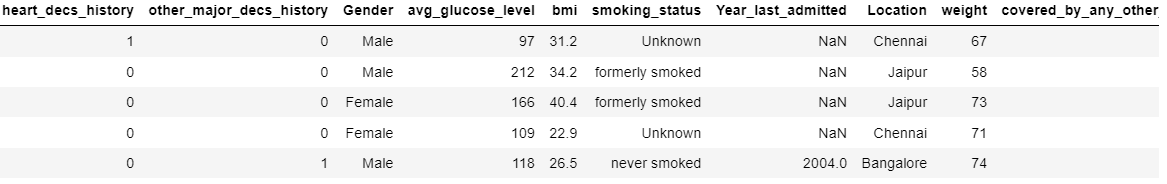
insurance\_cost Total Insurance cost

# Data Description

## Sample of the dataset:

As the dataset contains many columns so displayed all the columns in more than 1 row

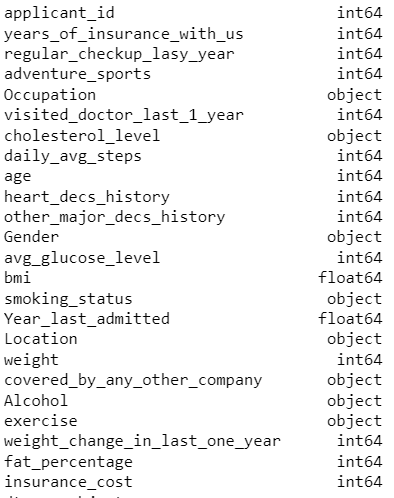






# Exploratory Data Analysis

## Let us check the types of variables in the data frame.



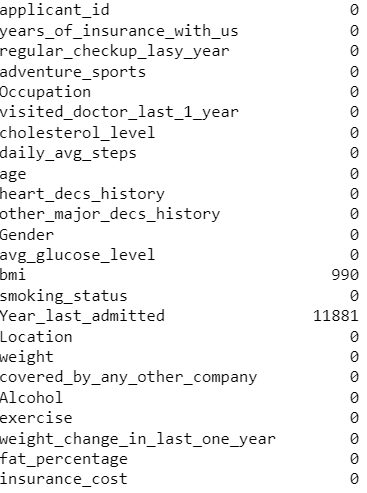
There is total of 25000 rows and 24 columns in the dataset.

We are dropping the **insurance\_id** column as it is not making any sense in the prediction of our goal.

Therefore, the new dataset contains 23 columns.

## Check for missing values in the dataset:

In order to check missing values in Pandas DataFrame, we use a **function isnull() and notnull()**. Both function help in checking whether a value is NaN or not. This function can also be used in Pandas Series in order to find null values in a series.



Column **BMI** and **Year\_last\_admitted** have null values.

**Null Value Treatment:**

As there are more than 30%of null values present for the year\_last\_admitted column so we have to drop the column whereas the BMI column has less than 30% so we are imputing null values using KNN.

(The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using **the mean value of the 'k'-neighbors found in the dataset**.)

**Univariate analysis:**

**Univariate analysis** is a form of quantitative, statistical, evaluation. This method of **analysis** separately studies the findings regarding each variable in a data set, and therefore each variable is summarised on its own.

1. Boxplot for years\_of\_insurance\_with\_us:

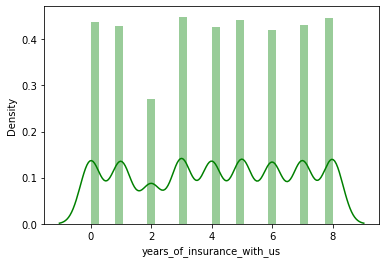
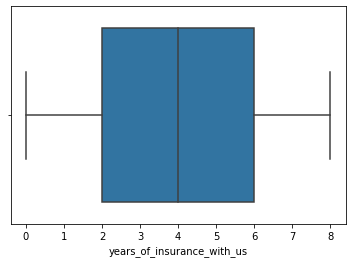
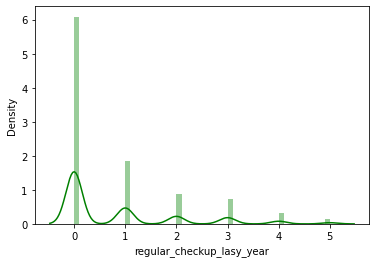
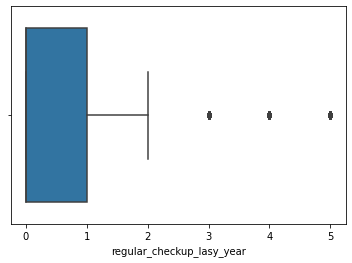


Figure :years\_of\_insurance\_with\_us

Boxplot shows there are no outliers present in the dataset.

The maximum years\_of\_insurance\_with\_us of the customers falls under the category on a scale of 0-8.

1. Boxplot and Distplot for regular\_checkup\_last\_year:

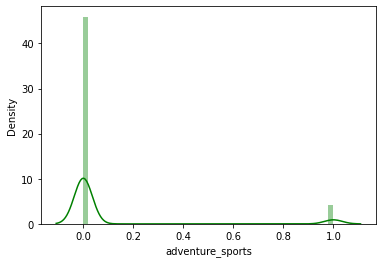


Figure

Boxplot for regular\_checkup\_last\_year variable shows few outliers.

The maximum regular\_checkup\_last\_year of the customers falls under the category on a scale of 0-5.

1. Boxplot and Distplot for adventure\_sports

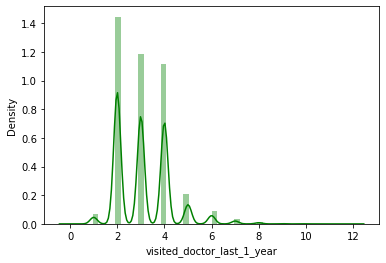
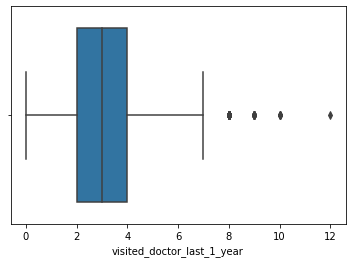


Figure

Boxplot for adventure\_sports variable shows negligible outlier.

The distplot for adventure\_sports is rightly skewed.

1. Boxplot for visited\_doctor\_last\_1\_year:

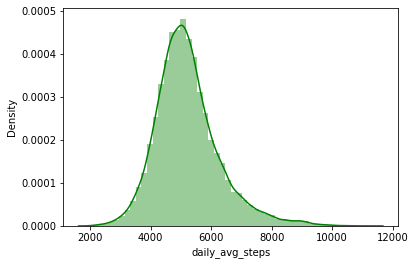
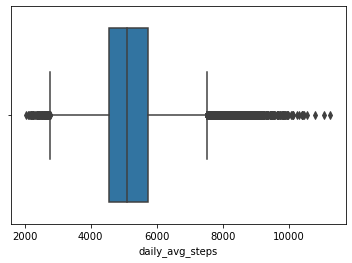


Figure

Boxplot for visited\_doctor\_last\_1\_year variable shows few outliers.

The maximum visited\_doctor\_last\_1\_year of the customer falls under the category on a scale of 0-12.

1. Boxplot and Distplot for daily\_avg\_steps:

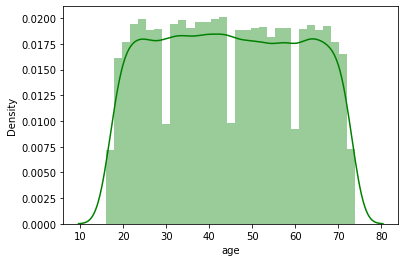
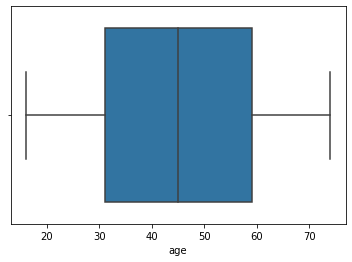


Figure

Boxplot for daily\_avg\_steps variable shows outliers.

Displot for daily-avg\_steps is normally distributed and ranges between 2034-11255.

1. Boxplot and Distplot for age:

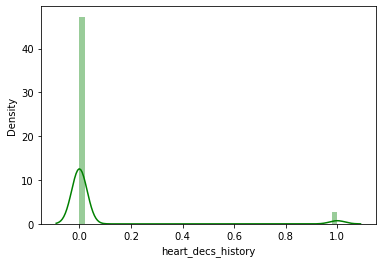


Figure

Boxplot for Age variable has no outliers.

The maximum Age of the customers falls under the category on a scale of 16-74.

1. **Boxplot and Distplot for heart\_decs\_history:**

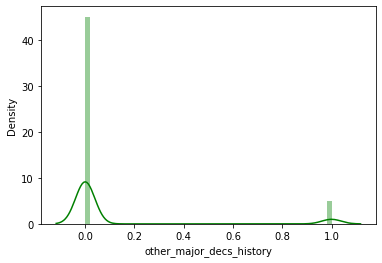


Figure

Boxplot for heart\_decs\_history variable shows negligible outliers.

Distplt for heart\_decs\_history is rightly skewed.

1. **Boxplot and Distplot for other\_major\_dec\_history:**

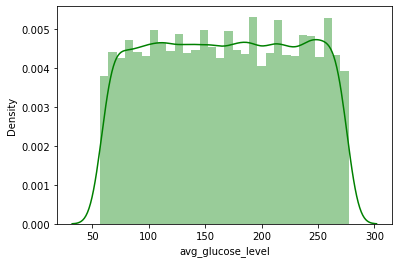
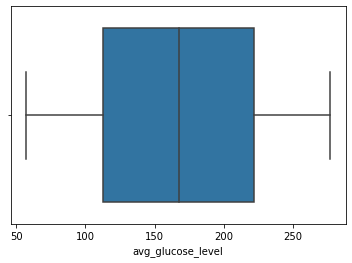


Figure

Boxplot for heart\_decs\_history shows negligible outliers.

Distplot for heart\_decs\_history is rightly skewed and ranges between 0-1.

1. **Boxplot and Distplot for avg\_glucose\_level:**

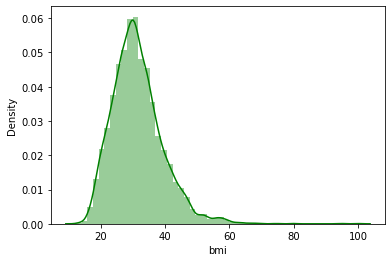
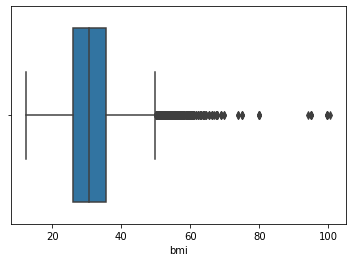


Figure

Boxplot for avg\_glucose\_level has 0 outliers.

Distplot for avg\_glucose\_level ranges between 57 -277.

1. **Boxplot and Distplot for bmi:**

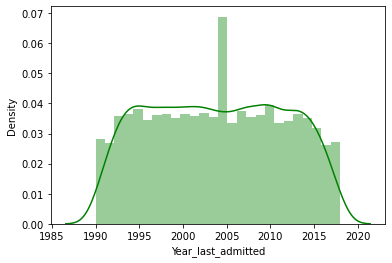
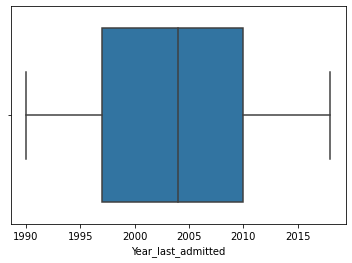


Figure

Boxplot for BMI has good number of outliers.

Distplot for BMI is normally distributed.

1. **Boxplot and Distplot for Year\_last\_admitted:**

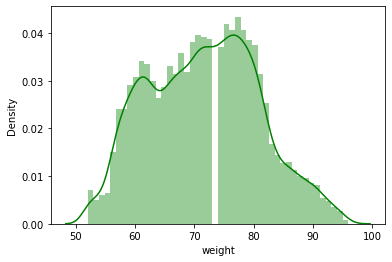
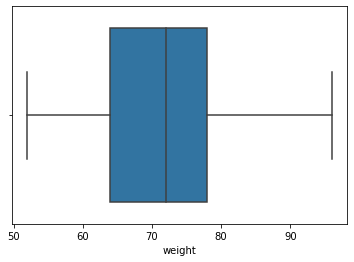


Figure

Boxplot for Year\_last\_admitted has no outliers.

Distplot for Year\_last\_admitted ranges between 1990-2018.

1. **Boxplot and Distplot for Weight:**

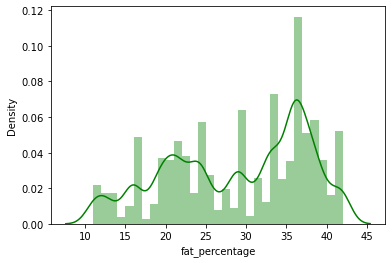
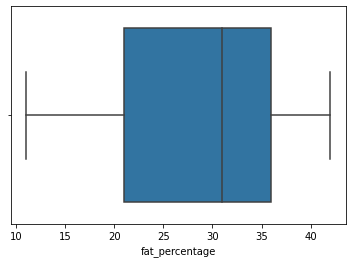


Figure

Boxplot for weight has no outliers.

Distplot for weight ranges between 52- 96 and is normally distributed.

1. **Boxplot and distplot for fat\_percentage:**

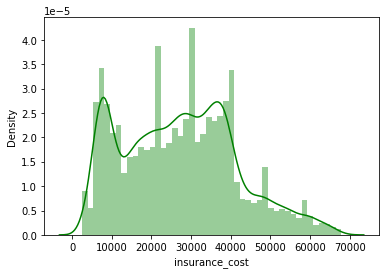
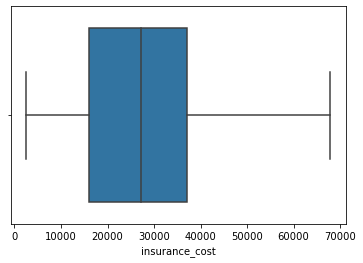


Figure

Boxplot for fat\_percentage has no outliers.

Distplot for fat\_percentage ranges between 11-42.

1. **Boxplot and Distplot for insurance\_cost:**



Figure

Boxplot for insurance\_cost has no outliers.

Dsitplot for insurance\_cost ranges between 2468-67870.

**Categorical Plot:**

1. **Occupation:**

Occupation Count for Student and Business is almost similar however, Student count is slightly more than Business and Occupation count for Salaried is lowest.

****

Figure

1. **Cholesterol level:**

As from the below plot we can see that most of the Customer’s Cholesterol level ranges between 150 to 175 and the lowest level ranges between 225 to 250.

****

Figure

1. **Gender:**

Count for Male gender is more than Female gender.



Figure

1. **Smoking\_Status:**

Count for Never\_Smoked customer is more than the customer who smokes.

****

Figure

1. **Location:**

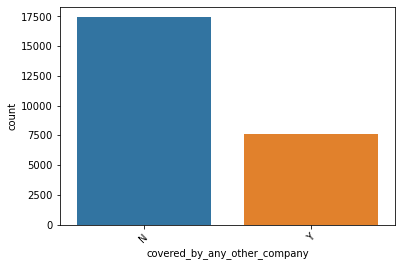
Customers from the Bangalore location are in more number.

****

Figure

1. **Covered\_by\_any\_other\_company:**

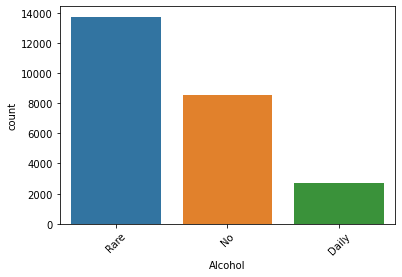
Count for Customers who are covered from any other insurance company is less than the customer who are covered from any other company.

****

Figure

1. **Alcohol:**

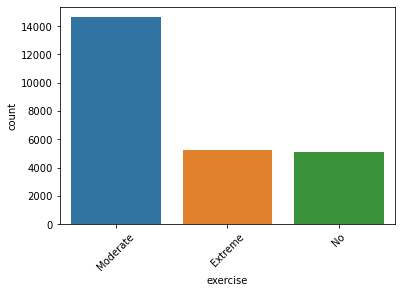
**Customer who drinks Alcohol rarely is more than those who drink daily.**

****

Figure

1. **Exercise:**

Moderate levels of exercise done by the customers were counted for extreme and no exercise have almost the same number of customers.

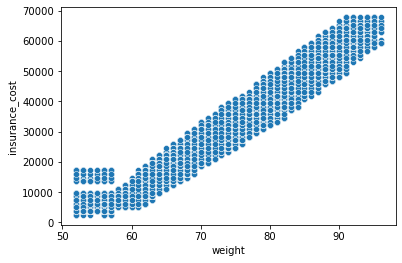
****

Figure

**BIVARIATE ANALYSIS:**

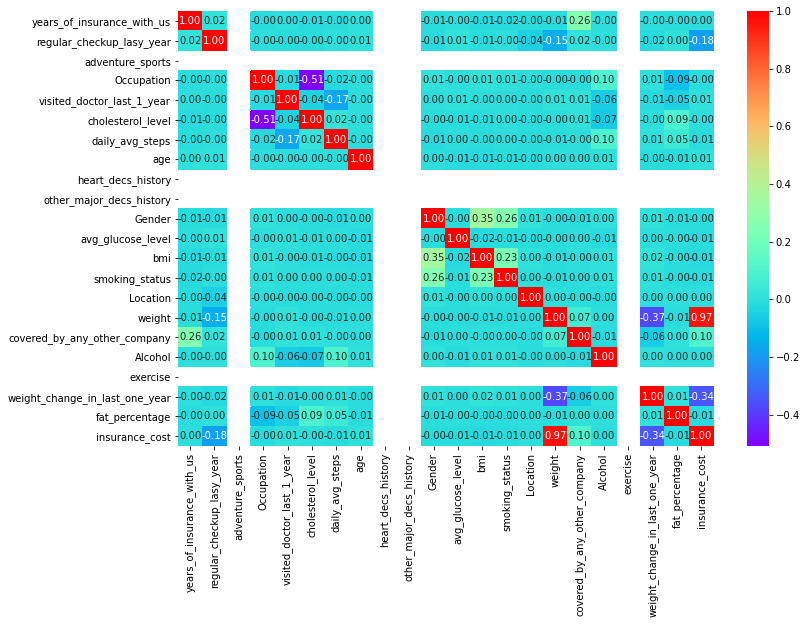
**insurance\_cost vs weight:**

Weight is highly correlated to insurance\_cost which means as the weight of customer increases insurance costs will also get increase.

****

Figure

**Multivariate Analysis:**

****

Figure

From the above heat map, we can see only the weight variable is highly correlated by a value of 0.97 and the rest variables are poorly correlated.

**Multicollinearity:**

Multicollinearity occurs **when two or more independent variables are highly correlated with one another in a** regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

**Outliers Treatment:**

We have identified outliers are present in few of our predictor variables using **boxplot** so we are treating the outliers using IQR methodology.

(**IQR (Interquartile Range)** is the difference between the third and the first quartile of a distribution (or the 75th percentile minus the 25th percentile). It is a measure of how wide our distribution is since this range contains half of the points of the dataset).

**Note: Our data is also not unbalanced. Since our target variable is continuous so we are not required to check data imbalance.**

**Model building and interpretation.**

We are using various types of linear and non-linear models to find the best results on finding the optimum insurance cost which is our target variable.

* Linear Regression
* Ridge Regression
* Lasso Regression
* Random Forest Regressor
* Adaboost Regressor

Linear Regressions and Adaboost Regressor are sensitive to outliers whereas Lasso, Ridge and Random Forest regressors models are robust to outliers and in case, we find high multicollinearity in the features. These are used when there is high multicollinearity among features. It adds a small squared biased factor to the variables but reduces the variances. Multilinear regression is common & most popular, yet impacted by outliers & multicollinearity. So, the mentioned models tried to overcome such problems. Adaboost is less prone to overfitting as the input parameters are not jointly optimized and it combines multiple “weak classifiers” into a single “strong classifier”.

**Scaling:**

Feature **Scaling** or Standardization: It is a step of Data Pre-Processing that is applied to independent variables or features of data. It helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm. Package Used: sklearn.preprocessing.

The dataset contains variables on a different scale. Therefore, it will be wiser to scale or standardize the data to make or compare all variables on the same scale.

**Does Scaling necessary:**

Scaling is not necessary. We will get the same result even if we apply to scale into the dataset. The only change we will get is in intercept and the coefficient of variables will get change. There will be no change in the accuracy score.

But recommended for regression techniques as well because it would help gradient descent to converge fast and reach the global minima. When several features become large, it helps is running the model quickly else the starting point would be very far from minima if the scaling is not done in preprocessing.

For now, we will process the model without scaling and later we will check the output with scaled data of regression model output.

**Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE.**

1. **Linear Regression:**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

We have split the data into train and test (70:30) and then applied linear regression.

After applying linear regression, the coefficients for each of the independent attributes are as follows:

The coefficient for years\_of\_insurance\_with\_us is -13.332115684590633

The coefficient for regular\_checkup\_lasy\_year is -621.3863666641173

The coefficient for adventure\_sports is -2.8990143619012088e-12

The coefficient for Occupation is 44.087091655973445

The coefficient for visited\_doctor\_last\_1\_year is -34.59082998404557

The coefficient for cholesterol\_level is 38.32161393808311

The coefficient for daily\_avg\_steps is -0.028729454495204414

The coefficient for age is 2.785596549296299

The coefficient for heart\_decs\_history is 2.1032064978498966e-12

The coefficient for other\_major\_decs\_history is 2.2737367544323206e-13

The coefficient for Gender is 47.62809559083542

The coefficient for avg\_glucose\_level is 0.35861728474397997

The coefficient for bmi is -0.44856924400874437

The coefficient for smoking\_status is -2.3924194268315513

The coefficient for Location is 9.135336389525415

The coefficient for weight is 1489.5365632965184

The coefficient for covered\_by\_any\_other\_company is 1210.7008412174446

The coefficient for Alcohol is 6.139718895206911

The coefficient for exercise is 0.0

The coefficient for weight\_change\_in\_last\_one\_year is 172.2466977295284

The coefficient for fat\_percentage is -1.0540647749231655

**Intercept:**

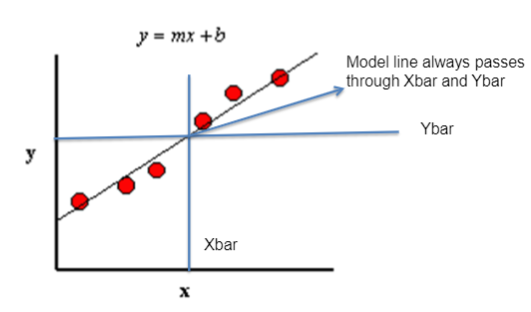
The intercept is always a constant value. We get the intercept value (Y, expected mean value) when all value of X is 0.

**Equation:** y= mx+c, where c is our intercept.



We got an intercept value of -79924.13 which means the insurance cost is -79924which is meaningless**.** To make an intercept near 0 we can do a Z score or scale the data.

**Coefficient of the determinant (RSquare)** –determines the fitness of a linear model. R-squared is always between 0 and 1**.** The closer the points get to the line, the R^2 (coeff of determinant) tends to 1, the better the model is and if the R^2 is 0 it means none of the variability of the response data around its mean.



**R-square = Explained variation / Total variation**

In this regression model, we got the

R-square value on

Training set: **0.9446898302914063 and**

Test set **: 0.931543712584074**.

**Root mean squared** **error (RMSE):** RMSE is a quadratic scoring rule that also measures the average magnitude of the error. It’s the square root of the average squared differences between prediction and actual observation.

In the scikit learn library, sklearn. metrics have a mean\_squared\_error function. The RMSE is just the square root of values it returns.

**RMSE on Training data:** 3379.3906460758326

**RMSE on Testing data:** 3336.465447198216

Below plot shows the predicted **y value vs actual y values** for the test data

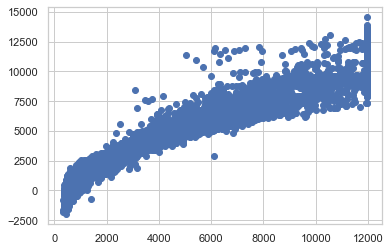


Figure 26:Scatterplot

From the plot, we can see that it is a linear plot with some kind of noise present on the data set which is Unexplained variances on the output. And it shows a very strong correlation between the predicted y and actual y.

**On applying Zscore:**

**Coefficient and intercept values got changed, below are the values:**

The coefficient for years\_of\_insurance\_with\_us is -0.0024187742207027116

The coefficient for regular\_checkup\_lasy\_year is -0.0396313595099337

The coefficient for Occupation is 0.002758032026856226

The coefficient for visited\_doctor\_last\_1\_year is -0.002682234480762974

The coefficient for cholesterol\_level is 0.0033590730156461304

The coefficient for daily\_avg\_steps is -0.0019329843550842064

The coefficient for age is 0.0031227330050908495

The coefficient for Gender is 0.0015729943800697603

The coefficient for avg\_glucose\_level is 0.0015657036387794721

The coefficient for bmi is -0.0002252960088277609

The coefficient for smoking\_status is -0.00017831817152631982

The coefficient for Location is 0.002752113153754733

The coefficient for weight is 0.9695420296334851

The coefficient for covered\_by\_any\_other\_company is 0.038825648208478544

The coefficient for Alcohol is 0.0002908604382110392

The coefficient for weight\_change\_in\_last\_one\_year is 0.020307020931331864

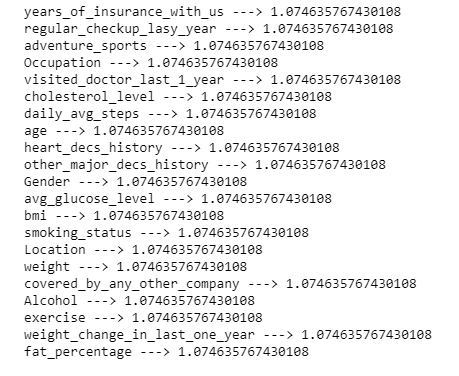
The coefficient for fat\_percentage is -0.0006327781944087147



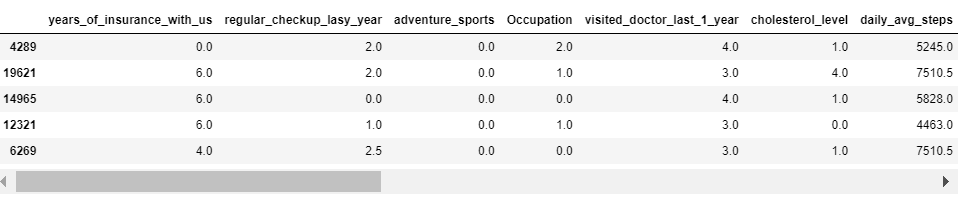
**Accuracy Score for Training data:** 0.9446898302914063

**Accuracy Score for Training data:** 0.9449125189257257

### **On checking Multi-collinearity using VIF**

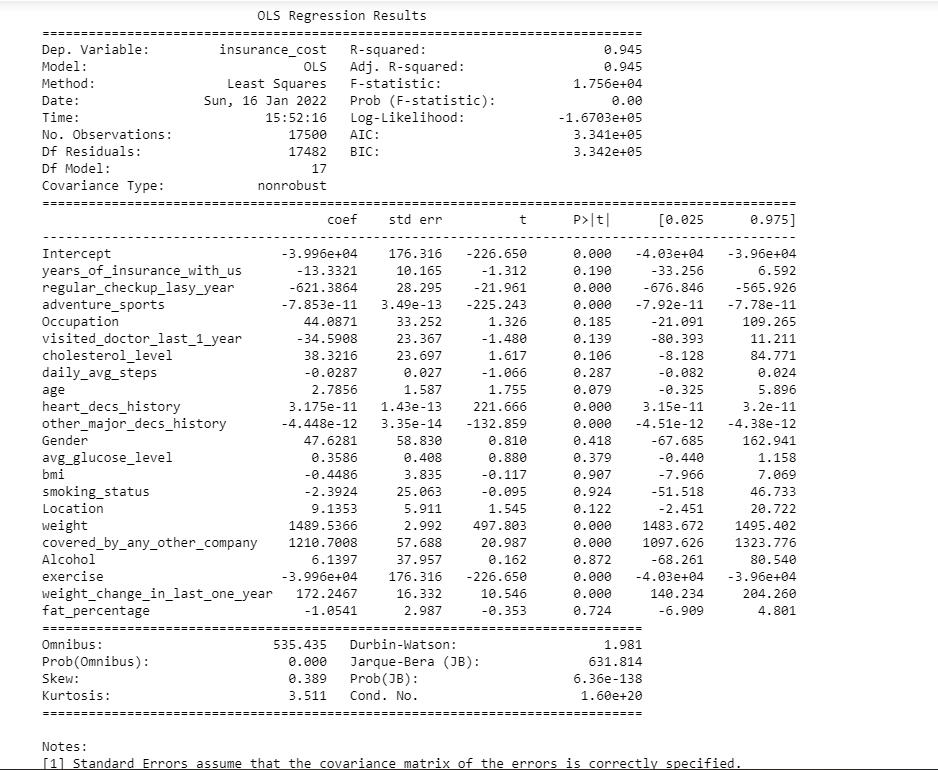


**LINEAR REGRESSION USING STATS MODEL:**



**Root Mean Squared Error** **Root Mean Squared Error**

**Root Mean Squared Error**



Let’s suppose the null hypothesis is true which means there is no relationship between independent variables with insurance cost. If we consider that we have drawn the sample from the universe and, on this sample, we have found many variables have p-value greater than alpha.

The variables whose p values is less than 0.05 are:

regular\_checkup\_last\_year

adventure\_sports

heart\_decs\_history

heart\_decs\_history

covered\_by\_any\_other\_company

exercise

weight\_change\_in\_last\_one\_year

We can see from above that the overall P-value is less than alpha, so rejecting the Null hypothesis (H0) and accepting the alternative hypothesis (Ha).

The p-value for those variables which is higher than 0.05 we can conclude those dimensions are useless and is a poor predictor for insurance cost variable as p-value greater than 0.05.

Root mean square error value for Train data: 3379.3906460758362

Root mean square error value for Test data: 3336.4654471982017

**Inference: Basis on these predictions, the business insights, and recommendations.**

**Inference:**

We have observed a linear plot with a strong correlation between the predicted y and actual y.

**Linear regression Performance Metrics:**

The intercept for the model: **-79924.13**

R square on training data: **0.9446898302914063**

R square on testing data: **0.931543712584074**

RMSE on Training data: **3379.3906460758326**

RMSE on Testing data: **3336.465447198216**

The value for training data & testing data score is fairly good enough, therefore, we can conclude this model is a **Right-Fit Model**.

**Impact of scaling:**

On applying z score the intercept became **5.944478035053907e-16** from -79924.13. And the co-efficient value has also got changed, the bias became nearly zero but the overall accuracy is still the same.

**Multi collinearity:**

We have observed there is no multicollinearity present in the data set.

**From statsmodels:**

On using the stats model

RMSE on Train data: 3379.3906460758362

RMSE on Test data: 3336.4654471982017

1. **Ridge Regression:**

Ridge regression is similar to linear regression where the objective is to find the best fit surface. The difference is in the way the best coefficients are found. Unlike linear regression where the optimization function is SSE, here it is slightly different.

The term is like a penalty term used to penalize large magnitude coefficients when it is set to a high number, coefficients are suppressed significantly. When it is set to 0, the cost function becomes the same as the linear regression cost function.

**Coefficient:**

The coefficient for years\_of\_insurance\_with\_us is -13.327254071635672

The coefficient for regular\_checkup\_lasy\_year is -621.372810758898

The coefficient for adventure\_sports is 0.0

The coefficient for Occupation is 44.085505710795466

The coefficient for visited\_doctor\_last\_1\_year is -34.59002549654029

The coefficient for cholesterol\_level is 38.32079257378632

The coefficient for daily\_avg\_steps is -0.028729017831163266

The coefficient for age is 2.785599270662597

The coefficient for heart\_decs\_history is 0.0

The coefficient for other\_major\_decs\_history is 0.0

The coefficient for Gender is 47.623532752751316

The coefficient for avg\_glucose\_level is 0.35861722292623427

The coefficient for bmi is -0.4484679271297991

The coefficient for smoking\_status is -2.3918501770905856

The coefficient for Location is 9.13534533881278

The coefficient for weight is 1489.5366745581177

The coefficient for covered\_by\_any\_other\_company is 1210.5946588654647

The coefficient for Alcohol is 6.138723310558234

The coefficient for exercise is 0.0

The coefficient for weight\_change\_in\_last\_one\_year is 172.2443110856351

The coefficient for fat\_percentage is -1.0540306052859587

**Ridge score for Train Data:** 0.9446898302799794

**Ridge score for Test Data:** 0.9449125348172092

1. **Lasso Regression:**

Lasso Regression is similar to the Ridge regression with a difference in the penalty term. Unlike Ridge, the penalty term here is raised to power 1. Also known as L1 norm.

Coefficient:

The coefficient for years\_of\_insurance\_with\_us is -13.292929828218638

The coefficient for regular\_checkup\_lasy\_year is -621.2730030946703

The coefficient for adventure\_sports is 0.0

The coefficient for Occupation is 43.87450750045001

The coefficient for visited\_doctor\_last\_1\_year is -34.52232682642922

The coefficient for cholesterol\_level is 38.174544203458424

The coefficient for daily\_avg\_steps is -0.028704526742215852

The coefficient for age is 2.785317220259223

The coefficient for heart\_decs\_history is 0.0

The coefficient for other\_major\_decs\_history is 0.0

The coefficient for Gender is 47.05009836103102

The coefficient for avg\_glucose\_level is 0.3585335974716412

The coefficient for bmi is -0.4374417368625487

The coefficient for smoking\_status is -2.2510697588030033

The coefficient for Location is 9.130992715224544

The coefficient for weight is 1489.5358996667512

The coefficient for covered\_by\_any\_other\_company is 1210.1476951895268

The coefficient for Alcohol is 5.933323704454971

The coefficient for exercise is 0.0

The coefficient for weight\_change\_in\_last\_one\_year is 172.20473678569735

The coefficient for fat\_percentage is -1.0522246197456646

**Lasso score for Train Data:** 0.944689829261956

**Lasso score for Test Data:** 0.9449127590479226

**Ensemble Modelling:**

The goal of ensemble regression is **to combine several models to improve the prediction accuracy in learning problems with a numerical target variable**. The process of ensemble learning can be divided into three phases: the generation phase, the pruning phase, and the integration phase.

**AdaBoostRegressor:**

An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases.

AdaBoostRegressor\_Score on Test data: 0.9454320614606616

**RandomForestRegressor:**

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

**RandomForestRegressor**\_Score on Test data: 0.9510682635093043

**Model Evaluations:**

In the performance matrix, Adjusted R2 should have been computed as R2 may have noise in it as it starts increasing. R2 & Adj R2 should be closer for a better model. Likewise, MAPE should be

considered as it is better than RMSE, coz it’s the square of absolute values and robust to outliers

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Train** | **Linear Regression** | **Random Forest** | **Ridge Regressor** | **Lasso Regressor** | **Adaboost Regressor** |
| **Adjusted R2 on train set** | 0.944643329 | 0.993325021 | 0.944643329 | 0.944643328 | 0.94776095 |
| **RMSE** | 3379.390646 | 1173.49 | 3379.390646 | 3379.390678 | 3282.85 |
| **MAPE** | 15.352914 | 4.568562 | 15.352898 | 15.352865 | 15.966052 |
|  |  |  |  |  |  |
| **Test** | **Linear Regression** | **Random Forest** | **Ridge Regressor** | **Lasso Regressor** | **Adaboost Regressor** |
| **Adjusted R2 on train set** | 0.944866205 | 0.951476608 | 0.944866221 | 0.944866445 | 0.946313425 |
| **RMSE** | 3336.465447 | 3131.38 | 3336.464966 | 3336.458175 | 3292.38 |
| **MAPE** | 15.098864 | 12.282444 | 15.098849 | 15.098744 | 15.900906 |

**After applying different models, we have found that all the models are giving almost**

**similar results except Random Forest as it’s overfitting because it gives very good results**

**for the training set and poor results for the test set. It’s therefore can be concluded that**

**we can use Linear Regression models for predicting the target variable.**

**Business Insights:**

1. Weight plays an important factor for increasing the Insurance cost hence to

optimize the insurance cost one should look over the weight of an individual,

more the weight more will be the insurance cost.

2. The Insurance cost would increase if the regular\_checkup\_last\_year decreases

which means that company should focus on regular check-up of customer in order

to reduce the insurance cost.

3. Insurance cost would also increase if an individual insurance is

covered\_by\_any\_other\_company.

4. As the weight\_change\_in\_last\_one\_year of an individual changes then that

would give rise to increase in insurance cost.

**Recommendations:**

1. Based on the analysis, Linear Regression is the most suited for modelling the given

dataset as it has better explanatory power. Also, Significant features contributing to

higher/lower cost of insurance should have been identified from linear regression

2. To estimate insurance with maximum accuracy the company should consider

below parameters.

* regular\_checkup\_last\_year
* adventure\_sports
* heart\_decs\_history
* heart\_decs\_history
* covered\_by\_any\_other\_company
* exercise
* weight\_change\_in\_last\_one\_year

-------------------------------------------------\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*-------------------------------------------------