Revanth Janapriyan

 PGP-DSBA ONLINE

 APRIL’ 22

23/08/2022

**DATA MINING**

Table of Contents

[PROBLEM 1 3](#_Toc112186078)

[***1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).*** 3](#_Toc112186079)

[EXECUTIVE SUMMARY 3](#_Toc112186080)

[INRODUCTION 3](#_Toc112186081)

[DATA DESCRIPTION 3](#_Toc112186082)

[DATA SAMPLE 4](#_Toc112186083)

[EXPLORATORY DATA ANALYSIS 4](#_Toc112186084)

[***1.2 Do you think scaling is necessary for clustering in this case? Justify.*** 13](#_Toc112186085)

[***1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.*** 15](#_Toc112186086)

[***1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.*** 16](#_Toc112186087)

[***1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.*** 18](#_Toc112186088)

[PROBLEM 2 20](#_Toc112186089)

[***2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).*** 20](#_Toc112186090)

[EXECUTIVE SUMMARY 20](#_Toc112186091)

[INRODUCTION 20](#_Toc112186092)

[DATA DESCRIPTION 20](#_Toc112186093)

[DATA SAMPLE 21](#_Toc112186094)

[EXPLORATORY DATA ANALYSIS 21](#_Toc112186095)

[***2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.*** 34](#_Toc112186096)

[***2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.*** 37](#_Toc112186097)

[***2.4 Final Model: Compare all the models and write an inference which model is best/optimized.*** 45](#_Toc112186098)

[***2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations*** 46](#_Toc112186099)

# PROBLEM 1

### ***1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).***

## eXECUTIVE SUMMARY

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

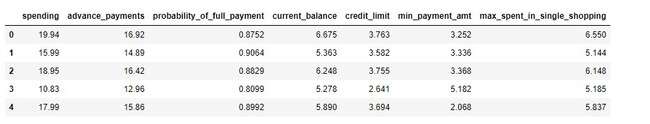
## INRODUCTION

In this first problem, we have taken various steps in order to categorize the customers of a bank based on credit card usage and help the marketing team to draft a plan to reach customers in best way possible. We have explored various aspects of the customer data such as spending, advance payment, credit limit, etc using data exploration techniques such as univariate analysis, bivariate analysis, graphical representations. Based on the data exploration results, techniques such as hierarchical clustering, dendrogram and K-means clustering are used to cluster the data. The clustering is checked for accuracy using methods such as elbow curve and silhouette score. Based on the clustering and its accuracy the marketing team can plan according to reach out the customers in specific ways.

## DATA DESCRIPTION

* **spending**: Amount spent by the customer per month (in 1000s)
* **advance\_payments**: Amount paid by the customer in advance by cash (in 100s)
* **probability\_of\_full\_payment**: Probability of payment done in full by the customer to the bank
* **current\_balance**: Balance amount left in the account to make purchases (in 1000s)
* **credit\_limit**: Limit of the amount in credit card (10000s)
* **min\_payment\_amt** : minimum paid by the customer while making payments for purchases made monthly (in 100s)
* **max\_spent\_in\_single\_shopping**: Maximum amount spent in one purchase (in 1000s)

## DATA SAMPLE

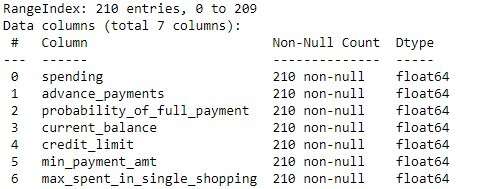


**TABLE 1.1**

There are 7 variables having various details of the customers activity on spending, probability of full payment, etc.

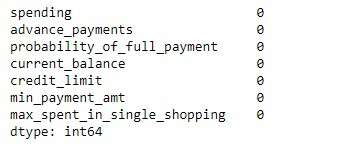
## **EXPLORATORY DATA ANALYSIS**

Let’s see the different types of variables and number of datapoints each has, data types each of variable and missing values.



**TABLE 1.2**

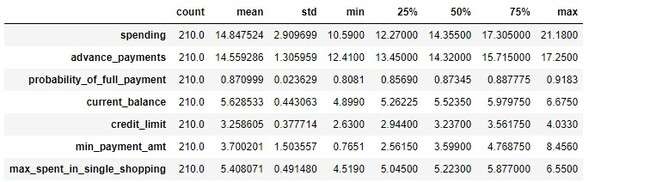
The table 1.2 shows that there are 7 float type features each having 210 entries.



**Table 1.3**

The table above confirms that there are no null/missing values present in the dataset.

Descriptive Analysis



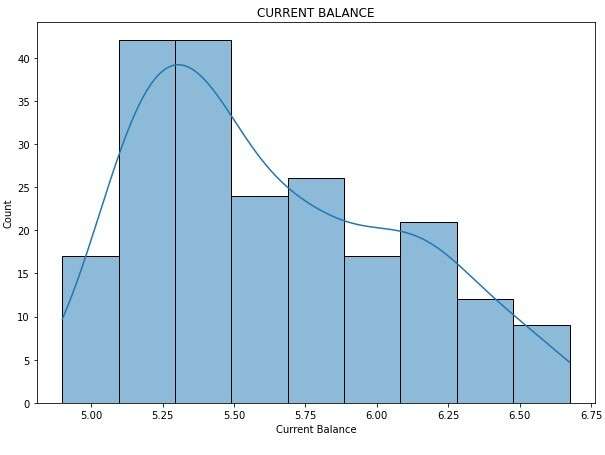
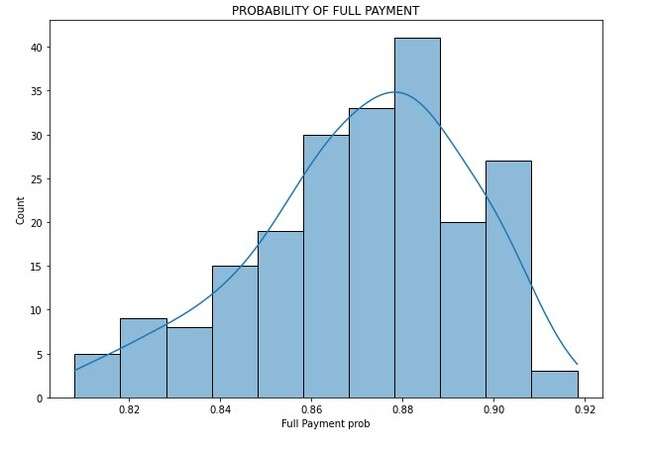
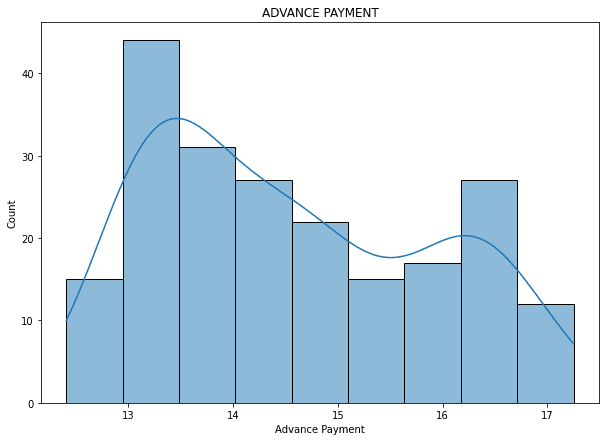
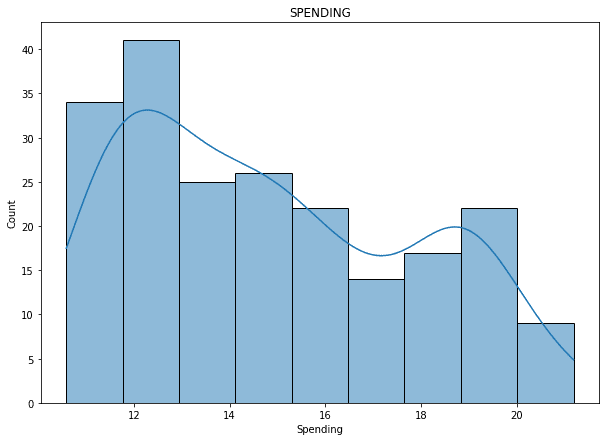
**Table 1.4**

With the help of the descriptive statistics, we can see various measure such as mean, Q1, Q2, Q3, min, max and standard deviation of each feature. Below are few measures:

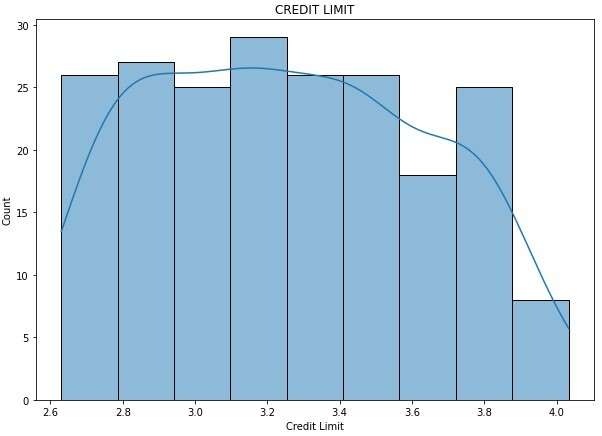
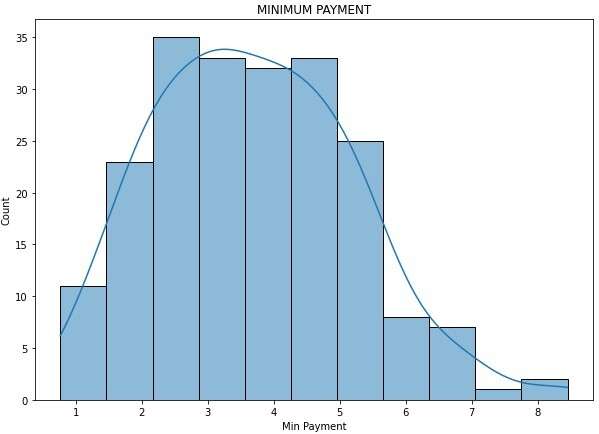
* From, the table 1.4 we can conclude that on average customers spend 14.8k using their credit card. Max spending is 21.1k and min spending is 10.59k.
* There is 87 % probability that on average customers will pay the full payment.

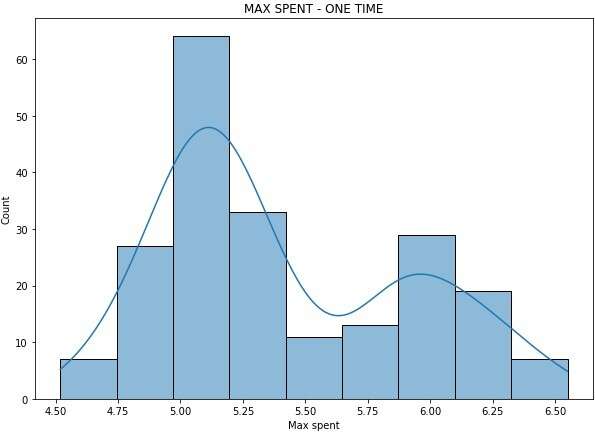
Visualizing spread of each variable:

The spread of the various features has been visualised with the help of histogram in the figure1.1, below.



**Fig 1.1**



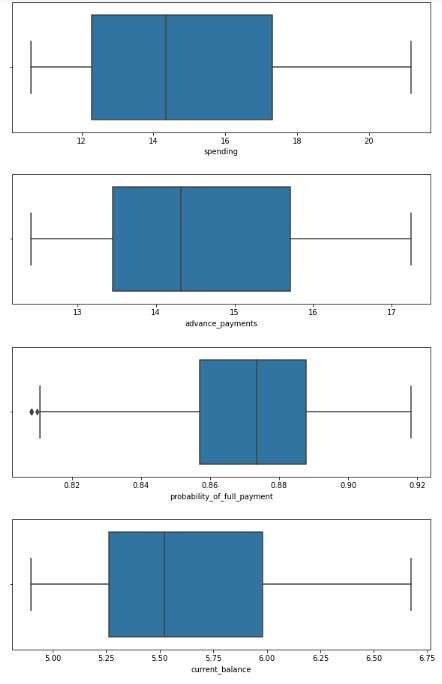
**Fig1.1**

With the help of the histograms created for each variable, we can interpret the following:

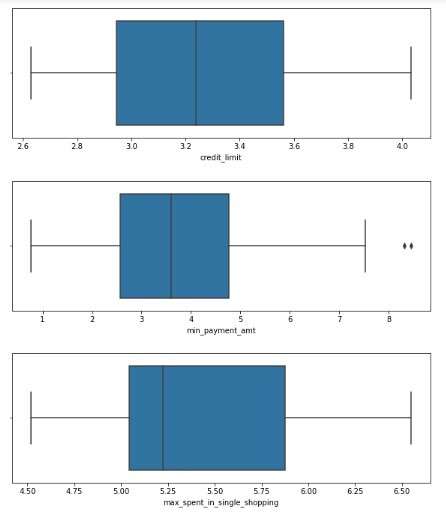
* Number of customers who spend around 12k-13k are greater than other spending caps. The least number customers spent over 20k.
* More than 40 customers pay 1300 -1350 in advance.
* There is 88% chance that more than 40 customers pay full payment.
* Most of the customers have a current balance ranging between 5.25k – 5.5k.
* Almost 30 customers have a credit limit of 320000, which is the highest. While, less than 10 customers have a credit limit of 400000.
* Most number customers pay a minimum balance between 200 -300.
* More than 60 customers spend a maximum of 5000 in a single purchase.

Outliers:

Outliers are the most crucial data points which affect the quality of the entire process. Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample. Visualizing the variable using boxplot can help us identify the outliers as below.



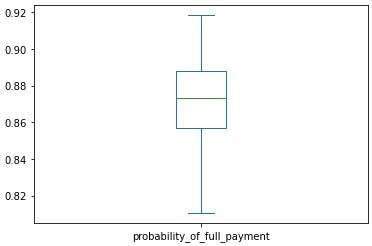
**Fig 1.2**

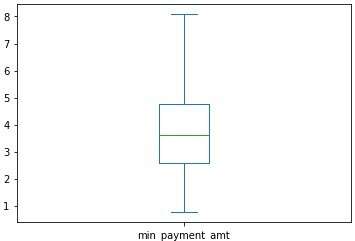


**Fig 1.2**

After visualizing all the features present in the data using the boxplot, it can clearly interpret that only the variables - probability\_of\_full\_payment and min\_payment\_amt have the outliers.

Further, the outliers are treated and both the variables are again checked for outliers using the box plot. The below Fig 1.3, shows the boxplots of the variables - probability\_of\_full\_payment and min\_payment\_amt after treating the outliers.





**Fig 1.3**

Skewness:

All the variables in the data are also checked for their skewness.

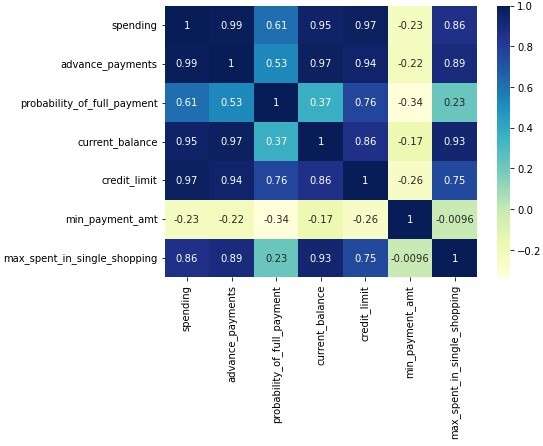


**Table 1.5**

Table 1.5 shows that all the variables except probability\_of\_full\_payment are right skewed.

Heatmap:

The correlation between each variable is visualized using the heat map in fig1.4.

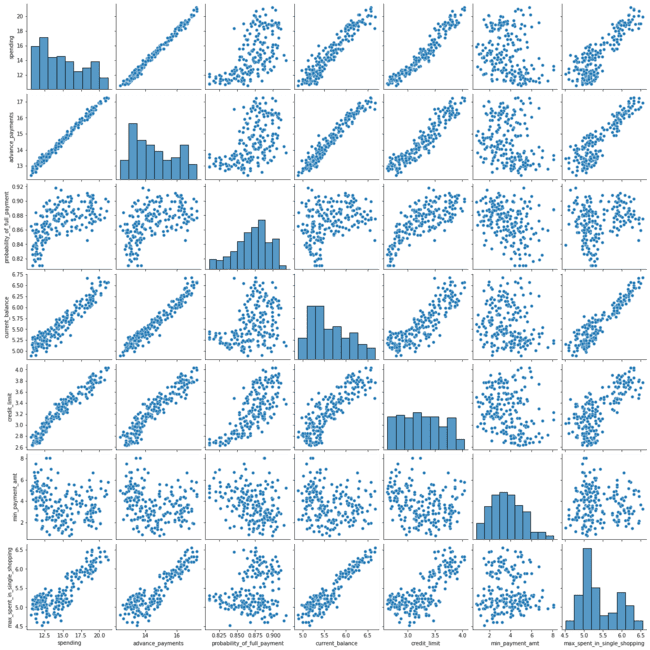


**Fig 1.4**

The heatmap shows that the variables spending, advance payment, current balance, credit limit and max spent in single shopping have higher correlation among each other than any other variables present in the dataset. The least correlation is between min payment amt and probability of full payment.

Pairplot:

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

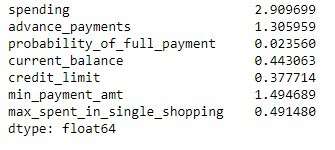


**Fig 1.5**

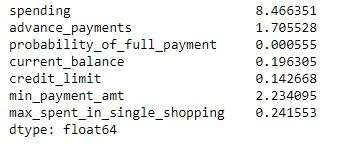
From the graph, we can see that there is positive linear relationship between variables between spending and advance payment. The histograms tell us that only the probability of full payment is left skewed.

### ***1.2 Do you think scaling is necessary for clustering in this case? Justify.***

Now, in order to decide whether to scale the data or to proceed with existing data, we have look upon few measures of central tendency. Below are few of those -standard deviation (Table1.6), variance (Table1.7) and covariance (Table1.8).



**Table 1.6**



**Table 1.7**

****

**Table 1.8**

As, it can be clearly seen that the standard deviation, variance and covariance are not very similar for each feature. This clearly suggest that scaling needs to done for each feature in order to make sure that no variable get more importance than other variable in the data and all variables are treated equally.

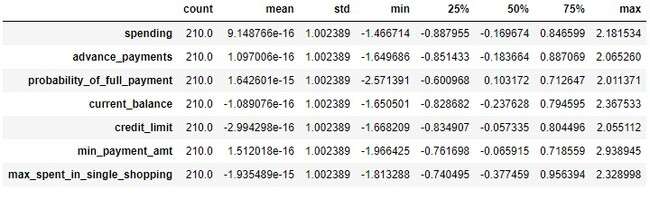
Scaling:

The features of the dataset are scaled using the Standscaler package present in the scikit learn. Below table shows the values of scaled dataset.



**Table 1.9**

Since, the dataset is scaled now, we can also use the descriptive statistic measure to know if it is properly scaled using the table 1.10.

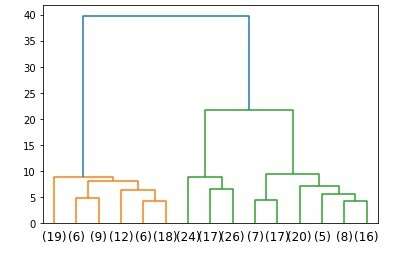


**Table 1.10**

Now, it can be confirmed that dataset is properly scaled as the standard deviation almost similar for all the variables. Also, other measure such as mean, min, max, Q1, Q2 and Q3 also proves that the dataset is scaled properly.

### ***1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.***

Now, since the data is scaled it can be used to identify the optimum clusters for the given data using Hierarchical clustering and dendrogram.



**Fig 1.6**

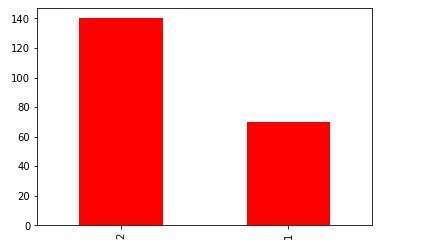
The dendrogram in the figure helps assume that data set can be clustered into two groups. Note: This dendrogram is obtaining after truncating with p value of 15.



**Table 1.11**

Table 1.11, shows the first 5 rows of the dataset after clustering using dendrogram and hierarchical clustering technique.

Further, the count of values in each cluster is visualized using the bar graph in the figure 1.7 below.



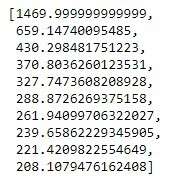
**Fig 1.7**

Here, we can clearly see that the second cluster has more value than the first. Thus, we can say that the data can be clustered into two groups, but still we have to verify with other method such as K-means in order to identify the best number of cluster.

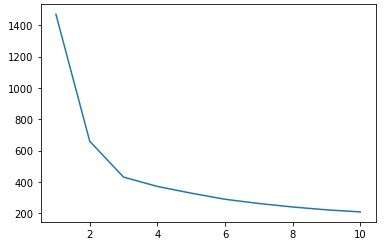
### ***1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.***

In K-means clustering methods such as Wss, Elbow curve and Silhouette score are used to find the optimum number of clustering. Since it uses two methods on confirming the number of clusters it is more effective and better in accuracy than the hierarchical clustering.

 WSS is the total distance of data points from their respective cluster centroids. The Elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.



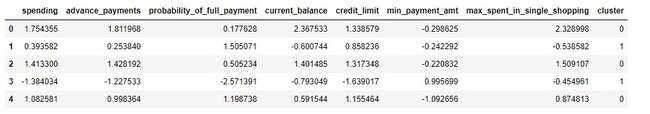
These are the values of with sum of the square(wss) obtained and used in plotting the elbow curve.



**Fig 1.8**

The elbow curve plot shows that the optimum number cluster can be either 2 or 3. In order to decide on that, silhouette score is used.

Before calculating silhouette score, two new datasets with cluster value of 2 and 3 are created individually. Those are shown in Tables 1.12 & 1.13.



**Table 1.12**

****

**Table 1.13**

Using the two data frames created for cluster values of two and three, the silhouette score and silhouette samples are calculated.

* The calculated silhouette score of clusters two and three are 0.46560 and 0.4008059 respectively.
* The calculated silhouette sample width of clusters two and three are -0.005677 and 0.0027685 respectively.

Even though, in terms of silhouette score one can go for the cluster size of 2, it is clearly evident that three is best clustering size in this case. The silhouette sample width always supports this since the value of cluster size two is negative.

### ***1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.***

Based on the results obtained from the clustering techniques, the clustering profiles can be of three types of customers based on their credit card usage. The clustered group can be – **BASIC, STANDARD** and **PREMIUM.**

**BASIC** – These are the credit card users who spend lesser than rest of the group.

**STANDARD** – The credit card users who falls between the basic and premium user groups (i.e.) average spending group.

**PREMIUM** - Most of credit card users present in this group use their credit cards more often for all kind purchase.

Recommendations:

* Basic category users spend lesser than the other group. These users are mostly new users who have recently credit cards or the are bachelors who have just started their career. We can pitch them with various specially offers on fuel, food delivery apps like Swiggy and Zomato, electronic accessories such as phone, laptop, bus booking, cab services, etc.
* Standard category users are usually of customers who are married, middle class or well earning. The bank can make the customers spent more by providing offers on grocery, stationery, food and dining, travel, fuel, paying rent, etc. Since, these customers are married bank can offer instant loans without processing charges at a considerable interest rate based on their spending nature.
* Premium customers are mostly from high earning group. The bank can offer them with discounts on top brands (clothing, electronics,etc), giving purchase coupons for online shopping sites or restaurants and bars, Personalised loan with interest rate lesser then other category users and other banks in the market can be provided, getting access to Hotels all around the world with discounted price and also for flights, allowing to redeem the points earned as cash, etc.

# PROBLEM 2

### ***2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).***

## eXECUTIVE SUMMARY

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

## INRODUCTION

The main intent of this exercise is to the dataset provided by the Insurance company and figure out what is the reason behind the higher claim frequency. In order provide solution to the insurance company for the higher claim frequency or to find out if there are any fraudulent claims being made models have been build to predict using various techniques. Decision tree classifier, Random Forest classifier and Artificial Neural Networks are the three machine learning models used in this problem. The performance of these models, which are evaluated using techniques such as Confusion Matrix, Classification report, Accuracy of the model, ROC curve and ROC\_AUC score. Based on the evaluation, the best performing model can be selected and used for solving the problem and providing recommendations.

## DATA DESCRIPTION

1. **Target**: Claim Status (Claimed)  
2. **Agency Code**: Code of tour firm of each firm  
3. **Type**: There are two types of tour insurance firms namely – Airlines and Travel Agency  
4. **Channel**: Contains the details Distribution channel of tour insurance agencies – Online and Offline

5. **Product Name**: Name of the tour insurance products - Customised Plan, Cancellation Plan, Gold plan, silver plan and bronze plan.  
6. **Duration**: Duration of the tour in days  
7. **Destination**: Destination of the tour  
8. **Sales**: Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)  
9. **Commission**: The commission received for tour insurance firm (Commission is in percentage of sales)  
10.**Age**: Age of insured

## DATA SAMPLE

The sample of the dataset provided by the insurance firm is shown in table 2.1.



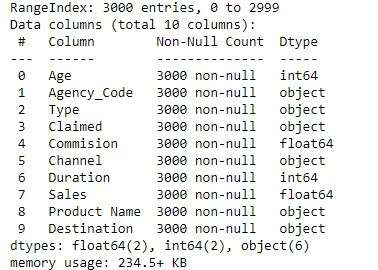
**Table 2.1**

There are 10 variables having various details of the customers, insurance agency, type of insurance, etc.

## **EXPLORATORY DATA ANALYSIS**

Let’s see the different types of variables and number of datapoints each has, data types each of variable and missing values.

The table 2.1 provides us with various information about the dataset and its data. There are total of 10 features present in the dataset, each of which has 3000 non-null data points.



**Table 2.1**

The dataset has six object columns, two float and two integer columns.

Descriptive Analysis:

Now, let us see the descriptive stats of each variable present in the dataset with the help of the table 2.2



**Table 2.2**

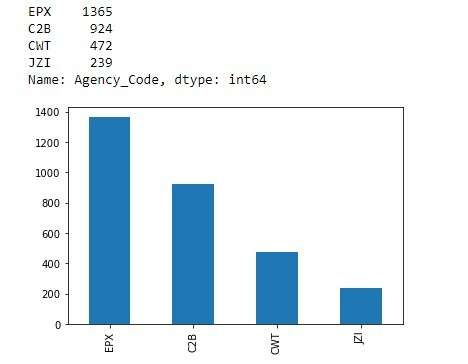
As, it can be seen that apart from Commision, Duration and Sales all other variables are categorical while these three variables are continuous. There are four unique values in Agency code, five in product name, three in destination and two each in type, claimed and Channel variable.

**Analysis of Categorical variables**

**As we know that there are seven categorical variable presents in the data set, let’s do some basic data analysis on those variables and see how the data points are spread across each variable starting with Agency code.**

**Agency Code:**

Below figure 2.1, shows the number of agency codes and its count in both graphical as well as numerical form.



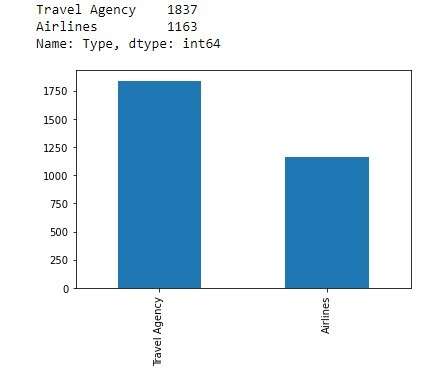
**Fig 2.1**

EPX tops the list with a massive count of 1365, while the least figure is owned by JZI.

Type:

There are two types of insurance firms namely “Travel Agency” and “Airlines”. Travel Agency

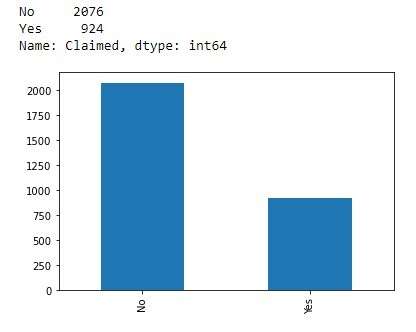
has upper hand in terms of counts here. (Refer Fig 2.2)



**Fig 2.2**

Claimed:

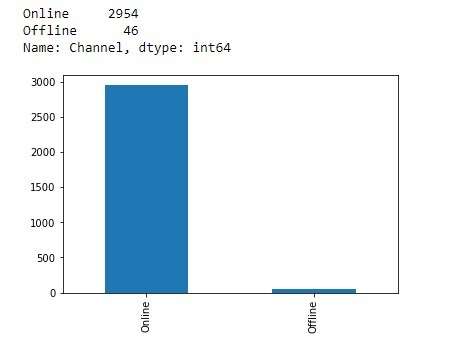
The Fig 2.3 shows the counts of claimed and unclaimed insurance. There are 2076 rejected insurance claims according to the data.



**Fig 2.3**

**Channel:**

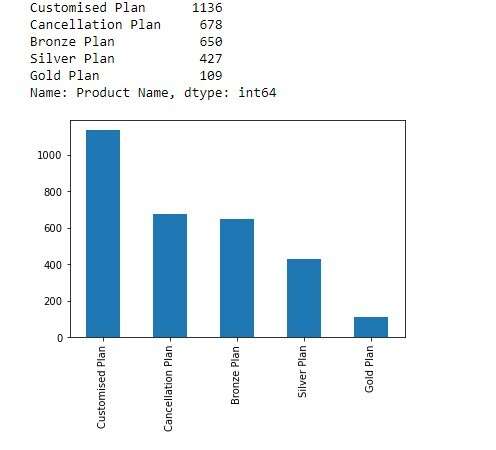
Channel is the way through which the agencies make claim to the insurance company. Online channel dominates this category. (Refer Fig 2.4).



**Fig 2.4**

*Product Name:*

The Company offers five types of products for its users.

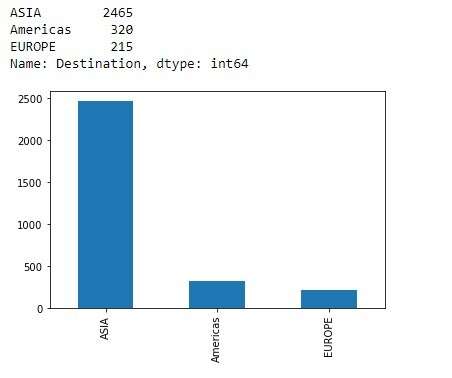


**Fig 2.5**

The different product types are Customised Plan, Cancellation Plan, Gold plan, silver plan and bronze plan. We can clearly see that most of the clients opt for customised plan and the least opted one is gold plan.

Destination:

Destination variable contains the information of tour destinations.



**Fig 2.6**

There are total of three destinations present in the variable namely – “Asia”, “Americas” and “Europe”. All of which are continents not country. Out of all three regions/destinations Asia is the most visited tourist spot.

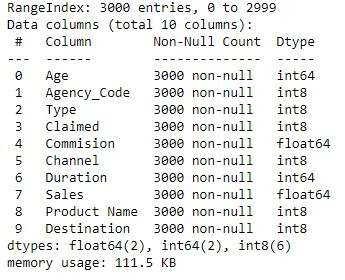
Now, we have seen the distribution across each variable, we noticed that all of the categorical variables have a data type ‘object’.

As, we know that Machine learning models require all input and output variables to be numeric. This means that if your data contains categorical data, you must encode it to numbers before you can fit and evaluate a model.

So, necessary processing is done into to convert the categorical variable in numeric and below tables shows the changes in the data points (Refer Table 2.3) and data types of those variables (Refer Table 2.4).



**Table 2.3**

****

**Table 2.4**

Now, it is confirmed that all of the categorical variables are converted into numeric datatype(integer).

Let us quickly see the descriptive stats of the converted variables before analysing continuous variables.



**Table 2.5**

In Table 2.5, values of categorical the variables are scaled to numeric values. Almost, all the categorical variables not have a mean and standard deviation close to 1.

**Analysis of Continuous variables**

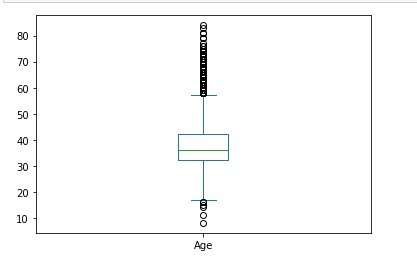
**Finishing the analysis of categorical variables, it’s time to analyse the continuous variables present in the data. The continuous variables present are:**

* **Age**
* **Commision**
* **Duration**
* **Sales**

**Let us start our analysis on continuous variable with “Age” variable. It is very crucial that we start with outliers and fix if there is any in those variables.**

**Age:**

The “Age” variable contained the age of all the insured persons.

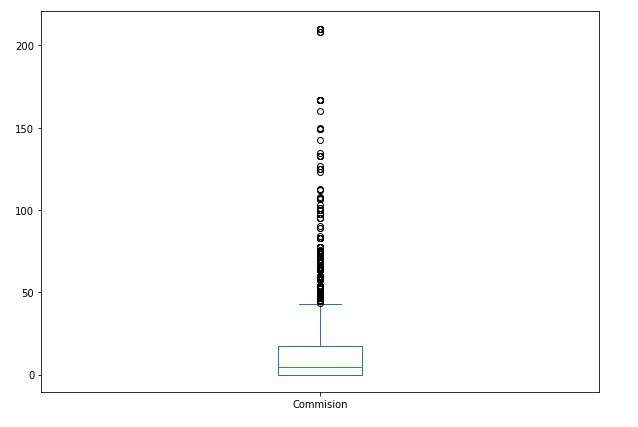


**Fig 2.7**

**Box plot in the Figure 2.7, shows that there is right skew and lots of outliers present in the age variable, which needs to be treated.**

**Commision:**

The boxplot below shows the distribution of commission variable’s data points.

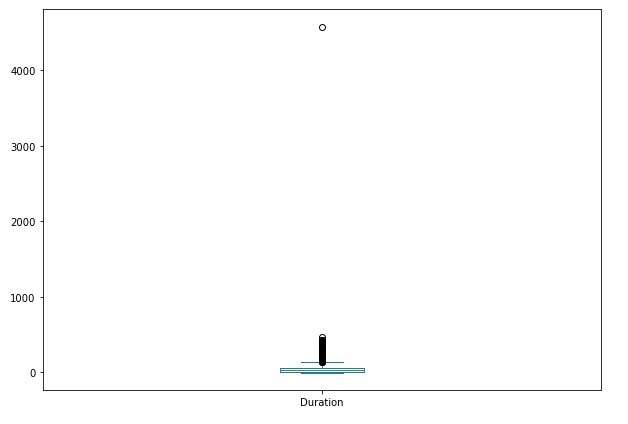


**Fig 2.8**

Similar to Age, Commision too has outliers and skewness. (Refer Figure 2.8)

Duration:

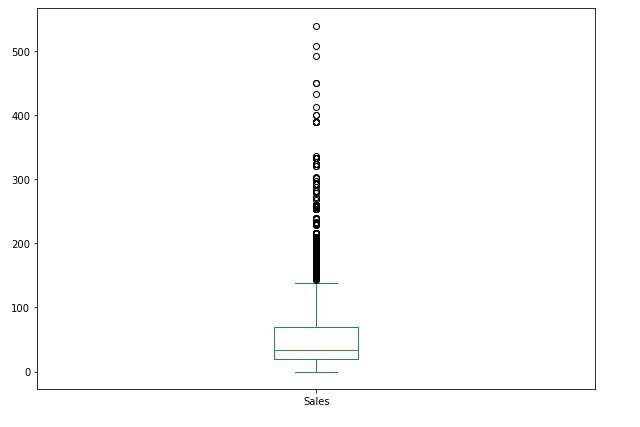
Duration is also with outliers similar to the one we saw previously. (Refer Figure 2.9)



**Fig 2.9**

Sales:

The distribution of Sales variables is shown in the boxplot below. (Refer Figure 2.10)

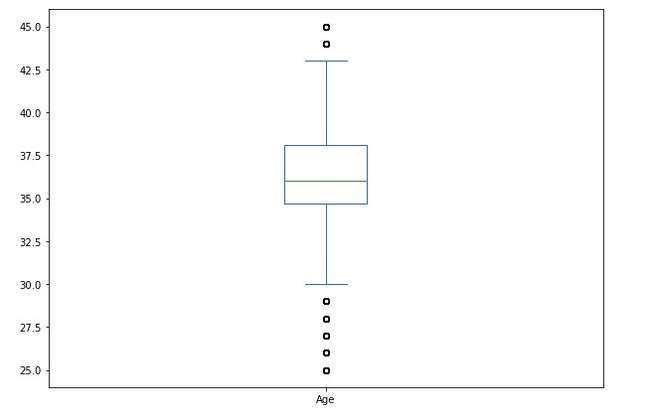


**Fig 2.10**

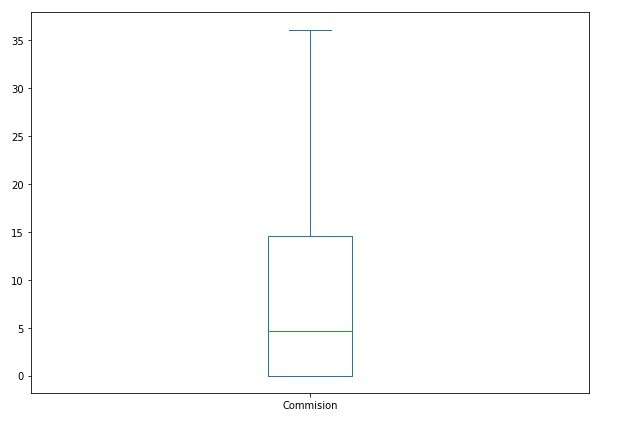
Sales variable is heavily right skewed and has lots of outliers.

***Outlier Treatment:***

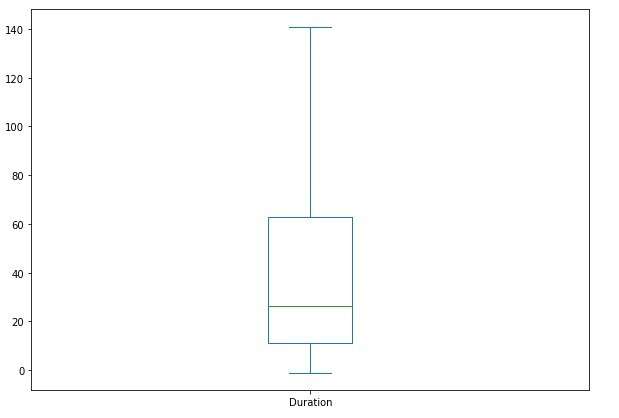
**As there are lots of outliers present in the continuous variables, necessary steps have been taken to eliminate the outliers as much as possible. Below figures (Refer Figures 2.11, 2.12, 2.13 and 2.14), shows the outlier treated continuous variables.**

****

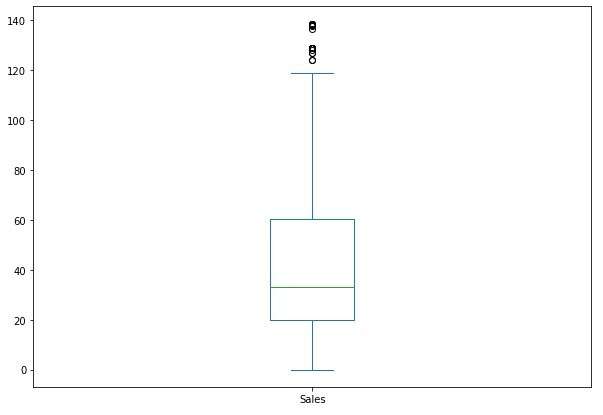
**Fig 2.11**

****

**Fig 2.12**

****

**Fig 2.13**

****

**Fig 2.14**

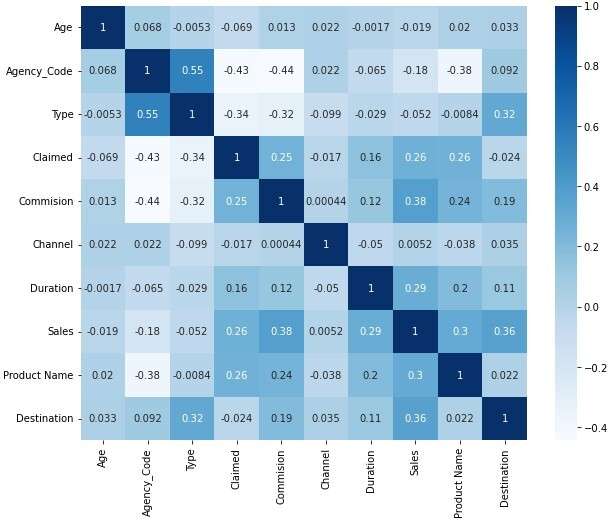
Now, after treating the outliers most the outlier has been eliminated. Only, few outliers are present in age and sales variable which can be ignore as it will not have much impact on the results.

**Correlation:**

**Before completing the exploratory data analysis, we need to see if there is any correlation among the variables using the correlation table (Refer Table 2.6) and Heatmap (Refer Figure 2.15).**



**Table 2.6**

****

**Fig 2.15**

Upon, carefully examining both the correlation table and Heat map there is no clear evidence that there is correlation among the variables.

### ***2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.***

Now, with exploratory data analysis is done, time to build the machine learning models. Before we build Machine learning model dataset is divided into two – training and testing data.

As Empirical studies shown that the best results are obtained if we use 20-30% of the data for testing, and the remaining 70-80% of the data for training. So, the data is divided as 70% for training and 30% for testing. Data has also been scaled before building the models.

ANN Model:

In order Neural network model, MLPClassifier has been imported from sklearn.neural\_network which a sub module of scikit learn package. Once the necessary packages are imported the ann model is build using the following parameters – Hidden layer, maximum iteration, random state, solver, verbose and tol.

**hidden\_layer\_sizes** – This parameter is used to define the number of hidden layers(neurons) the model should have. Here, the hidden layer is initialized to be 500.

**max\_iteration** – Number of iterations. In this model it is set to a random value of 5000.

**random\_state** – Random state is defined in order to ensure that the model’s output does not even if it runs on different system. Random state for this model is 1.

**solver** – Solver is used for weight optimization. Solver used in this model is “sgd” (stochastic gradient descent).

**verbose** – Verbose is used to print the progress of the model as output. Verbose is set as True to print it in output.

**tol** – Tolerance is used to optimize model for desired output. tol value used in this model is 0.01.

Now as, the model is built the training and test data has been used and the score of the model on train and test data are 76.90% and 73.7% respectively.

Later, the grid search cv is imported and used for finding the best parameters and the best estimators. Finally, best parameters obtained by grid search cv is 'hidden\_layer\_sizes': 500, 'max\_iter': 2500, 'solver': 'adam', 'tol': 0.01 and the best estimators are hidden\_layer\_sizes=500, max\_iter=2500, random\_state=5, tol=0.01.

Even after using the best parameters and estimators the model’s performance did not improve.

Decision Tree:

The Decision tree classifier has the following parameters

**criterion** –This is a used to specify the quality of the split across the tree. In this model ‘gini’ is used as criterion.

**max\_depth** – This parameter is used to define the depth of the tree.

**min\_samples\_split** – The minimum number split required at each node is mentioned here.

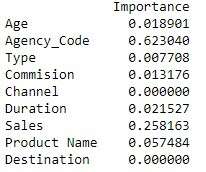
**min\_samples\_leaf** – This is used specify the minimum number samples required at each leaf node.

Unlike, Ann model here the grid search cv is approached first in order to save time and effort in building the model.

The best parameters obtained using grid search cv are 'criterion': 'gini', 'max\_depth': 10, 'min\_samples\_leaf': 50, 'min\_samples\_split': 150 and the best estimators are max\_depth=10, min\_samples\_leaf=50, min\_samples\_split=150, random\_state = 1.

Using the best parameters and best estimators a decision tree model has been built. Graphviz is used to prune the unwanted branch of the decision tree model.

The table below shows the importance given to each feature while making the decision tree.



**Table 2.7**

Out of all the features Agency code and Sales are the top two important features, while destination is the least.

Random Forest:

Random Forest classifier can be defined as the name suggests it is collection(forest) of decision trees. The Random Forest classifier has the following parameters:

**n\_estimators** – Used for defining number of trees.

**oob\_score** – It is used to define whether to use out-of-bag samples to estimate the generalization score.

**max\_depth** – Used to define maximum depth of the tree.

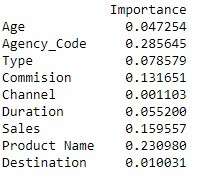
**max\_features** – Used to specify the number of features to consider when looking for the best split.

**min\_samples\_leaf** – Used to specify the minimum of samples required to be at a leaf node.

**min\_samples\_split** – Used to define the minimum number of samples required to split an internal node.

Using, the grid search cv the best parameters obtained are 'max\_depth': 5, 'max\_features':3,

'min\_samples\_leaf': 5, 'min\_samples\_split': 15, 'n\_estimators': 250 and the best estimators are max\_depth=5, max\_features=3, min\_samples\_leaf=5, min\_samples\_split=15, estimators=250.



**Table 2.8**

The table 2.8, shows that the Agency code and Product Name are the top two important features consider in this random forest model and the least considered feature is Channel.

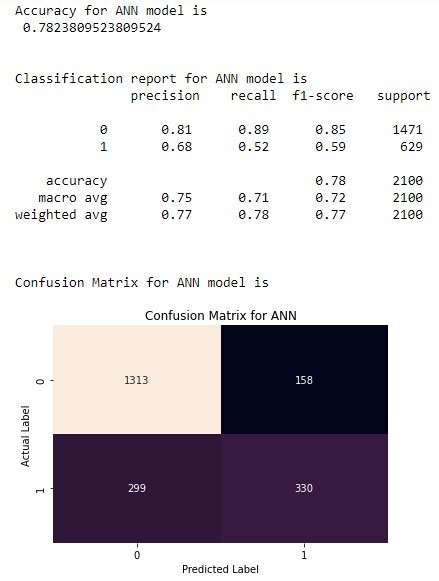
### ***2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.***

In this section, the models built are tested for its performance based on Accuracy, Confusion Matrix, ROC curve, ROC\_AUC Score and Classification report. All the models are trained using the training dataset and predictions are made using the test dataset.

**ANN**:

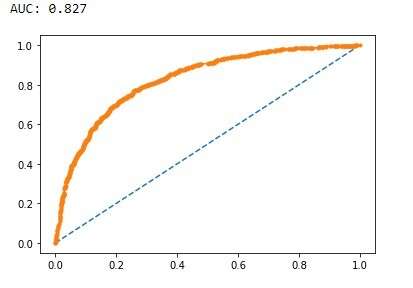
*Train dataset:*

Let us look at the various performance metrics of model trained using training dataset.



**Fig 2.16**

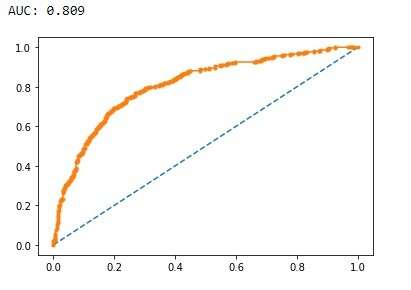
Fig 2.16, shows that the accuracy the model is 78%. Precision, recall, f1-score of the model is also found in the figure. There are 158 false positives and 299 false negative present according to the confusion matrix. The area under curve for the training model is 82.7%. (Refer fig 2.17)



**Fig 2.17**

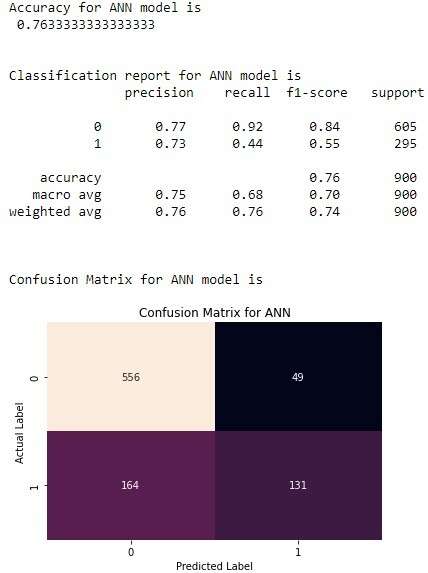
*Test dataset:*

Let us look at the various performance metrics of model trained using test dataset.



**Fig 2.18**

The area under the curve for test model is 80.9%.



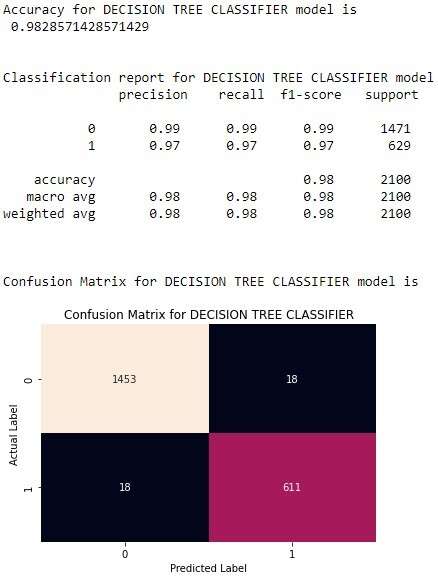
**Fig 2.19**

The accuracy of model on test data has dropped to 76.3%. There is an improvement in the FP and FN count as well. Also, you can notice that there is improvement in the recall too.

**Decision Tree**:

*Train Dataset:*

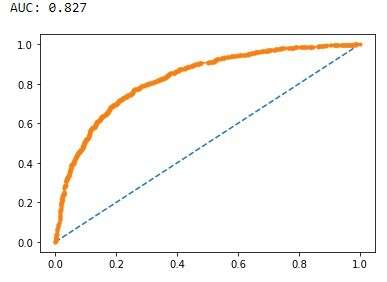
Let us look at the various performance metrics of model trained using training dataset.



**Fig 2.2**

The accuracy of the model on training data is 98.2%. This model has a good precision, recall and f1 score. FP and FN are very low when compared to ANN model. (Refer Fig 2.2)

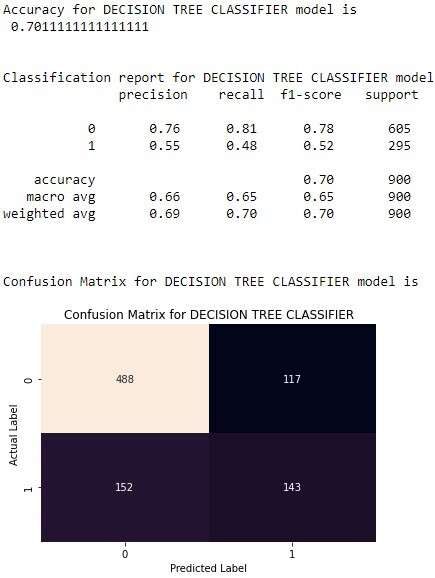
Let us, also see the ROC\_AUC curve and the AUC score to see how the model have performed. Referring Fig 2.21, the area under the curve for the model is 82.7 %.



**Fig 2.21**

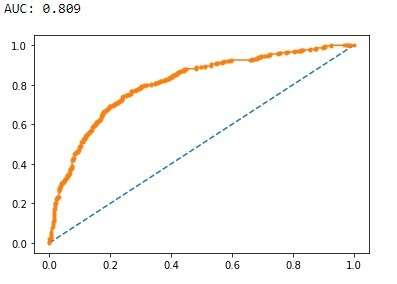
*Test Dataset:*

Let us look at the various performance metrics of model trained using test dataset.



**Fig 2.22**

The fig 2.2, shows that the accuracy of model drastically dropped to 70%. Also, the precision, recall and f1-score has also dropped along with an increased FP and FN.



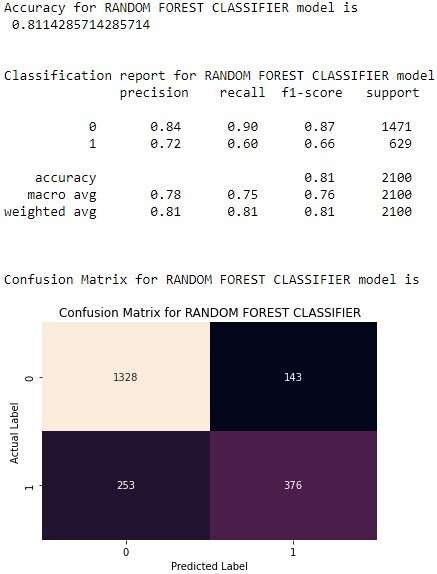
**Fig 2.23**

The area under the curve for the model on test data also dropped to 80.9. (Refer fig 2.23)

**Random Forest:**

*Train Dataset:*

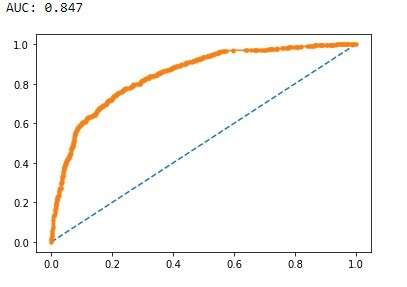
Let us look at the various performance metrics of model trained using training dataset.



**Fig 2.24**

Fig 2.24, has a detailed information about various performance metrics of the model. The accuracy of the training model is 81.1%. The precision, recall and F1-score are also visible in the figure. FP and FN present are 143 and 253 respectively.

Let us also see the AUC score and ROC\_AUC Curve for the model.

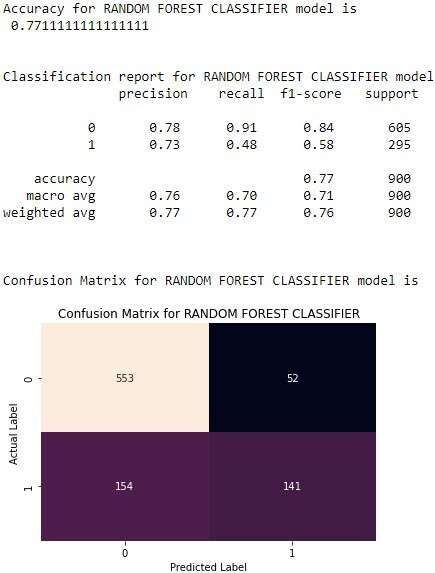


**Fig 2.25**

The auc for the model is 84.7%.

*Test Dataset:*

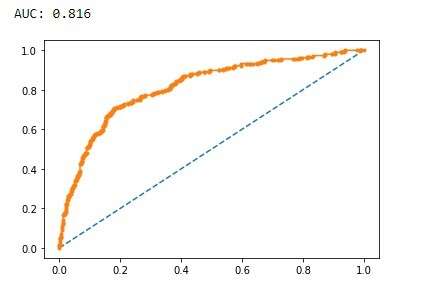
Let us look at the various performance metrics of model trained using test dataset.



**Fig 2.25**

The Accuracy of the model is 77%. There is a decrease in the FP and FN when compared to the model’s performance on the training dataset.

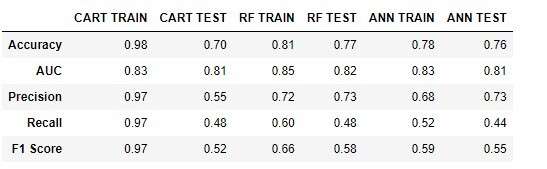
The Area under the curve for the model is 81.6%. (Refer Fig 2.26)



**Fig 2.26**

### ***2.4 Final Model: Compare all the models and write an inference which model is best/optimized.***

As, all the models (ANN, Decision Tree and Random Forest) are trained using the train dataset and predicted using the test dataset, now its time to compare all the model’s performance metrics and choose the best model for this business scenario using the table below.



**Table 2.9**

Comparing all the metrics of each models following inferences can be:

* Though the accuracy, auc, precision, recall and f1 score of decision tree(CART) model on training dataset is high, it’s performance on test dataset is so poor. So, it better to reject the model in this case.
* Secondly, both the random forest model and the neural network models have performed well on both the training as well test dataset. So, it would be a best option to choose between these two models.

### ***2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations***

Based on the whole analysis, certain important insights and recommendation are made:

* In terms of selecting the model for this business scenario, I would suggest to go with Random Forest Model. Even though the Artificial Neural Network model’s performance is close to Random Forest Classifier, the later best suits this scenario as the data we have is limited. ANN model performance is best when there are plenty of data, so it is advised not go for with less data.
* More monitoring and deep dive is need for the online claims made by Travel agency Firms in EPX. These are three which accounted for most number of claims. There are chances that there may be clerical error or wrongfully updated claim status and also fraudulent claims may also happen. So, it is necessary to channelise and monitor this agency given more importance.
* In terms of product, the customized plan and cancellation plan out performs the rest of the plans. A deep dive is needed to understand what lacks in the poorly performing plans and what attracts the customer towards the other two plans. Based on the deep dive results, we can customise gold, silver and bronze plan to perform better than its past performances.