GREAT LEARNING

BUSINESS REPORT

CAPSTONE PROJECT - HEALTH CARE

BALAJI G

BATCH : JAN 2021 – B

EMAIL : balaji.gdp@gmail.com

LIST OF TABLE

LIST OF FIGURES

1) INTRODUCTION :

PROBLEM STATEMENT

Here we are going to deal with insurance policy and analyze the optimum insurance cost for an individual based on various criteria's with respect to health and habit related parameters of an individual. The medical insurance companies looks to optimize the insurance cost, so that it becomes affordable for every individual to opt for insurance policy

NEED OF THE STUDY/PROJECT

We all know that healthcare is most important domain and insurance policy is essential for every individual since the medical cost incurred during uncertainties are very high. Acquiring insurance policy is the best way to save an individual from high medical bills. So the company wants to optimize the insurance cost to make affordable premium for insurance policy, which will also attracts public to invest on it

UNDERSTANDING BUSINESS/SOCIAL OPPORTUNITY

The insurance policy is nothing but the sum of amount collected as premium from an individual is accumulated and invested in the market to produce income. The accumulated premium and profit through investing in the market are used for distributing claims to treat individuals during uncertain medical situation, it's nothing but sharing the risk. So optimizing the insurance cost to make affordable insurance policy premium which attracts the public to opt for it. Medical insurance is calculated based on health and habit related parameters of an individual. There are various tools used to calculated the premium in which machine learning becomes handy and efficient in calculating the insurance cost which also reduces the resource and human effort and improves the profitability.

DATA DICTIONARY :

S.NO	FIELD NAME	DESCRIPTION
1	APPLICANT_ID	APPLICANT UNIQUE ID
2	YEARS_OF_INSURANCE_WITH_U S	SINCE HOW MANY YEARS CUSTOMER IS TAKING POLICY FROM THE SAME COMPANY ONLY
3	REGULAR_CHECKUP_LASY_YEAR	NUMBER OF TIMES CUSTOMERS HAS DONE THE REGULAR HEALTH CHECK UP IN LAST ONE YEAR
4	ADVENTURE_SPORTS	CUSTOMER IS INVOLVED WITH ADVENTURE SPORTS LIKE CLIMBING, DIVING ETC.
5	OCCUPATION	OCCUPATION OF THE CUSTOMER
6	VISITED_DOCTOR_LAST_1_YEAR	NUMBER OF TIMES CUSTOMER HAS VISITED DOCTOR IN LAST ONE YEAR
7	CHOLESTEROL_LEVEL	CHOLESTEROL LEVEL OF THE CUSTOMERS WHILE APPLYING FOR INSURANCE
8	DAILY_AVG_STEPS	AVERAGE DAILY STEPS WALKED BY CUSTOMERS
9	AGE	AGE OF THE CUSTOMER
10	HEART_DECS_HISTORY	ANY PAST HEART DISEASES
11	OTHER_MAJOR_DECS_HISTORY	ANY PAST MAJOR DISEASES APART FROM HEART LIKE ANY OPERATION

12	GENDER	GENDER OF THE CUSTOMER
13	AVG_GLUCOSE_LEVEL	AVERAGE GLUCOSE LEVEL OF THE CUSTOMER WHILE APPLYING THE INSURANCE
14	BMI	BMI OF THE CUSTOMER WHILE APPLYING THE INSURANCE
15	SMOKING_STATUS	SMOKING STATUS OF THE CUSTOMER
16	YEAR_LAST_ADMITTED	WHEN CUSTOMER HAVE BEEN ADMITTED IN THE HOSPITAL LAST TIME
17	LOCATION	LOCATION OF THE HOSPITAL
18	WEIGHT	WEIGHT OF THE CUSTOMER
19	COVERED_BY_ANY_OTHER_CO MPANY	CUSTOMER IS COVERED FROM ANY OTHER INSURANCE COMPANY
20	ALCOHOL	ALCOHOL CONSUMPTION STATUS OF THE CUSTOMER
21	EXERCISE	REGULAR EXERCISE STATUS OF THE CUSTOMER
22	WEIGHT_CHANGE_IN_LAST_ONE_YEAR	HOW MUCH VARIATION HAS BEEN SEEN IN THE WEIGHT OF THE CUSTOMER IN LAST YEAR
23	FAT_PERCENTAGE	FAT PERCENTAGE OF THE CUSTOMER WHILE APPLYING THE INSURANCE
24	INSURANCE_COST	TOTAL INSURANCE COST

2) EDA AND BUSINESS IMPLICATION :

- The company has the data of customers up to 8 years with respect to years of insurance that the customer holds insurance policy with the company.
- The data was collected from the various cities of India includes Ahmedabad, Bangalore, Bhubaneswar, Chennai, Delhi, Guwahati, Jaipur, Kanpur, Kolkata, Lucknow, Mangalore, Mumbai, Nagpur, Pune, Surat.
- The company conducted the survey using their regular questionnaire and collected data with various parameters required to arrive the insurance cost
- The dataset has 25000 rows and 24 columns
- There are no duplicates in the data frame
- The descriptive details of dataset
- There are outlier present in the dataset. The attributes with outliers are regular_check_up_last_year, adventure_sports, visited_doctor_last_1_year, daily_avg_steps, heart_disease_history, other_major_disease_history and bmi
- There are totally 12871 null values in the data frame
- The data type of various attributes are 14 no's of integer, 8 no's of object and 2 no's of float data types
- Insurance cost is the target variable
- The original name of attribute is retained to understand the variable, since it explains the field names. but there are some spelling mistakes, which is corrected by renaming it
- The attributes which are rename are regular_checkup_lasy_year to regular_checkup_last_year, heart_decs_history to heart_disease_history, other_major_decs_history' to other_major_disease_history

UNIVARIATE ANALYSIS

There are 24 attributes in which 16 are numerical and 8 are categorical variables. Pandas profiling is used to explore the data.

i. years_of_insurance_with_us :

Year of insurance with us is the number of years the customer holds insurance with the company

- ➢ Years of insurance ranges from 0 to 8 years
- ➢ The mean is 4.08904, median is 4 and standard deviation is 2.6066
- ➢ The 25% percentile Q1 is 2 and 75% percentile Q3 is 6
- ➢ Though the skewness is not clearly visible in plot, since the value is negative the distribution is left skewed

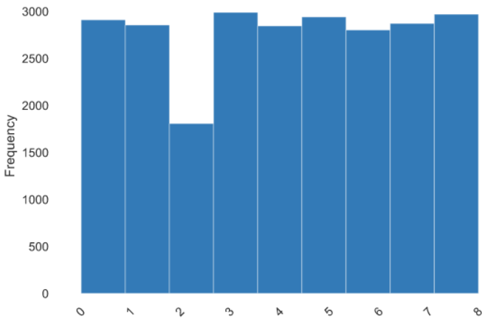

FIGURE 1 : HISTOGRAM - YEARS_OF_INSURANCE_WITH_US

ii. regular_checkup_last_year :

Regular check up last year is the number of times the customer undergone regular health check up in last year

- ➢ Regular checkup last year ranges from 0 to 5 times
- ➢ The mean is 0.77368, median is 0 and standard deviation is 1.19944
- ➢ The 25% percentile Q1 is 0 and 75% percentile Q3 is 1
- ➢ From plot we could interpret the plot is right skewed, also the skewness is positive

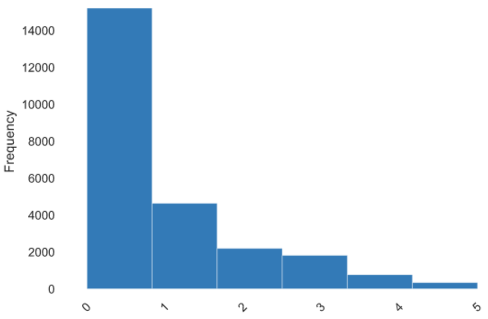

FIGURE 2 : HISTOGRAM - REGULAR_CHECKUP_LAST_YEAR

iii. visited_doctor_last_1_year :

Visited doctor last 1 year means the customer visited doctor for any illness

- ➢ Visited doctor last 1 year ranges from 0 to 12 times
- ➢ The mean is 3.1042, median is 3 and standard deviation is 1.14166
- ➢ The 25% percentile Q1 is 0 and 75% percentile Q3 is 1
- ➢ From plot we could interpret the plot is right skewed, also the skewness is positive

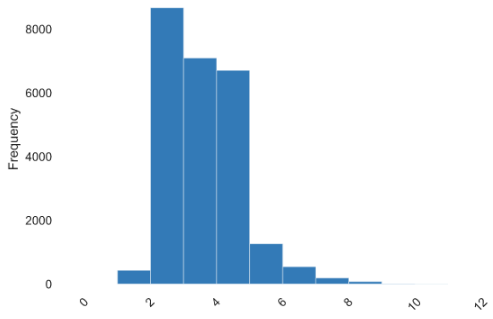

FIGURE 3 : HISTOGRAM - VISITED DOCTOR LAST 1 YEAR

iv. daily_avg_steps :

Daily average steps is nothing but the average steps that the customer walks everyday

- ➢ Daily average steps ranges from 2034 to 11255 steps
- ➢ The mean is 5215.88932, median is 5089 and standard deviation is 1053.1797
- ➢ The 25% percentile Q1 is 4543 and 75% percentile Q3 is 5730
- ➢ From plot we could interpret the plot is right skewed, also the skewness is positive

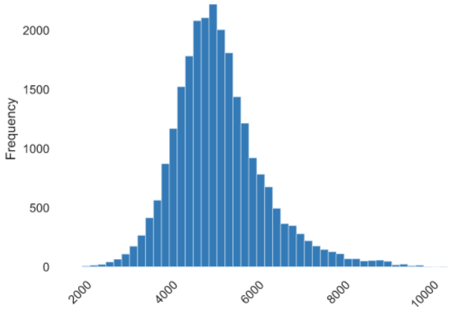

FIGURE 4 : HISTOGRAM - DAILY AVERAGE STEPS

v. age :

Age of the customer

- Age ranges from 16 to 74 years
- The mean is 44.91832, median is 45 and standard deviation is 16.1074
- The 25% percentile Q1 is 31 and 75% percentile Q3 is 59
- Though the skewness is not clearly visible in plot, since the value is positive the distribution is right skewed

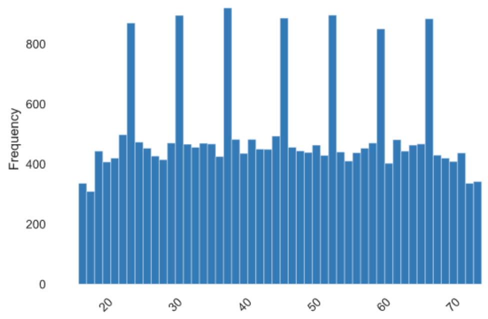

FIGURE 5 : HISTOGRAM - AGE

vi. avg_glucose_level :

Avg glucose level is one of the parameter which leads to various diseases such as heart disease, vision loss and kidney disease

- Average glucose level ranges from 57 to 277 mmol/ltr
- The mean is 167.53, median is 168 and standard deviation is 62.7297
- The 25% percentile Q1 is 113 and 75% percentile Q3 is 222
- Though the skewness is not clearly visible in plot, since the value is negative the distribution is left skewed

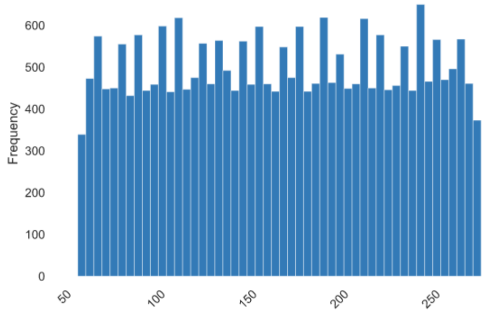

FIGURE 6 : HISTOGRAM - AVG GLUCOSE LEVEL

vii. bmi :

BMI of the customer

- BMI ranges from 12.3 to 100.6 times
- The mean is 31.3933, median is 30.5 and standard deviation is 7.8765
- The 25% percentile Q1 is 26.1 and 75% percentile Q3 is 35.6
- From plot we could interpret the plot is right skewed, also the skewness is positive

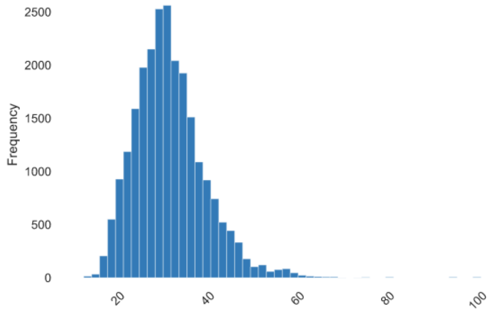

FIGURE 7 : HISTOGRAM - BMI

viii. Year last admitted :

Year last admitted is the detail pertaining to history of customer hospitalization for treatment

- Years last admitted ranges from 1990 to 2018
- The mean is 2003.892, median is 2004 and standard deviation is 7.5815
- The 25% percentile Q1 is 1997 and 75% percentile Q3 is 2010
- Though the skewness is not clearly visible in plot, since the value is positive the distribution is right skewed

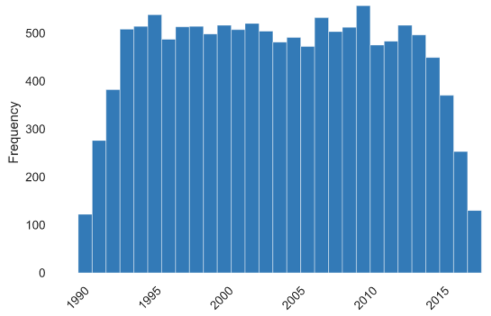

FIGURE 8 : HISTOGRAM - YEAR LAST ADMITTED

ix. Weight :

Weight of the customer

- Weight ranges from 52 to 96 times
- The mean is 71.61048, median is 72 and standard deviation is 9.32518
- The 25% percentile Q1 is 64 and 75% percentile Q3 is 78
- From plot we could interpret the plot is right skewed, also the skewness is positive

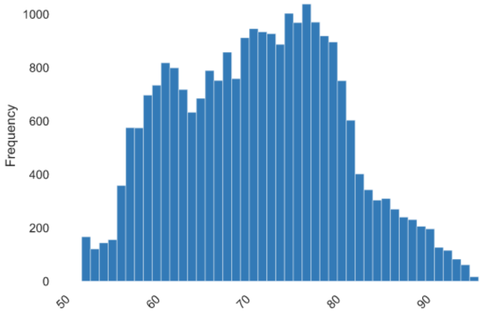

FIGURE 9 : HISTOGRAM - WEIGHT

x. Weight change for last one year :

Customer weight change in last one year

- Weight change in last one year ranges from 0 to 6 times
- The mean is 2.51796, median is 3 and standard deviation is 1.69033
- The 25% percentile Q1 is 1 and 75% percentile Q3 is 4
- From plot we could interpret the plot is right skewed, also the skewness is positive

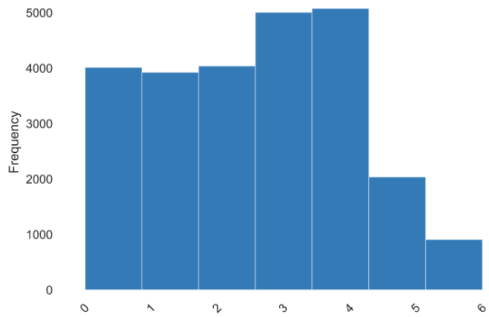

FIGURE 10 : HISTOGRAM - WEIGHT CHANGE IN LAST ONE YEAR

xi. fat_percentage :

Fat percentage is the amount of fat in customer body, the insurance company requires fat percentage to know the health condition of customer because high fat prone to diabetes, heart disease, stroke, artery disease.

- Fat percentage ranges from 11 to 42 times
- The mean is 28.81228, median is 31 and standard deviation is 8.63238
- The 25% percentile Q1 is 21 and 75% percentile Q3 is 36
- From plot we could interpret the plot is left skewed, also the skewness is negative

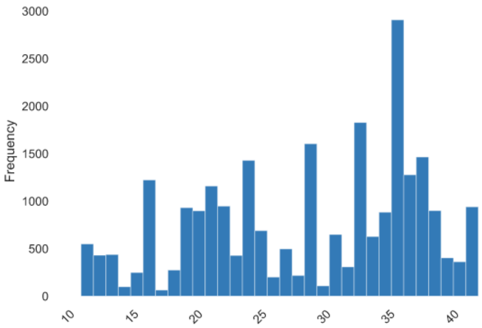

FIGURE 11 : HISTOGRAM - FAT PERCENTAGE

xii. insurance_cost :

Insurance cost is the premium amount of insurance

- Insurance cost ranges from 2468 to 67870 times
- The mean is 27147.40768, median is 27148 and standard deviation is 14323.69183
- The 25% percentile Q1 is 16042 and 75% percentile Q3 is 37020
- From plot we could interpret the plot is right skewed, also the skewness is positive

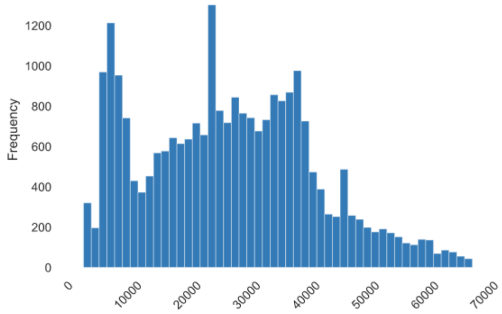

FIGURE 12 : HISTOGRAM - INSURANCE COST

xiii. adventure sports :

Adventure sports is the detail pertaining to the customer involves in any sports activity

- ➢ Adventure sports is categorical variable with 2 distinct value 0 and 1
- ➢ There are 91.83% customers without any sport activity and 8.17% customers involve in some sport activity
- ➢ There are 22957 customers without any sport activity and 2043 customers involve in some sport activity

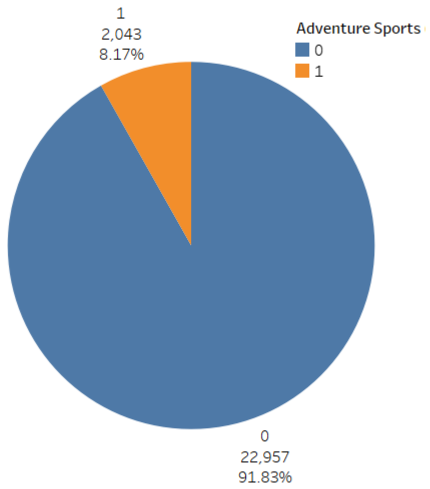

FIGURE 13 : PIE CHART- ADVENTURE SPORTS

xiv. Occupation :

Occupation of customer

- ➢ Occupation is categorical variable with 3 distinct category student, business and salaried
- ➢ There are 40.68% of students, 40.08% customers are into business and 19.24% customers are salaried
- ➢ There are 10169 of students, 10020 customers are into business and 4811 customers are salaried

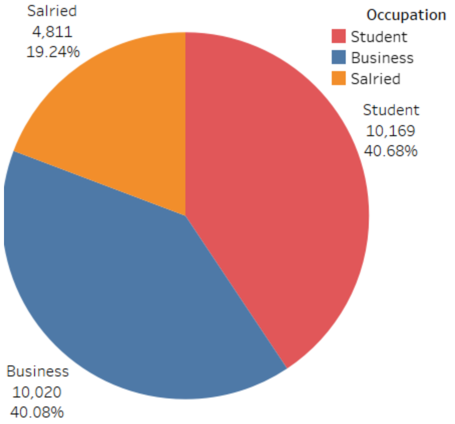

FIGURE 14 : PIE CHART- OCCUPATION

xv. cholesterol_level :

Cholesterol level of customer is one of the criteria that insurance companies wants to know about customer health condition. Since high cholesterol leads to heart disease

- ➢ Cholesterol level is categorical variable with 5 distinct ranges 125 to 150, 150 to 175, 175 to 200, 200 to 225 and 225 to 250
- ➢ The percentage of customers with cholesterol level 125 to 150 is 33.36%, 150 to 175 is 35.05%, 175 to 200 is 11.52%, 200 to 225 is 11.85% and 225 to 250 is 8.22%
- ➢ The total number of customers with cholesterol level 125 to 150 is 8339, 150 to 175 is 8763, 175 to 200 is 2881, 200 to 225 is 2963 and 225 to 250 is 2054

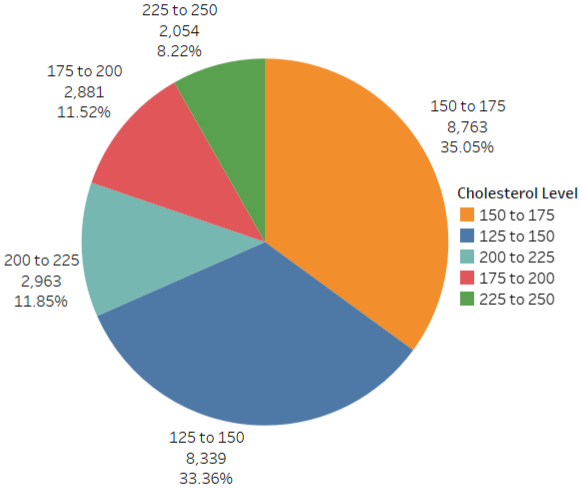

FIGURE 15 : PIE CHART - CHOLESTEROL LEVEL

xvi. heart_disease_history :

History of customer with heart disease or not

- Heart disease history is categorical variable with 2 distinct values 0 and 1
- There are 94.54% customers has no history of any heart disease and 5.46% customers with had heart disease
- There are 23634 customers has no history of any heart disease and 1366 customers with had heart disease

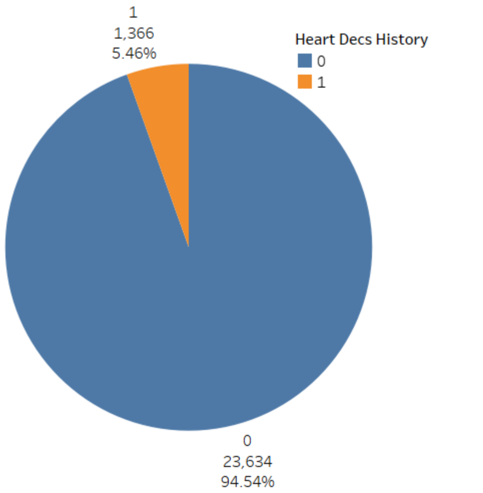

FIGURE 16 : PIE CHART - HEART DISEASE HISTORY

xvii. other_major_disease_history :

History of customer with any other major disease

- Other major disease history is categorical variable with 2 distinct values 0 and 1
- There are 90.18% customers has no history of any other disease and 9.82% customers with history of any other disease
- There are 22546 customers has no history of any other disease and 2454 customers with history of any other disease

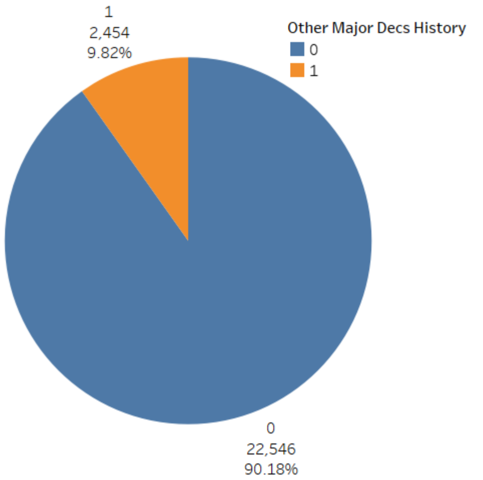

FIGURE 17 : PIE CHART - OTHER MAJOR DISEASE HISTORY

xviii. Gender :

Gender of the customer

- Gender of the customer is categorical variable with 2 distinct values male and female
- There are 65.69% of customers are male and 34.31% of customers are female
- There are 16422 of customers are male and 8578 of customers are female

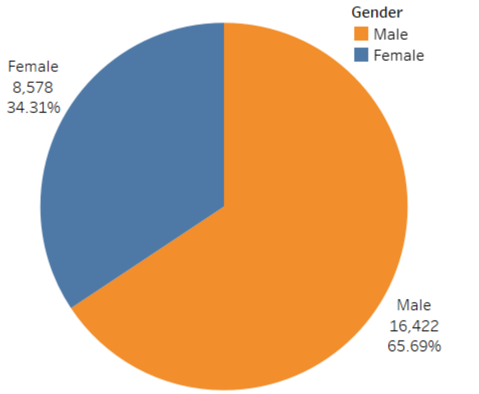

FIGURE 18 : PIE CHART - GENDER

xix. smoking_status :

Smoking status of the customer

- Smoking status of the customer is categorical variable with 4 distinct category never smoked, unknown, formerly smoked and smokes
- There are 37% of customers never smoked, 30.22% customers status is unknown, 17.32% customers formerly smoked and 15.46% of customers smokes
- There are 9249 of customers never smoked, 7555 customers status is unknown, 4329 customers formerly smoked and 3867 of customers smokes

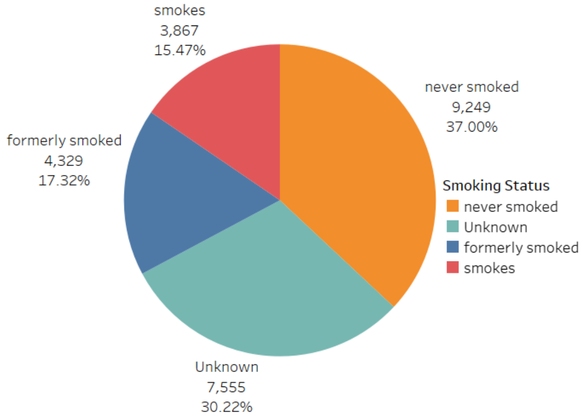

FIGURE 19 : PIE CHART - SMOKING STATUS

xx. covered_by_any_other_company :

This is pertaining to customer who had insurance with other company previously

- Covered by any other company is a Boolean variable
- There are 69.67% of customers were not covered by other company insurance policy and 30.33% of customers were covered by other company insurance policy
- There are 17418 of customers were not covered by other company insurance policy and 7582 of customers were covered by other company insurance policy

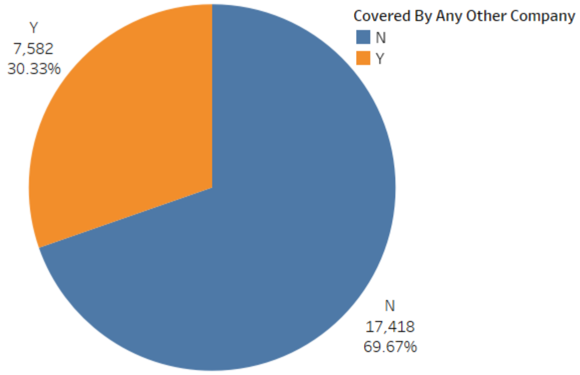

FIGURE 20 : PIE CHART - COVERED BY ANY OTHER COMPANY

xxi. Alcohol :

This is pertaining to alcohol habit of customer

- Alcohol is a categorical variable with 3 distinct category rare, no and daily
- The frequency alcohol habit by a customer is 55.01% of customers rarely drinks, 34.16% of customers doesn't drink and 10.83% of customers drinks daily
- The frequency alcohol habit by a customer is 13752 of customers rarely drinks, 8541 of customers doesn't drink and 2707 of customers drinks daily

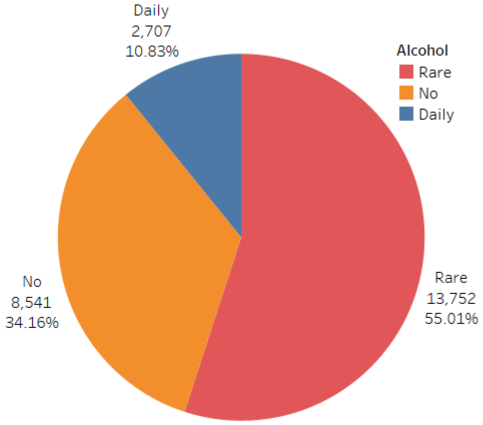

FIGURE 21 : PIE CHART - ALCOHOL

xxii. exercise :

This is pertaining to customer will exercise or not

- Exercise is a categorical variable with 3 distinct category moderate, extreme and no
- The customer with exercise habit are 58.6% of customers exercise moderately, 21% of customers exercise extremely and 20.5% of customers doesn't exercise
- The customer with exercise habit are 14638 of customers exercise moderately, 5248 of customers exercise extremely and 5114 of customers doesn't exercise

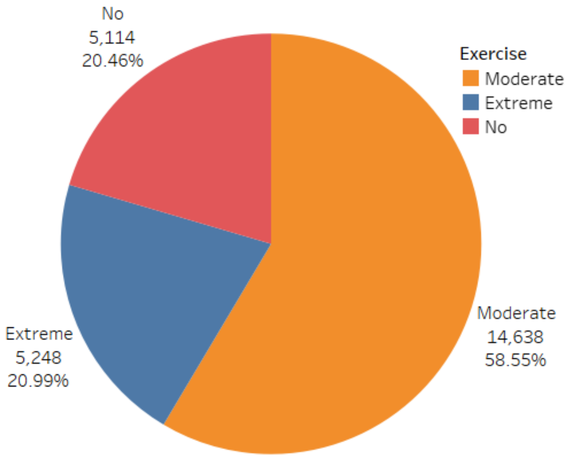

FIGURE 22 : PIE CHART - EXERCISE

xxiii. Location :

Location of insurance policy opted city wise

- The data available for customers located in 15 cities of India includes Ahmedabad, Bangalore, Bhubaneswar, Chennai, Delhi, Guwahati, Jaipur, Kanpur, Kolkata, Lucknow, Mangalore, Mumbai, Nagpur, Pune, Surat

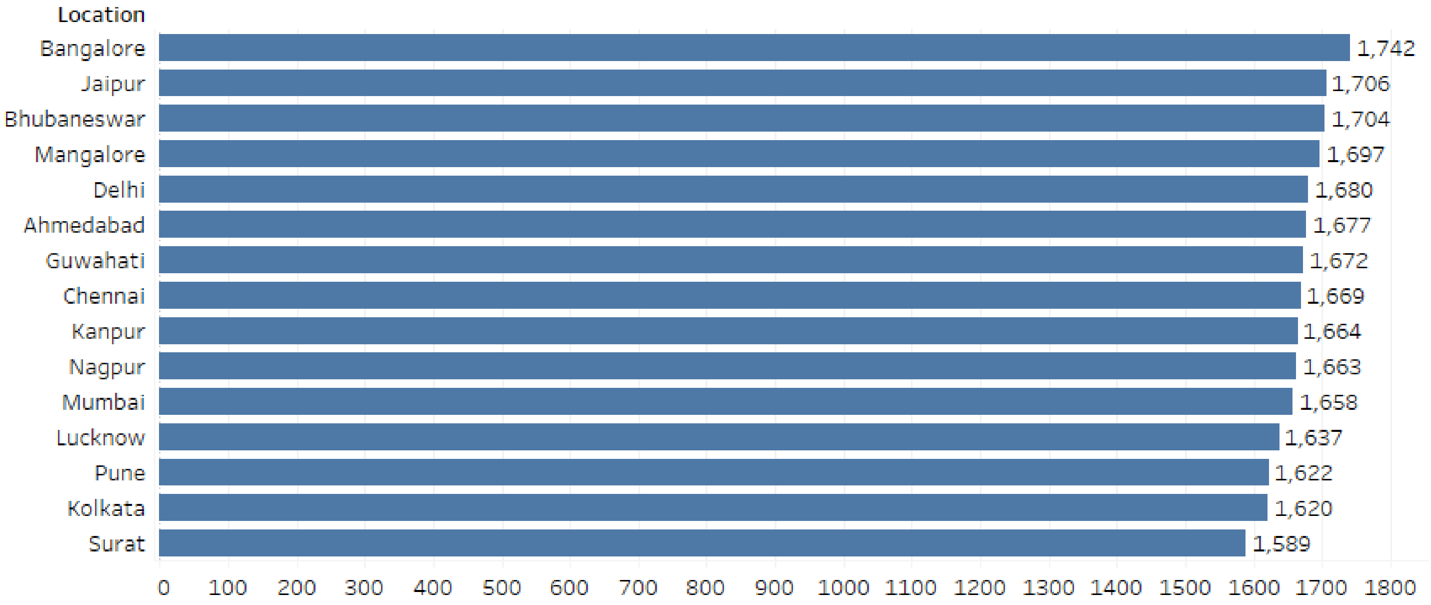

FIGURE 23 : BAR CHART - LOCATION

BIVARIATE ANALYSIS (RELATIONSHIP BETWEEN DIFFERENT VARIABLES , CORRELATIONS)

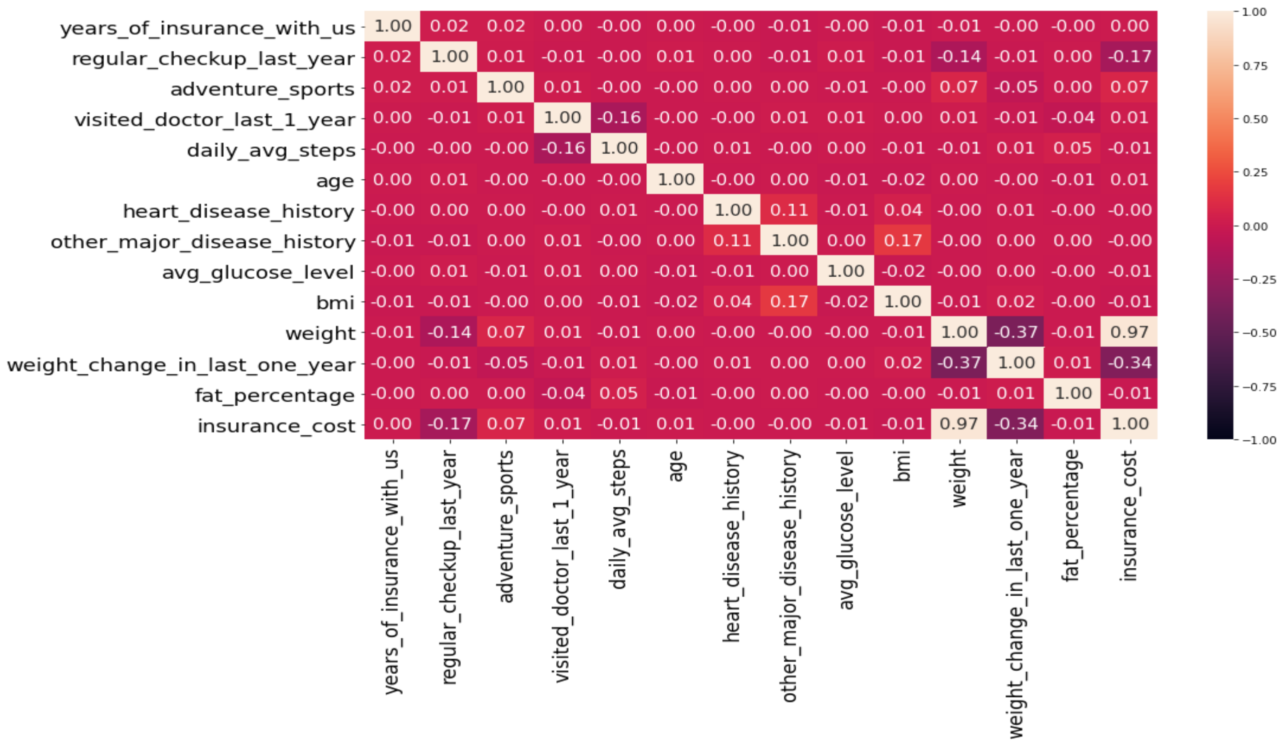

FIGURE 24 : HEATPLOT - RELATIONSHIP BETWEEN VARIABLES

- From above heat plot, we could infer that there is only one independent variable is correlated with target variable.
- The correlation between weight and insurance cost is 97%

i. Weight vs Insurance Cost :

- From above heat plot, we could infer that there is there is linear correlation between weight and insurance cost
- The correlation between weight and insurance cost is 97%

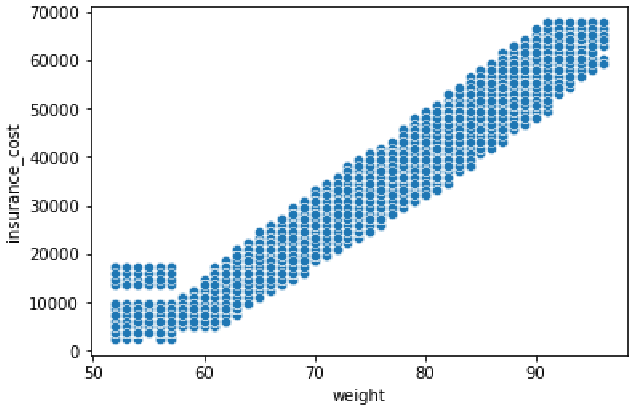

FIGURE 25 : SCATTER PLOT - WEIGHT VS INSURANCE COST

BUSINESS IMPACT FROM EDA

From EDA we could infer the variables which will impact the business

1) year of insurance with company defines the customer relationship with particular company
2) Regular check up last year shows the history of regular medical checkups carried by the customers, this shows the present health condition of a customer
3) Visited doctor last 1 year defines the history of diseases in last 1 year of customer
4) Exercise and Daily average steps is the habitual activities of customer to maintain the good health condition
5) The age is one the major factor where the claim frequency is determined, based on which the insurance cost varies for different range age group. The age range of age group is classified into 0 - 18 years, 19 - 35 years, 36 - 45 years, 46 - 50 years, 50-55 years, 56 - 60 years, 61 - 65 years, 65 - 70 years and 71+. The insurance cost is different for different age group
6) Avg glucose level is one of the major parameter for calculating the insurance cost because the customer with high glucose level are prone to heart disease, kidney disease, eye problems etc. , which will impact the business
7) BMI is another important factor is proportional to weight of the customer. The ideal range of BMI is 18.5 - 24.9. The customer with BMI < 18.5 and BMI > 24.9 are considered as underweight and obese. There are around 76.9% of customers are overweight and 5.9% are underweight. The over weighted customers are prone to diabetes, heart disease, stroke, some type of cancer etc. There is risk in business and BMI is directly proportional to weight. Already weight is most significant variable towards insurance cost
8) Year last admitted shows the history of customer got hospitalized for any major treatments
9) Weight is most important and highly significant in calculating the insurance cost, because the consequence of obesity are diabetes, heart disease, stroke, some type of cancer etc., which will impact the business
10) Cholesterol level and Fat percentage are nothing but the percentage of fat in body, where high fat may results in hypertension, diabetes and gallbladder diseases, which will impact the business. The ideal range of Cholesterol is 125 - 200 mg/dl. There are around 8.25% customers with Cholesterol level > 200, so the business risk is lesser with respect to cholesterol level
11) Adventure sports is the customer who involves in adventure sports, they are most vulnerable towards accidents which is impact the business
12) Heart disease history and other major disease are another important variables where the customers are high risk towards disease and hospitalize, which will impact the business
13) Smoking status is defines the smoking habit of customer. The smokers are vulnerable to cancer, heart disease, stroke, lung diseases, diabetes, and chronic obstructive pulmonary disease, which will impact the business
14) Alcohol habit of customer prone to stomach ailments, brain damage, serious memory loss and liver cirrhosis, which will impact the business
15) Location is another important variable where the hospital expenses are different in different zones, so the insurance cost will be different for different cities. The location of customer is another important variable because the insurance cost varies for different cities. It is segregated into 3 zones based on the medical expense, the medical expense in Zone3 > Zone2 >Zone1

3) DATA CLEANING AND PREPROCESSING :

IDENTIFYING MISSING VALUE AND OUTLIER IN THE DATASET

There are missing values in the dataset. Find the following list of missing values before outlier treatment

- BMI has 990 missing values which is 4% of total values
- Year last admitted has 11881 missing data which is 47.5% of total values

There are outlier present in the dataset. The variables with outliers are

- Outliers are calculated with 25th percentile Q1 and 75th percentile Q3
- Then calculate Inter Quartile Range(IQR), which is difference between Q3 and Q1
- Then calculate upper limit, UL
  UL = Q3 + 1.5 * IQR
- Then calculate lower limit, LL
  LL = Q1 - 1.5 * IQR
- The value above upper limit and less than lower limit are identified as outliers, following are the variables with outliers

a) adventure_sports : Adventure sports is a Boolean value 0 and 1. The outlier treatment is not required

b) bmi : There are some extreme values in BMI, which is nothing but anomaly because those are abnormal observations and it has to be treated

c) daily_avg_steps : There are some extreme values in average steps walked by a customer and it has to be treated

d) heart_desc_history : History of heart disease is a Boolean value 0 and 1. The outlier treatment is not required

e) other_major_desc_history : History of other major disease is a Boolean value 0 and 1. The outlier treatment is not required

f) regular_check_last_year : No. of times the customer have done regular medical checkup in last 1 year is a continuous variable. The outlier treatment is not mandatory

g) visiting_doctor_last_1_year : No. of times the customer have visiting doctor in last 1 year is a continuous variable. The outlier treatment is not mandatory

MISSING VALUE AND OUTLIER TREATMENT

The missing values and outliers are treated by replacing the outlier with Nan and then Nan is imputed using KNN imputer. The steps involve in treating outliers and missing values are

- The identified outliers are replaced with NaN values
- Checking for missing value after replacing outliers with NaN value
- BMI has 1539 missing values. Earlier it was 990 which is increased to 1539 missing values
- Daily average steps has 952 missing values after replacing outliers with NaN value
- Transforming the categorical variable into numerical variable using manual encoding and one hot encoding

- Then scaling the transformed data, since we are using KNN imputer to treat the missing values.
- KNN imputer is used to impute the missing value by considering 10 nearest neighbors
- The NaN values are imputed using KNN imputer

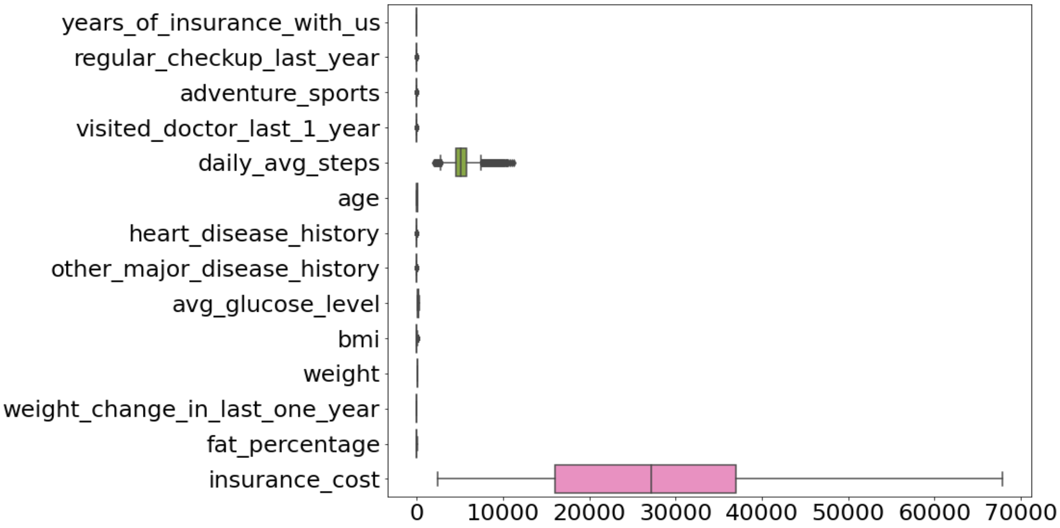

FIGURE 26 : BOXPLOT BEFORE AND AFTER OUTLIER TREATMENT

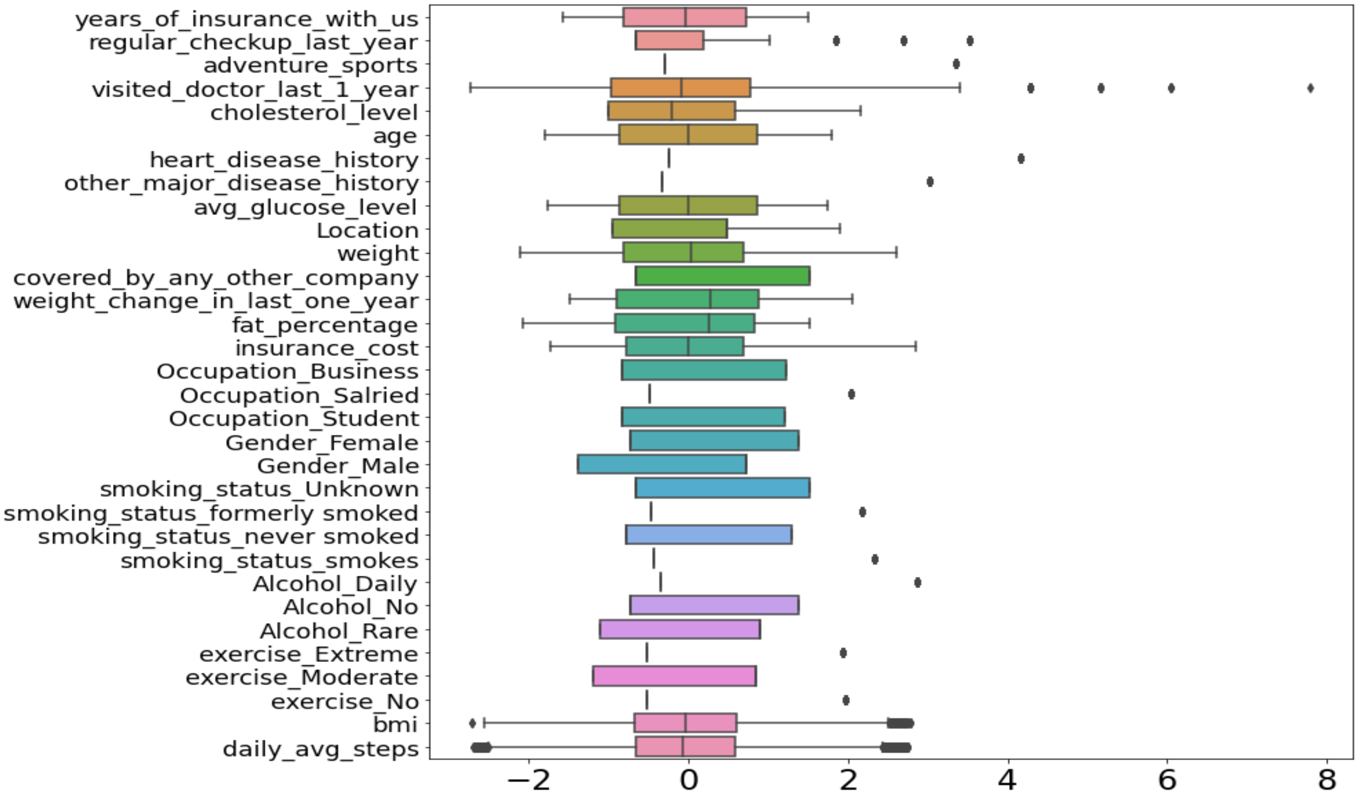

FIGURE 27: BOX PLOT AFTER OUTLIER AND NULL VALUE TREATMENT

There are 8 object type variables. All these object type variables has to be transformed to numeric variable. The transformed variables are Cholesterol_level, location, covered_by_any_other_company, occupation, gender, smoking_status, alcohol and exercise

The manual encoding peformed on Cholesterol_level, location and covered_by_any_other_company

i. Encoding cholesterol level :

The various range of cholesterol level is encoded into numeric value

Cholesterol level	**Encoded Cholesterol level**
125 to 150	125
150 to 175	150
175 to 200	175
200 to 225	200
225 to 250	225

TABLE 1 : ENCODED CHOLESTEROL VALUE

ii. Encoding location :

The location is encoded into zones, since the location where the customer reside plays an important to determine insurance cost. Based on treatment cost incurred, the cities are segregated into zones

Location	**Encoded Location into Zones**
Delhi and Mumbai	3
Bangalore, Chennai, Kolkata, Ahmedabad, Pune and Surat	2
Jaipur, Bhubaneswar, Mangalore, Guwahati, Kanpur, Nagpur and Lucknow	1

TABLE 2: ENCODED LOCATION INTO ZONES

iii. Encoding covered by any other company :

The insurance covered by any other company is encoded into 0 and 1

Covered by any other company	**Encoded Covered by any other company**
N	0
Y	1

TABLE 3: ENCODED COVERED BY ANY OTHER COMPANY

iv. One hot encoding on remaining categorical variable :

Occupation, gender, smoking_status, alcohol and exercise variables are encoding using one hot encoding

REMOVAL OF UNWANTED VARIABLES

The variables which doesn't contribute for prediction shall be considered as unwanted variables. This dataset has some unwanted variables which is removed from the dataset, they are

i. Applicant ID is a unique variable used to identify the customer
ii. Year last admitted is one of the criteria for insurance company wants to know whether the customer is hospitalized for illness. But year last admitted has 11881 missing data which is 47.5% of total values. So we drop year last admitted since there is lot of missing values
iii. Performed VIF on final processed dataset and check for significant variables to build model. The variables with VIF factor < 5 are highly significant and they are

SIGNIFICANT VARIABLE	VIF FACTOR
years_of_insurance_with_us	1.1
regular_checkup_last_year	1.0
adventure_sports	1.0
visited_doctor_last_1_year	1.0
cholesterol_level	1.4
age	1.0
heart_disease_history	1.0
other_major_disease_history	1.1
avg_glucose_level	1.0
Location	1.0
weight	1.2
covered_by_any_other_company	1.1
weight_change_in_last_one_year	1.2
fat_percentage	1.1
bmi	1.2
daily_avg_steps	1.1

TABLE 4 : SIGNIFICANT VARIABLE AFTER VIF

iv. Splitting Data Into Train And Test Dataset after VIF With respect to VIF the most significant variables are identified and considered as independent variables and Insurance cost as target variable. Both independent and target variables are split into train and test dataset at 70 : 30 ratio at random state '0'. The shape train and test dataset are
    i. x_train - 17500 rows and 15 columns
    ii. y_train - 17500 rows and a column
    iii. x_test - 7500 rows and 15 columns
    iv. y_test - 7500 rows and a column

v. Performed Hypothesis testing using Stats model Linear Regression and identify the most significant variable with p<0.05. The most significant independent variables after eliminating the insignificant variables are
    i. regular_checkup_last_year
    ii. adventure_sports

iii. visited_doctor_last_1_year
iv. heart_disease_history
v. weight
vi. covered_by_any_other_company
vii. weight_change_in_last_one_year

ADDITION OF NEW VARIABLES

As a result of one hot encoding the new variables got added. Since the variables are straight forward abides the basic requirement of insurance company to arrive insurance cost the feature engineering is not necessary for this dataset.

The insurance premium amount has huge difference between the customer, the details pertaining to individual / family cover and sum of amount insured are required. These details shall be added if we have the company brochure with premium details

4) MODEL BUILDING:

The target variable is insurance_cost and it is continuous , we are going to predict the insurance cost using various regressor and evaluate using R^2 value, adjusted R^2 and RMSE value. The following regressors are used to build models

a) CART model
b) Random forest model
c) Linear Regression using sklearn model
d) Artificial Neural Network
e) Ada boosting
f) XG boosting
g) Gradient Boosting
h) Bagging
i) Voting regressor

SPLITTING DATA INTO TRAIN AND TEST DATASET WITH SIGNIFICANT VARIABLES

Now the dataset is split into train and test dataset using the most significant variables obtained after eliminating insigficant variables using stats model linear regression

Both independent and target variables are split into train and test dataset at 70 : 30 ratio at random state '0'. The shape of train and test dataset are

i. x_train - 17500 rows and 7 columns
ii. y_train - 17500 rows and a column
iii. x_test - 7500 rows and 7 columns
iv. y_test - 7500 rows and a column

a) CART Model :

Performing Decision tree regression to predict the insurance cost. The steps involved in modeling are

- Selecting the best parameters through grid search
- The parameters obtained through grid search are max_depth - 7, min_sample_leaf - 30, min_sample_split - 2 and splitter is best at random state - 0
- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value
- The features that contributes in predicting insurance cost are

weight	0.995670
covered_by_any_other_company	0.002338
regular_checkup_last_year	0.001498
weight_change_in_last_one_year	0.000438
visited_doctor_last_1_year	0.000055
adventure_sports	0.000000
heart_disease_history	0.000000

- From above feature importance we could infer that weight contributes more in predicting the insurance cost

b) Random Forest Model :

Performing Random Forest regression to predict the insurance cost. The steps involved in modeling are

- Selecting the best parameters through grid search
- The parameters obtained through grid search are n_estimators - 350, max_depth - 7, max_features - 7, min_sample_leaf - 1 and min_sample_split - 3 at random state - 0
- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value
- The features that contributes in predicting insurance cost are

weight	0.994660
covered_by_any_other_company	0.002410
regular_checkup_last_year	0.001743
weight_change_in_last_one_year	0.000696
visited_doctor_last_1_year	0.000344
adventure_sports	0.000089
heart_disease_history	0.000059

- From above feature importance we could infer that weight contributes more in predicting the insurance cost

c) Linear Regression Model :

Performing Linear regression to predict the insurance cost. The steps involved in modeling are

- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value
- As a result of linear regression model, the insurance cost is predicted using the linear equation

insurance_cost = -0.000538 + (-0.037 * regular_checkup_last_year) + (0.005 * adventure_sports) + (-0.004 * visited_doctor_last_1_year) + (0.004 * heart_disease_history) + (0.967 * weight) + (0.038 * covered_by_any_other_company) + (0.02 * weight_change_in_last_one_year)

- From above equation we could infer that the coefficient of weight contributes more in predicting the insurance cost and regular check up last year and visited doctor last 1 year has negative impact on prediction but it is negligible

d) Neural Network Model :

Performing Neural network model to predict the insurance cost. The steps involved in modeling are

- Selecting the best parameters through grid search
- The parameters obtained through grid search are hidden_layer_sizes - 100, solver - adam, max_iter - 200 and tol - 0.0001 at random state - 0
- Fit the training set into model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value

e) Ada Boosting

- Ada boosting is built using random forest as base_estimator
- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value
- The features that contributes in predicting insurance cost are

weight	0.990456
covered_by_any_other_company	0.003079
regular_checkup_last_year	0.002899
weight_change_in_last_one_year	0.001611
visited_doctor_last_1_year	0.001330
adventure_sports	0.000347
heart_disease_history	0.000279

f) XG Boosting

- XG boost is built at random state '0'
- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value

g) Gradient Boosting

- Gradient boost is built at random state '0'
- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value

h) Bagging model

- Bagging model is built using random forest as base_estimator
- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value

i) Voting Regression model

- Voting regressor model is built using all the models built above as base_estimators
- Fit the training set into model and validate the model and predict the insurance cost for both train and test dataset and validate the model using R^2, adjusted R^2 and RMSE value

5) MODEL VALIDATION:

All the models includes CART, Random Forest, Linear Regression, Neural Networks, Ada Boositng, XG Boosting, Gradient Boosting, Bagging and Voting Regressor are validates using R^2, Adjusted R^2 and RMSE value. Find the following comparison table of performance metrics

PERFORMANCE MATRICS	R SQUARE TRAIN	R SQUARE TEST	ADJ R SQUARE TRAIN	ADJ R SQUARE TEST	RMSE TRAIN	RMSE TEST
CART	0.9560	0.9546	0.9139	0.9112	0.2088	0.2153
RF	0.9563	0.9552	0.9145	0.9124	0.2081	0.2138
LR SKLEARN	0.9439	0.9463	0.8909	0.8953	0.2358	0.2342
NN	0.9533	0.9525	0.9087	0.9073	0.2153	0.2201
AB	0.9558	0.9543	0.9135	0.9107	0.2093	0.2159
XGB	0.9615	0.9529	0.9244	0.9079	0.1954	0.2193
GB	0.9557	0.9556	0.9133	0.9132	0.2095	0.2127
BAG	0.9564	0.9553	0.9147	0.9126	0.2079	0.2135
VOTE_REG	0.9564	0.9570	0.9146	0.9158	0.2080	0.2095

TABLE 5 : COMPARISION OF PERFORMANCE METRICS

From above table we could infer

- R^2 , adjusted R^2 and RMSE is good for all the models but we need the select the best model with highest score. The best value is highlighted in above table
- R^2 of random forest regressor is comparatively better than other models on both train and test dataset
- Adjusted R^2 of random forest regressor is comparatively better than other models on both train and test dataset
- RMSE of random forest regressor is comparatively better than other models on both train and test dataset
- So above comparison table we could infer that random forest model performs better than other models. Let us perform some boosting and bagging regressor
- R^2 , adjusted R^2 and RMSE is good for all the models but we need the select the best model with highest score. The best value is highlighted in above table
- R^2, adjusted R^2 and RMSE for train dataset is comparatively better for XG boost Regressor than other models
- R^2, adjusted R^2 and RMSE for test dataset is better for Voting Regressor compare to other models
- From above comparison we could infer that all model performs relatively close and there is no much difference between the model, but **VOTING REGRESSOR** which has better score compare to other models with least error and found to be best model

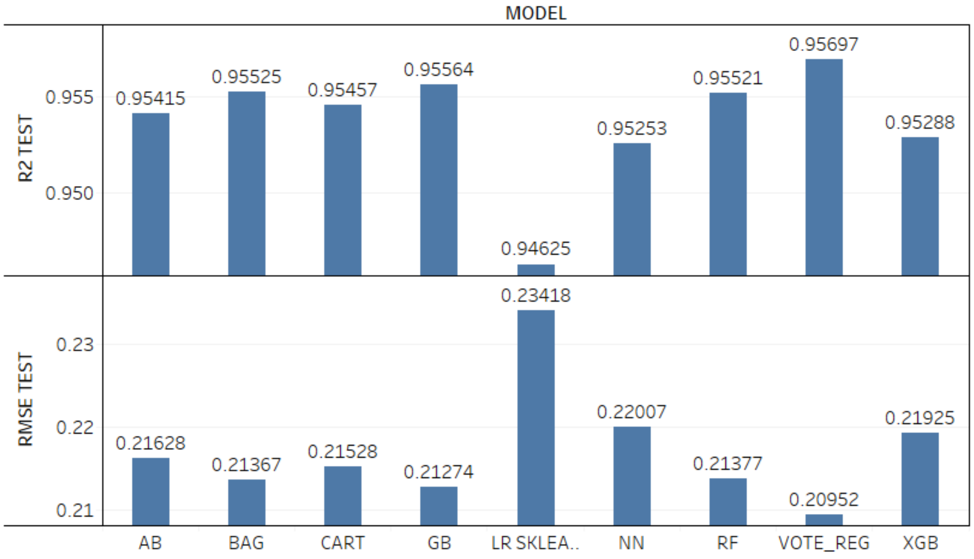

FIGURE 28 : PERFORMANCE COMPARISION BAR CHART OF ALL MODELS

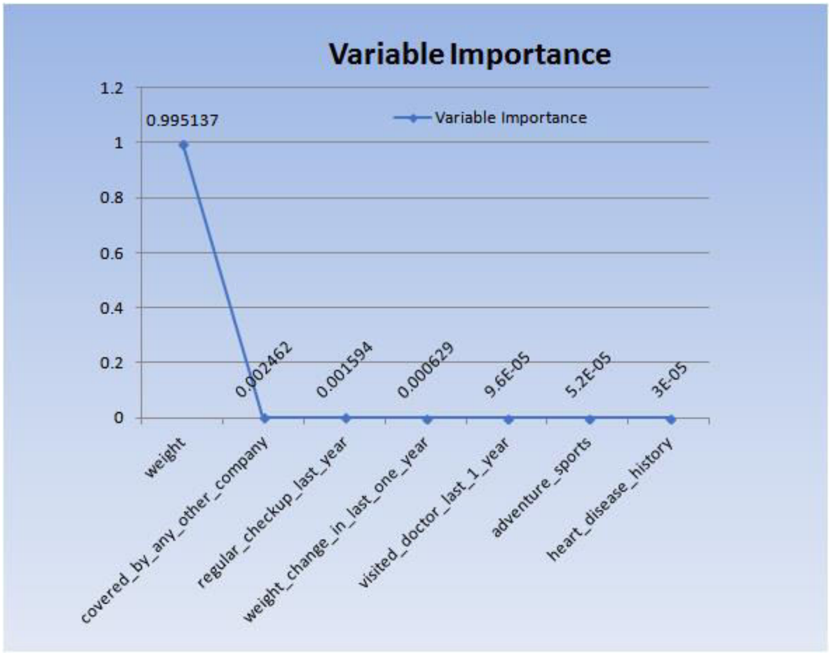

FIGURE 29 : VARIABLE IMPORTANCE

From above figure we could infer that weight is the most significant variable that impact the insurance cost and other variables has negligible impact on insurance cost.

Significant Variables	Variable Importance
weight	0.995137
covered_by_any_other_company	0.002462
regular_checkup_last_year	0.001594
weight_change_in_last_one_year	0.000629
visited_doctor_last_1_year	0.000096
adventure_sports	0.000052
heart_disease_history	0.00003

TABLE 6 : VARIABLE IMPORTANCE

6) FINAL INTERPRETATION / RECOMMENDATION:

INTERPRETATION

1) Weight is the most significant variable that contributes around 99.51% in predicting the insurance cost.
2) Covered by any other company detail will the help the customer to retain his existing terms and condition for waiting period and it will not be considered as new policy, so there is a risk for insurance company. But impact on insurance cost is around 0.0025%
3) The customer involve in adventurous sport have high possible for accidents. The impact of adventure sports on insurance cost is around 5.2 x e^{-05} %which is negligible
4) History of heart disease will have impact on insurance cost and there are separate policies are available for customers with heart disease. The impact of adventure sports on insurance cost is around 3 x e^{-05} %which is negligible
5) Regular check up last year have around 0.006% impact on insurance cost, which is negligible
6) Visited doctor last 1 year have around 9.6 x e^{-05} % impact on insurance cost, which is negligible
7) Weight change in last 1 year have around 0.0006% impact on insurance cost, which is negligible
8) Voting regressor have minimum RMSE value and found to be best model for predicting the insurance cost

RECOMMENDATION

1) The insurance cost should be high for the customer with overweight which has risk towards various diseases. Also change in customer weight has to be monitored and insurance cost has to be worked during renewal
2) The insurance cost should be charged more for customers in zone 3 followed by zone 2 and zone 1, since the medical expense incurred in zone 3 are high compare to other zones
3) The customers involve is adventurous sports should be charged high, since they prone to accident possible for frequent claims
4) The customer with heart disease history are considered to be high risk category and insurance cost should be charged high
5) The customer with history of other major disease are considered to be high risk category and insurance cost should be charged high
6) Insurance companies should frame the plans with add on coverage packages along with general insurance packages. The parameters like adventurous sports, heart disease history, other major disease history shall be considered as add on with general insurance policy. So that the risk on insurance company shared

From above analysis we could infer that VOTING REGRESSOR performs better than other models. Predicting insurance cost by abiding basic criteria's using ML helps the medical insurance companies to attract customers and save time in calculating the insurance cost for every individual. So insurance policy providers shall define the policy easily and ML make the cost calculation quickly compare to manual calculation by considering various criteria's, also handling and managing the data will become convenient.