**CAPSTONE PROJECT ON HEALTHCARE**

**BUSINESS REPORT**

**SULOCHANA**

|  |
| --- |
| **CONTENTS**  **TABLE OF CONTENTS………………………………………………………2**  **LIST OF FIGURES…………………………………………………………....4**  **LIST OF TABLES……………………………………………………………..6**   1. **Introduction…………………………………………………………….7** 2. **Business problem…………………………………………………………………………..7** 3. **Defining problem statement…………………………………………………………..7** 4. **Scope……………………………………………………………………………………………..7** 5. **Need of the study/project……………………………………………………………..7** 6. **Understanding business/social opportunity………………………………….7** 7. **EDA and Business Implication……………………………………..8** 8. **Understanding how data was collected in terms of time,**   **Frequency and methodology………………………………………………………………8**   1. **Visual inspection of data**   **(rows, columns, descriptive details)……………………………………………….9**   1. **Descriptive details…………………………………………………………………………10** 2. **Understanding of attributes**   **(variable info, renaming if required)………………………………………………11**   1. **Univariate analysis (distribution and spread for every continuous**   **Attribute, distribution of data in categories for**  **categorical ones)……………………………………………………………………………12**   1. **Bivariate analysis (relationship between**   **Different variables, relations)…....................................................24**   1. **Multivariate analysis……………………………………………………………….29** 2. **Business insights………………………………………………………………………….29** 3. **Data Cleaning and Pre-processing ………………………………….30** 4. **Removal of unwanted variables (if applicable)……………………………….30** 5. **Missing Value treatment (if applicable)…………………………………………..31** 6. **Outlier treatment (if required)………………………………………………………..32** 7. **Variable transformation (if applicable)…………………………………………….33** 8. **Model building…………………………………………………………34** 9. **Training and Testing Data……………………………………………………………….36** 10. **Model building……………………………………………………………………………….36**     * + **Linear Regression Model…………………………………………………..36**  Linear Regression using stats models………………………………..38Building Decision Tree Regression Basic modelWith all variables ………………………………………….42Building Random Forest Regressor basic modelwith all variables…………………………………………………………….43  * + - **Model Building with XGBoost Regressor basic model…….44**  1. **Model tuning measures……………………………………………………………….46**    * + **Create a regularized RIDGE model………………………………….46**  Create a regularized LASSO model………………………………….48Decision Tree Regression tuned model…………………………..49Random Forest Regressor tuned model………………………….50XGBoost Regressor model Tunining……………………………….52  1. **Model validation………………………………………………………53** 2. **Final interpretation / recommendation…………………………...55** |
| ***List of Figures***  ***Figure 1: Count Plot of Occupation variable………………………………12***  ***Figure 2: Count plot of cholesterol level variable………………………..13***  ***Figure 3: Count plot of Gender variable……………………………………13***  ***Figure 4: Count plot of smoking\_status variable………………………...14***  ***Figure 5: Count plot of Location variable………………………………….14***  ***Figure 6: Count plot of covered\_by\_any\_other\_company variable….15***  ***Figure 7: Count Plot of Alcohol variable……………………………………15***  ***Figure 8: Count plot of exercise variable…………………………………..16***  ***Figure 9: Years\_of\_insurance\_with\_us distplot and boxplot………….16***  ***Figure 10: regular\_checkup\_last\_year distplot and boxplot…………..17***  ***Figure 11: adventure\_sports observations distpolt and boxplot…….17***  ***Figure 12: Visited\_doctor\_last\_1\_year distplot and boxplot………….18***  ***Figure 13: Daily\_avg\_steps distplot and boxplot………………………..18***  ***Figure 14: Age observations distplot and boxplot………………………19***  ***Figure 15: Heart\_decs\_history distplot and boxplot……………………19***  ***Figure 16: Other\_major\_decs\_history distplot and boxplot…………..20***  ***Figure 17: Avg\_glucose\_level distplot and box plot……………………20***  ***Figure 18: bmi observations distplot and boxplot………………………21***  ***Figure 19: Year\_last\_admitted distplot and boxplot……………………21***  ***Figure 20: Weigh distplot and boxplot…………………………………….22***  ***Figure 21: Weight\_change\_in\_last\_one\_year distplot and boxplot…22***  ***Figure 22: Fat\_percentage distplot and boxplot…………………………23***  ***Figure 23: Insurance\_cost distplot and boxplot…………………………23***  ***Figure 24: Target Variable vs. Categorical Independent Variables…..24***  ***Figure 25: Target Variable vs Numeric Independent Variables………..27***  ***Figure 26: Correlation Plot…………………………………………………….29***  ***Figure 27: outliers detection………………………………………………….32***  ***Figure 28: outliers after treatment……………………………………………33***  ***Figure 29: Plot for predicted vs actual values on train data……………43***  ***Figure 30: Plot for predicted vs actual values on test data…………….43***  ***Figure 31: Plot for predicted vs actual values on train data……………44***  ***Figure 32: Plot for predicted vs actual values on test data…………….44***  ***Figure 33: Plot for predicted vs actual values on train data……………45***  ***Figure 34: Plot for predicted vs actual values on test data…………….45***  ***Figure 35: Plot for predicted vs actual values on train data……………47***  ***Figure 36: Plot for predicted vs actual values on test data…………….47***  ***Figure 37: Plot for predicted vs actual values on train data……………49***  ***Figure 38: Plot for predicted vs actual values on train data……………49***  ***Figure 39: Plot for predicted vs actual values on train data………….50***  ***Figure 40: Plot for predicted vs actual values on test data…………..50***  ***Figure 41: Plot for predicted vs actual values on train data………….51***  ***Figure 42: Plot for predicted vs actual values on test data…………..51***  ***Figure 43: Plot for predicted vs actual values on train data………….53***  ***Figure 44: Plot for predicted vs actual values on train data………….53*** |
| **List of Tables**  ***Table1: First 5 rows of dataset……………………………………………9***  ***Table2: data types…………………………………………………………..9***  ***Table3: descriptive statistics of the data………………………………10***  ***Table 4: missing values…………………………………………………….31***  ***Table 5: one hot encoding…………………………………………………34***  ***Table 6: coefficients………………………………………………………...37***  ***Table 7: OLS summary……………………………………………………...38***  ***Table 8: VIF values ………………………………………………………….39***  ***Table 9: OLS of model 33…………………………………………………..40***  ***Table 10: OLS of model 33 final equation………………………………41***  ***Table 11: coefficients of ridge model……………………………………47***  ***Table 12: coefficients of Lasso model………………………………….48***  ***Table 13: performance metrics of basic models………………………53***  ***Table 14: performance metrics of tuned models……………………..54***  ***Table 15: model comparison………………………………………………55*** |

1. **INTRODUCTION**
2. **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.

1. **Defining problem statement:**

The goal of this analysis is to understand the relationship between the different variables given in this data and how these features can predict the optimum insurance cost for individual.

1. **Scope:**

To find the optimal model and contributing variables impacting target variable, perform EDA and predictive model analysis using a variety of models.

1. **Need of the study/project**

* The objective of this exercise is to build a model, using data that provide the optimum insurance cost for an individual.
* The analysis deals with the prediction of best insurance cost based on the health and habit related parameters given in the data.

1. **Understanding business/social opportunity**

This section aims at understandings that how will such kind of a project or a study generate business profitability or social benefits.

The demand for health insurance is increasing because of population aging and the growing prevalence of chronic disease, as well as rising incomes in the developing world.

Health insurance as it is different from other segments of insurance business is more complex because of serious conflicts arising out of adverse selection, moral hazard, and information gap problems. If not properly regulated, does have adverse consequences for the costs of care, equity, consumer satisfaction, fraud and ethical standards. Concerning the value of insurance in the lives of individuals, it becomes important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it. Various variables estimate these charges. Each factor of these is important. If any factor is omitted when the amounts are computed, the policy changes overall. It is therefore critical that these tasks are performed with high accuracy.

Increase in health insurance knowledge is helping people to increase their awareness about the risk to be covered through insurance. Change in lifestyle is leading to increase in risk thereby giving an opportunity to insurance companies to innovate newer products. Societal benefit is derived by transfer of risk through insurance due to improved socio-cultural environment. For insurance companies understand the factors that impact user’s health insurance premium would be very essential to make the accurate charge, premium always be a user’s priority consideration to make appropriate decisions.

Hence, as a data scientist, it’s our duty to predict optimum insurance cost based on the given attributes. Therefore, this project becomes an imperative to the lives of people, as well as to the profits of the companies of the nation and aboard.

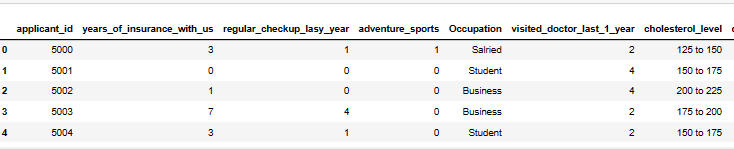
**2. EDA and Business Implication**

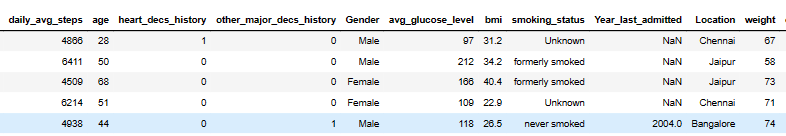
**a) Understanding how data was collected in terms of time, frequency and methodology**

This section aims at giving how the data is collected. This is the capstone project driven by the great Learning, hence the data of “Health care project” is provided to us from the learning platform.

**b) Visual inspection of data (rows, columns, descriptive details)**

We can see the initial look of the data.



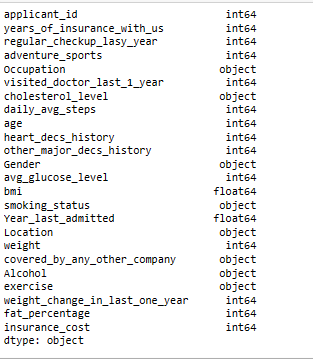




*Table1: First 5 rows of dataset*

In the data set, we have 25000 records and 24 columns, out of which

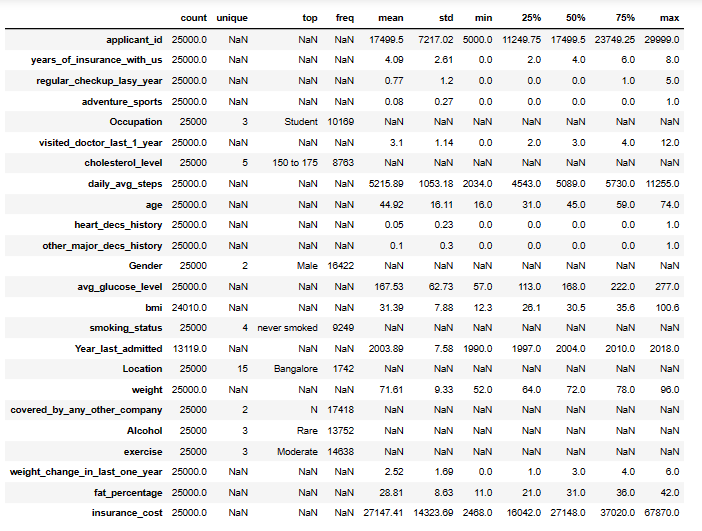
* 14 features are of integer type
* 8 features are of object data type
* 2 features are of float type



*Table2: data types*

**c) Descriptive details:**

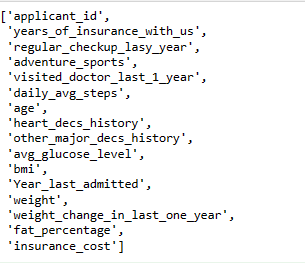
Below graph shows descriptive statistics of the data.



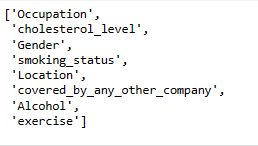
*Table3: descriptive statistics of the data*

Besides graphs, statistics that summarize the distribution of the data are used to transform data into information. The above table is summary statistics of the data.

Numerical data – there are 16 parameters such as



Categorical data – remaining 8 out of 24 are categorical data.



**d) Understanding of attributes (variable info, renaming if required)**

The various attributes provided are

* **applicant\_id**  - Applicant unique ID
* **years\_of\_insurance\_with\_us** - Since how many years customer is taking policy from the same company only
* **regular\_checkup\_lasy\_year** - Number of times customers has done the regular health checkup 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 **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
* **heart\_decs\_history** - Any past heart diseases
* **other\_major\_decs\_history** - 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 have been admitted in 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. This is given target variable for this dataset

There are 16 continuous variables and 8 categorical variables available in this data set. Out of these, 17 variables are health related parameters. Other variables are habit related variables.

All the variable names seem to be perfect except **regular\_checkup\_lasy\_year**. There is a small spelling mistake. So we need rename this to **regular\_checkup\_last\_year.**

**E) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)**

**Univariate Analysis categorical variables**:

**Occupation observation:**

|  |
| --- |
|  |

*Figure 1: Count Plot of Occupation variable*

* Based on above graph we can say most of the customers are students followed by business and lowest members are salaried.

**Cholesterol\_level observation:**

|  |
| --- |
|  |

*Figure 2: Count plot of cholesterol level variable*

* We can say highest number of customers have 150 to 175 cholesterol levels followed by 125 to 150.
* The lowest percentages of customers have 225 to 250 levels.
* Overall most of the customers have normal cholesterol level.

**Gender observation:**

|  |
| --- |
|  |

*Figure 3:* ***Count plot of Gender variable***

* Based on above graph highest percentage of the customers are male.

**Smoking \_status observation**

|  |
| --- |
|  |

*Figure 4: Count plot of smoking\_status variable*

* **From above plot we have observed highest numbers of customers are never smoked followed by unknown. The least numbers persons are smoked.**

**Location observations:**

|  |
| --- |
|  |

*Figure 5: Count plot of Location variable*

* From above output we have identified almost all cities have equal number of customers. Bangalore has slightly highest number followed by Mangalore. Surat has slightly lowest when compared to other cities.

**Covered\_by\_any\_other\_company observations:**

|  |
| --- |
|  |

*Figure 6: Count plot of covered\_by\_any\_other\_company variable*

* For this variable, highest number of the customers do not enrolled to other insurance company. Very few numbers of customers have insurance in other company.

**Alcohol observations:**

|  |
| --- |
|  |

*Figure 7: Count Plot of Alcohol variable*

* From above graph we have found highest numbers of customers take alcohol rarely and very few members have alcohol daily.

**Exercise observation**:

|  |
| --- |
|  |

*Figure 8: Count plot of exercise variable*

* Based on above output highest numbers of customers do moderate exercise and a very few members do extreme exercises.

# Univariate Analysis for numeric variable:

**Box plot & density plot of each numerical column as a subplot:**

**Years\_of\_insurance\_with\_us observations**

|  |
| --- |
|  |

*Figure 9: Years\_of\_insurance\_with\_us distplot and boxplot*

* For feature 'years\_of\_insurance\_with\_us' seems to be uniformly distributed. Most of the values lie in between 0 to 3 and 5 to 8. Fewer values lie 3 to 5.there is no outliers in this variable.

**regular\_checkup\_last\_year observations**

|  |
| --- |
|  |

*Figure 10: regular\_checkup\_last\_year distplot and boxplot*

* For 'regular\_checkup\_last\_year', data is right skewed distribution. There is a long tail on left side this indicates outliers are present in the data. The box plot confirms outliers are present in upper bond of box plot.

**adventure\_sports observations**

|  |
| --- |
|  |

*Figure 11: adventure\_sports observations distpolt and boxplot*

* For 'adventure\_sports' variable data is right skewed distribution. And most of the values are lie in between 0.0 to 0.2. Very fewer values lie in 0.8 to 1.0. So this indicates most of the values in this variable are zeros.
* Outliers are present in this data.

**Visited\_doctor\_last\_1\_year observations**

|  |
| --- |
|  |

*Figure 12: Visited\_doctor\_last\_1\_year distplot and boxplot*

* For 'visited\_doctor\_last\_1\_year', the data is right skewed and long tail present in left side. Most of the values lie in between 0.0 to 5.0. Low values present in 0.5 to 7.5. Box plot shows outliers are present in upper bond side.

**Daily\_avg\_steps observations**

|  |
| --- |
|  |

*Figure 13: Daily\_avg\_steps distplot and boxplot*

* For 'daily\_avg\_steps' feature, the data seems too distributed symmetrical. Most of the values present between in 2000 to 10000. The outliers are present in this variable.

**Age observations**

|  |
| --- |
|  |

*Figure 14: Age observations distplot and boxplot*

* For 'age', data distributed uniformly. The age distribution of people aged between 20 to 80.

**Heart\_decs\_history observations**

|  |
| --- |
|  |

*Figure 15: Heart\_decs\_history distplot and boxplot*

* For 'heart\_decs\_history', right skewed data is present. All the values lie between in 0.0 to 0.2. This will indicate most of the values are 0. There are outliers are present.

**Other\_major\_decs\_history observations**

|  |
| --- |
|  |

*Figure 16: Other\_major\_decs\_history distplot and boxplot*

* For 'other\_moajor\_decs\_history', right skewed data is present. All the values lie between in 0.0 to 0.2. This will indicate most of the values are 0. There are outliers are present.

**Avg\_glucose\_level observations**

|  |
| --- |
|  |

*Figure 17: Avg\_glucose\_level distplot and box plot*

* For 'avg\_glucose\_level', data distributed uniformly. There are no outliers present in this data set.

**bmi observations**

|  |
| --- |
|  |

*Figure 18: bmi observations distplot and boxplot*

* For bmi, it is a right skewed data. Most of the values present in between 15 to 45. There are outliers are present in the data set.

**Year\_last\_admitted observations**

|  |
| --- |
|  |

*Figure 19:Year\_last\_admitted distplot and boxplot*

* For year\_last\_admitted , data distributed uniformly.

**Weight observations**

|  |
| --- |
|  |

*Figure 20: Weigh distplot and boxplot*

* For ‘weight', the data seems too distributed symmetrical.

**Weight\_change\_in\_last\_one\_year observations**

|  |
| --- |
|  |

*Figure 21: Weight\_change\_in\_last\_one\_year distplot and boxplot*

* For 'weight\_change\_in\_last\_one\_year', data distributed uniformly. Most of the values present 0 to 1.

**Fat\_percentage observations**

|  |
| --- |
|  |

*Figure 22: Fat\_percentage distplot and boxplot*

* For 'fat\_percentage', this variable has right skewed data. There are no outliers present in this data.

**Insurance\_cost observations**

|  |
| --- |
|  |

*Figure 23:* ***Insurance\_cost distplot and boxplot***

* For 'insurance\_cost', it is slightly right skewed data. Slightly tail is present on left side. There are no outliers present in this data.

**f) Bivariate analysis (relationship between different variables, correlations)**

## Target Variable vs. Categorical Independent Variables

|  |  |  |
| --- | --- | --- |
| **Plot of insurance\_cost with occupation** | **Plot of insurance\_cost with cholesterol\_level** | **Plot of insurance\_cost with gender** |
| **Plot of insurance\_cost with smoking\_status** | **Plot of insurance\_cost with Location** | **Plot of insurance\_cost with covered\_by\_any\_other\_company** |

|  |  |
| --- | --- |
| **Plot of insurance\_cost with alcohol** | **Plot of insurance\_cost with exercises** |

## *Figure 24: Target Variable vs. Categorical Independent Variables*

**Insights from Bivariate analysis of Target vs categorical variable observations:**

When we compare categorical variable with target variable insurance cost, we do not see any difference in insurance cost. Insurance cost is same for all subcategories, except insurance covered \_by \_any \_other Company. There is slight difference between ‘Y’ and ‘N’. Who have enrolled with other company got high insurance cost.

## Target Variable vs Numeric Independent Variables

|  |  |  |
| --- | --- | --- |
| **Plot of insurance\_cost with years\_of\_insurance\_with\_us** | **Plot of insurance\_cost with regular\_checkup\_last\_year** | **Plot of insurance\_cost with adventure\_sports** |
| **Plot of insurance\_cost with visited\_doctor\_last\_1\_year** | **Plot of insurance\_cost with years\_of\_insurance\_with\_us** | **Plot of insurance\_cost with heart\_decs\_history** |
| **Plot of insurance\_cost with other\_major\_decs\_history** | **Plot of insurance\_cost with weight\_change\_in\_last\_one\_year** | **Plot of insurance\_cost with fat\_percentage** |

|  |  |
| --- | --- |
| **Plot of insurance\_cost with daily\_avg\_steps** | **Plot of insurance\_cost with avg\_glucose\_level** |
| **Plot of insurance\_cost with age** | **Plot of insurance\_cost with bmi** |
| **Plot of insurance\_cost with year\_last\_admitted** | **Plot of insurance\_cost with weight** |

## *Figure 25: Target Variable vs Numeric Independent Variables*

**Insights from Bivariate analysis of Numeric variable vs target variable:**

* 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 check up, 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.

# h)Multivariate analysis:

## Correlation Plot:

|  |
| --- |
|  |

## *Figure 26: Correlation Plot*

**Observations:**

With the help of heat map we can find correlation between 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.

**i) Business insights**

* 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. We can focus on these areas by providing awareness about insurance policies.
* 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 programmes and encourage them to follow these programs. And creating awareness of healthy weight will prevent health risks.

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

1. **Removal of unwanted variables (if applicable):**

Based on above analysis, now we can remove **"applicant\_id"** variable. This variable explains basic information about customer as well used to check duplicate values. We don’t use for our model. So we will drop this column.

For variables **'regular\_checkup\_last\_year'**, 60% of data have zeros and in similar way **'adventure\_sports'**, **'heart\_decs\_history'**, **'other\_major\_decs\_history'** variables have 91%, 94%, 90% zeros.

So we can imagine these all are outliers. But we can’t treat these outliers in conventional route.

Because here Q1 = 0

Q3 = 0

IQR = Q3-Q1 = 0

If we calculate UB and LB

UB = Q3+ (1.5\*IQR) = 0

LB = Q1- (1.5\*IQR) = 0

These all values are '0'

If we go to treating these outliers in these columns in conventional way like replacing with UB and LB values end of the result all values replaced with zero.

So it’s better to remove these variables from the data set. Moreover, this features absolutely useless for predicting models.

After removing unwanted variables:

Data shape is

The numbers of observations are 25000

The numbers of columns are 19

1. **Missing Value treatment (if applicable)**

|  |
| --- |
|  |

*Table 4: missing values*

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 variable contain 0.03% of missing values and this variable has outliers. So these missing values imputed with median value.
* The Column “year\_ last \_admitted “has about 47% of 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.
* But now these missing values imputed with median value. Later In the analysis, after performing VIF is performed; it can be removed if necessary.

There are no duplicates in this dataset.

1. **Outlier treatment (if required)**

|  |
| --- |
|  |

*Figure 27: outliers detection*

* We have seen outliers for columns **visited\_doctor\_last\_1\_year**, **daily\_avg\_steps, bmi, year\_last\_admitted.**
* One of the most important steps as part of data preprocessing is detecting and treating the outliers as they can negatively affect the statistical analysis and the training process of a machine learning algorithm resulting in lower accuracy.
* If there are outliers in the data, they should not be removed or ignored without a good reason. Whatever final model is fit to the data would not be very helpful if it ignores the most exceptional cases.
* We used user defined function to get upper and lower bounds of numeric columns for outlier capping and flooring.
* After that we treated outlier column in capping and flooring method.

|  |
| --- |
|  |

*Figure 28: outliers after treatment*

1. **Variable transformation (if applicable)**

**Converting categorical to dummy variables (Encoding:**

* 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 onehot 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:

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

*Table 5: one hot encoding*

**4. Model building**

For this dataset we use regression models because the dependent variable is a continuous variable.

The regression analysis is a predictive method that explores the relationship between a dependent (target) and the Independent variable (predictor). In this analysis I want to analyze the relationship between insurance Cost (target variable) and six independent variables. I 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.

**Models used**:

* **Linear Regression Model**
* **Linear Regression Model using stats Models**

# Decision Tree Regression basic Model

# Random forest Regression basic Model

# XGBoost Regression basic Model

**Models tuning done:**

# Ridge regression model

# Lasso regression model

# Decision Tree Regression tuned model

# Random forest Regression tuned model

# XGBoost Regression tuned model

**Metrics Considered:**

* **R square score**
* **Mean Squared Error (MSE)**
* **Root Mean Squared Error (RMSE)**
* **Mean Absolute Error(MAE)**

1. **Training and Testing Data**

* Splitting the dataset into training (70%) and testing (30%) subsets helps to assess the performance of the model over an independent dataset.
* We train the model using training data set and then evaluate the models performance using the testing data set, which is independent of the training data set.

1. **Model building**

**Linear Regression Model**

Linear regression is used to estimate the association of independent (predictor) variables with a continuous dependent (outcome) variable.

* 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 R2, describes the proportion of the variability in the outcome variable that can be explained by the independent variables.

**The coefficients for each of the independent attributes:**

|  |
| --- |
|  |
|  |

*Table 6: coefficients*

**The intercept for the model:** 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**

* **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**

The model worked really good on both train and test dataset with R square value being 94%.

# Linear Regression using stats models

**Basic Model– Considering All Independent Variables:**

* With the treated model, a basic model has been ran using ordinary least squares Method (OLS) where all the independent variables are taken in.

|  |
| --- |
|  |

*Table 7: OLS summary*

* 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.

**The VIF of the predictors:**

|  |
| --- |
|  |

*Table 8: VIF values*

* There is no multi collinearity within all the variables subjected to base model.
* Since all the VIF values are within the range (not more than 6), it is better to check P-values obtained and make decision for elimination of variables in the consecutive model.

**After removing variables one by one those have high p-values this is the final model named as 33 models.**

**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:**

OLS of model 33 are given below,

|  |
| --- |
|  |

*Table 9: OLS of model 33*

Final Equation from Model 33 will have 7 variables as follows,

|  |
| --- |
| *Table 10: OLS of model 33 final equation* |

## Assumptions 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:-

|  |  |
| --- | --- |
| 1. Linearity 2. Independence 3. Homoscedasticity | 4.Normality of error terms  5.No strong Multicollinearity |

### All the assumptions of linear regression are now satisfied.

## Now we can make predictions on data.

We can now use the model for making predictions on the test data.

* **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.

# Building Decision Tree Regression Basic model with all variables

Decision Tree is a supervised learning techniquethat can be used for both classification and Regression problems. It is a tree-structured classifier, where**internal 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.

**Fit the model:**



**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 29: Plot for predicted vs actual values on train data* | *Figure 30: Plot for 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.

# Building Random Forest Regressor basic model with all variables

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.

**Fit the model:**



**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 31: Plot for predicted vs actual values on train data* | *Figure 32: Plot for predicted vs actual values on test data* |

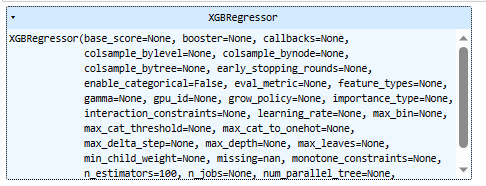
* As expected, an overgrown tree has been obtained with a train accuracy of 99.33% while the test data accuracy was 95.34%.
* We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
* MAE indicates that our current model is able to predict insurance 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.

**Model Building with XGBoost Regressor basic 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.

**Fit the model:**



**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 33: Plot for predicted vs actual values on train data* | *Figure 34: Plot for predicted vs actual values on test data* |

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.

**c) Model tuning measures**

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:**

**Regularization:**

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. **If the 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 coeff 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

# . Create a regularized RIDGE model:

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

**Fit the model:**



**Coefficients:**

|  |
| --- |
|  |

*Table 11: coefficients of 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 35: Plot for predicted vs actual values on train data:* | *Figure 36: Plot for predicted vs actual values on test data:* |

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 visualised. Model prediction on dependent variable seems like it is predicting better both in train and test data.

# Create a 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.

Fit the model:

****

# Coefficients:

|  |
| --- |
|  |

*Table 12: coefficients of Lasso model*

Based above coefficients this model explains the weight variable is more significant to predict insurance cost. The remaining variables are insignificant.

**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**

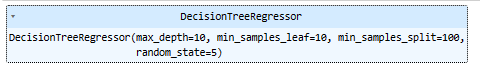
|  |  |
| --- | --- |
| *Figure 37: Plot for predicted vs actual values on train data:* | *Figure 38: Plot for predicted vs actual values on train data:* |

* 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 Regression tuned model

The **parameters** used in this model is **criterion = squared\_error, max\_depth = 10, min\_samples\_leaf=10, min\_samples\_split=100, random\_state=5**.

Fit the model:

****

**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 39: Plot for predicted vs actual values on train data* | *Figure 40: Plot for predicted vs actual values on test data* |

* 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 prevent our model from overfitting.

# Random Forest Regressor tuned model

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

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

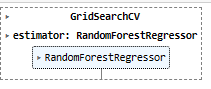
**These are the parameters were passed in GridSearchCV with CV value of 3,**

|  |  |
| --- | --- |
| Criterion = squared error, mse  Max depth = 5,7,10  Min sample leaf = 5,10 | Min sample split = 50, 100, 150  N\_estimetors = 100,200,300 |

**Out of this these are best parameters from GridSearch CV,**

|  |  |
| --- | --- |
| Criterion = squared error  Max depth = 7  Min sample leaf = 10 | Min sample split = 50  N\_estimetors = 100 |

**Fit the model:**

****

**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 41: Plot for predicted vs actual values on train data:* | *Figure 42: Plot for predicted vs actual values on test data* |

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 Regressor model Tunining

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**

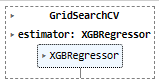
**max\_depth = 5,6,7, 8**

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

**n\_estimators = 100**

**max\_depth =5**

**Fit the model:**

****

**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 43: Plot for predicted vs actual values on train data* | *Figure 44: Plot for predicted vs actual values on train data* |

* 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

**5) Model validation**

**Test predictive model against the test set using various appropriate performance metrics**

**For every model building procedure used in the project, a few steps are used to get the output generated by each algorithm. The performance metrics opted to analyse these models are r-square score, MSE, RMSE, MAE in it for each model.**

|  |
| --- |
|  |

*Table 13: performance metrics of basic models*

Based on above analysis,

* From liner regression, this model performs well on both train and test dataset. For the regression models three error metrics (MSE, RMSE and MAE) that are commonly used for evaluating and reporting the performance of a regression model. In liner regression the MSE, RMSE and MAE values are low and it is comparable on the train and test datasets. R-squared of the model is 0.94 and adjusted R-squared is 0.94, which shows that the model is able to explain 94% variance in the data. This is quite good. This model is not suffering from over fitting.
* By using liner regression stats model we have found years\_of\_insurance\_with\_us,visited\_doctor\_last\_1\_year,Age,Year\_last\_admitted,weight,weight\_change\_in\_last\_one\_year,covered\_by\_any\_other\_company\_Y play vital role in predicting insurance cost.
* For decision regression model perform well on train but the model is over fitting. The other metrics (MSE, RMSE and MAE) are completely different on both train and test data. This model needs to be tuned to prevent from over fitting.
* For random forest base model performs well on this dataset. This model is able to explain 99% variance in the data and also MSE, RMSE and MAE scores explains the error between predicted and actual values are very low. So we can say this model also good for predicting insurance cost.
* The XGBoost model explains 97% of variance in the data set as well other error metric scores also low. This model also performed well on dataset.
* Before selecting best model we need to be performing model tuning by using GridCv method.

**Below is the output, after performed tuning on models:**

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

*Table 14: performance metrics of tuned models*

* Based on above output all models performed well. We can see now we have prevented decision tree model from overfitting.

**Final model comparison**

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

*Table 15: model comparison*

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.

**5) Final interpretation / recommendation**

**Insights**

* 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 increases 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.

**END**