**Business Report on**

***Data Mining***

***Submitted to***



**Great Learning Olympus**

***Prepared by***

**Mrinmoy Majumdar**

**June-D Batch**

**Post Graduate Program in Data Science & Business Analytics**

**From**

****

UT Austin

**October, 2021**

TABLE OF CONTENT

**Problem 1-Clustering ……………………………………………....4**

**1.** Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)............................................4

**2.**Do you think scaling is necessary for clustering in this case? Justify……..13

**3.**Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them………………………..14

**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…………….15

**5.** Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters…………………………………..17

**Problem 2-CART-RF-ANN………………………………………...18**

**1.** Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).............................................18

2. Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network........................................................30

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.……………………..33

4. Final Model: Compare all the models and write an inference which model is best/optimized.………………………………………………………………....40

5. Inference: Based on the whole Analysis, what are the business insights and recommendations................................................................................................41

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **Images** | **Page No** |
| **Info and null value checking** | **6** |
| **Histogram plot** | **7-9** |
| **Box plot** | **10-11** |
| **Skewness** | **12** |
| **Heat map** | **12** |
| **Pair plot** | **13** |
| **Variance pre and post scaling** | **14** |
| **Dendrogram** | **15** |
| **Cluster output for K-Means** | **16** |
| **Point plot** | **17** |
| **WSS plot** | **17** |
| **Datatype information and null count** | **20** |
| **Numerical box plot** | **21** |
| **Histogram, box plot and skewness** | **22-23** |
| **Bar plot** | **23-25** |
| **Correlation plot** | **26** |
| **Categorical vs categorical count plot** | **26-27** |
| **Categorical vs numeric box plot** | **28** |
| **Facet Grid** | **31** |
| **Datatype changes post encoding** | **31** |
| **Decision Tree** | **32** |
| **Feature importance** | **33** |
| **ROC curve** | **34** |
| **Confusion Matrix** | **35** |
| **Classification report** | **35-36** |

**LIST OF TABLES**

|  |  |
| --- | --- |
| Table 1 | 4 |
| Table 2 | 6 |
| Table 3 | 13 |
| Table 4 | 13 |
| Table 5 | 15 |
| Table 6 | 15 |
| Table 7 | 16 |
| Table 8 | 17 |
| Table 9 | 17 |
| Table 10 | 18 |
| Table 11 | 20 |

**Problem 1-Clustering**

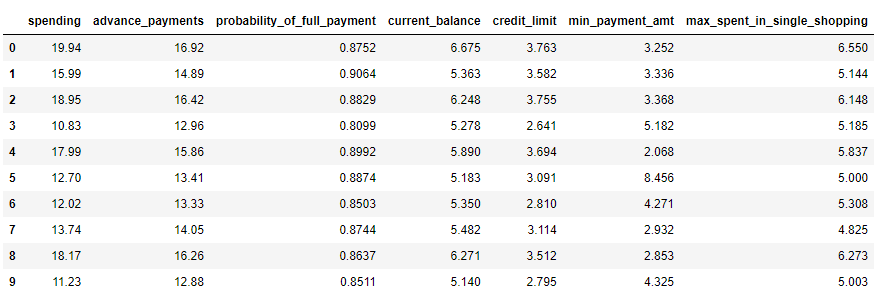
**Problem Statement:**

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.

# Attribute Information:

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

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

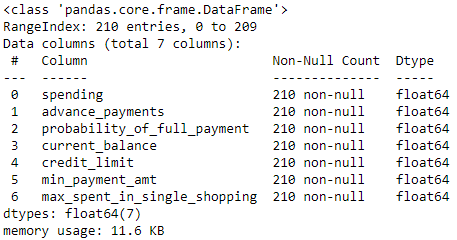
Displaying Marketing Data:

**Table 1: Top 10 rows of marketing data Frame**

## Basic EDA:

* Checking shape and information of data Frame

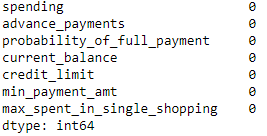
(210,7) – The data set contains 210 observations of data and 7 variables.



**Image 1: Information on marketing dataset**

### The data has 210 instances with 1 attribute that is of float type.

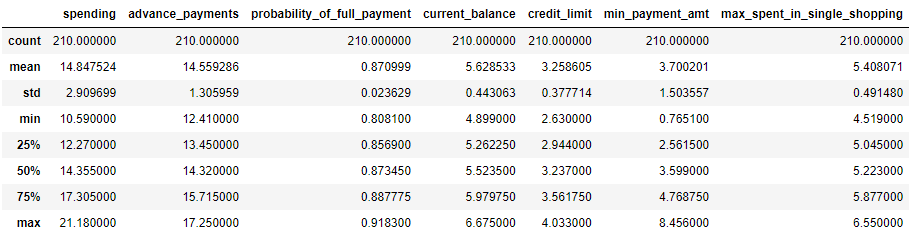
* Check the presence of missing values



**Image 2: Checking null values in data**

### There are no null values in any of the columns. All the variables in the dataset are continuous. There are no duplicate values in the dataset.

* Checking summary of data Frame



**Table 2: Description of marketing dataset**

Looking at the 5 point summary, we can probably conclude that data is normally distributed as the mean and median of the columns are almost identical. The claims can be further solidified with the help of univariate, bivariate and multivariate analysis of feature columns along with its associated skewness.

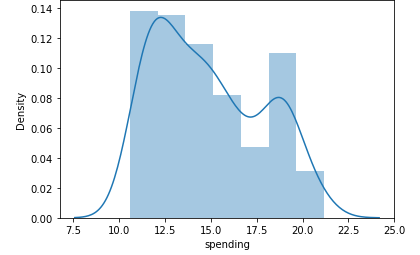
The minimum amount of spending per month is 10.59 and maximum spending amount is 21.18.

Average amount paid by the customer in advance is 12.41.

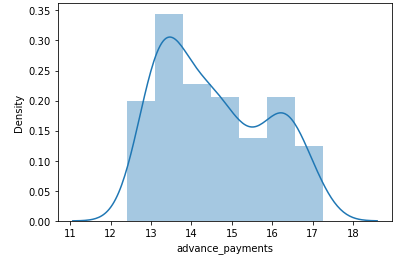
Maximum probability of full payment by the customer is approximately 92% which is quite impressive.

The standard deviation of spending variable is higher compared to other variables.

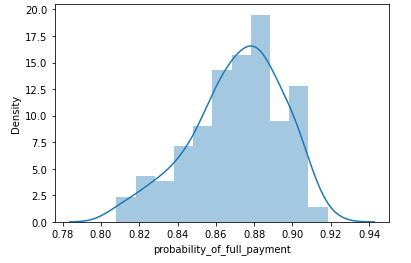
**Univariate Analysis**



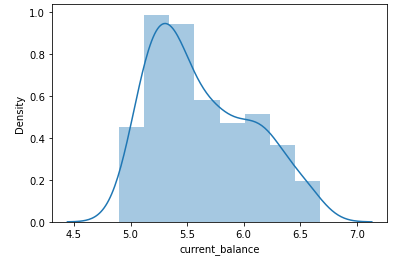
**Image 3: Histogram plot on spending**



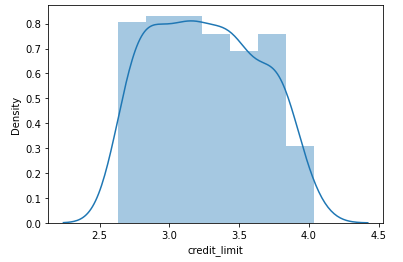
**Image 4: Histogram plot on advance payments**



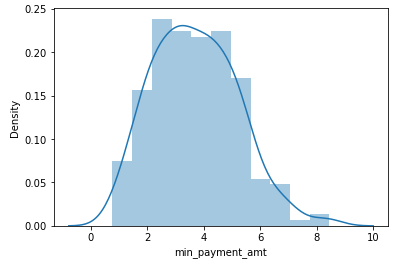
**Image 5: Histogram plot on probability of full payment**



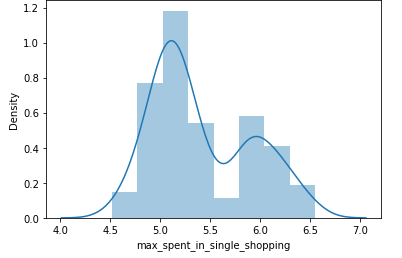
**Image 6: Histogram plot on current balance**



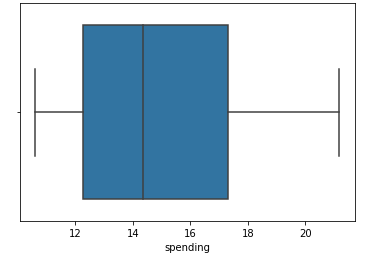
**Image 7: Histogram plot on credit limit**



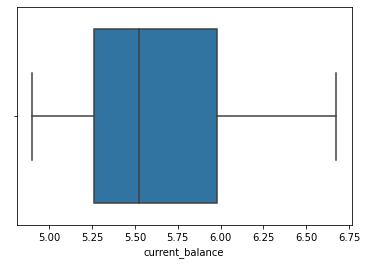
**Image 8: Histogram plot on minimum payment amount**



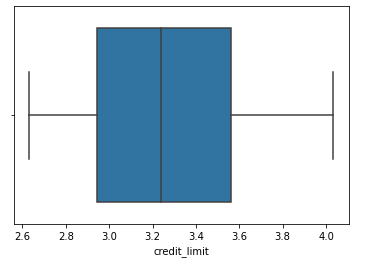
**Image 9: Histogram plot on maximum spent in single shopping**



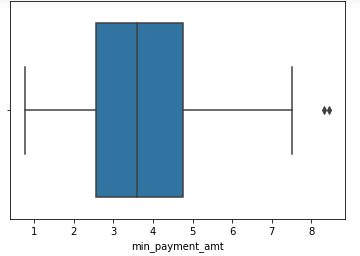
**Image 10: Boxplot on spending**



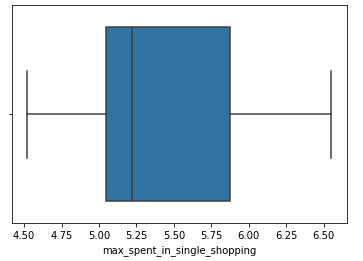
**Image 11: Boxplot on current balance**



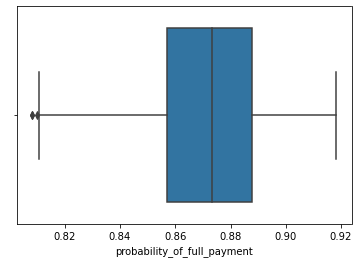
**Image 12: Boxplot on credit limit**



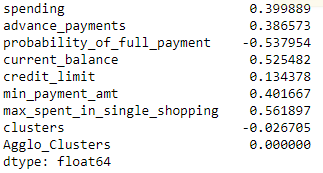
**Image 13: Boxplot on minimum payment amount**



**Image 14: Boxplot on maximum spent in single shopping**



**Image 15: Boxplot on probability of full payment**



**Image 16: Skewness of columns in dataset**

**Inferences-**

* Spending, advance payments, current balance, credit limit, minimum payment amount and maximum spent in single shopping are positively skewed.
* Probability of full payment is negatively skewed.
* Minimum payment amount and probability of full payment are the only variable with outliers.

**Multivariate Analysis**

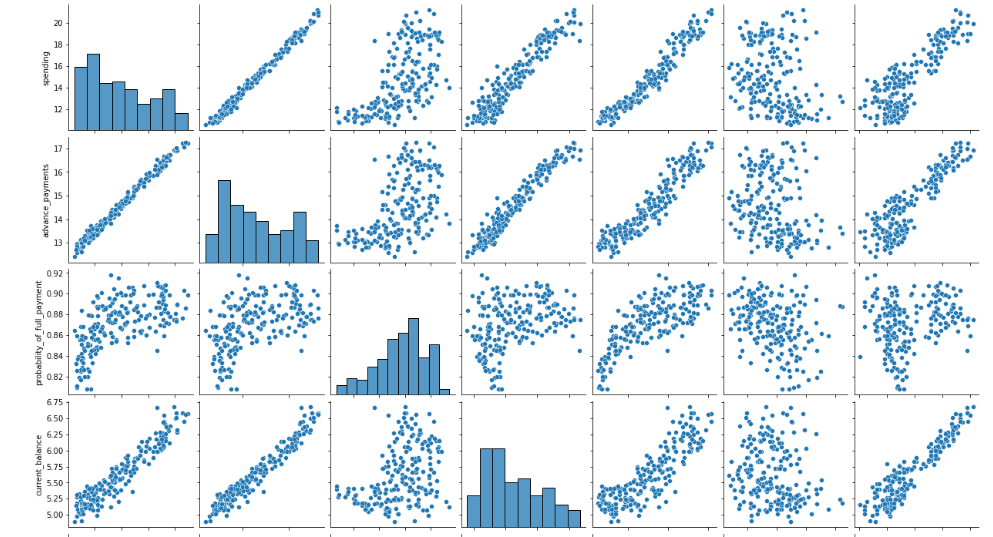


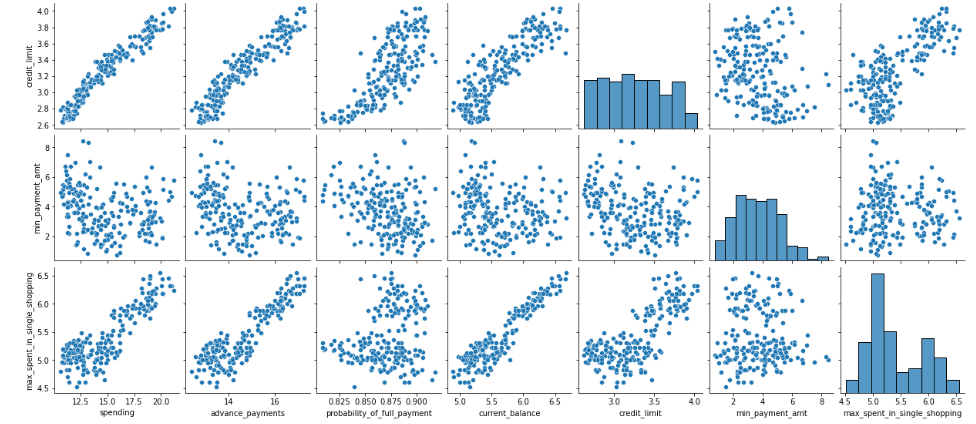
**Image 17: Heat map of marketing data**

**Observations**

We can see high positive correlation among following variables –

* Spending and advance payments
* Current balance and spending
* Spending and credit limit
* Advance payments and credit limit
* Maximum spent in single shopping and current balance
* Current balance and advance payments



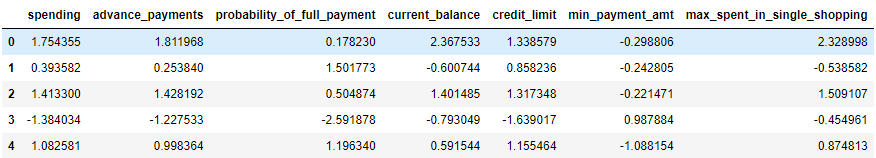


**Image 18: Pair plot of marketing dataset**

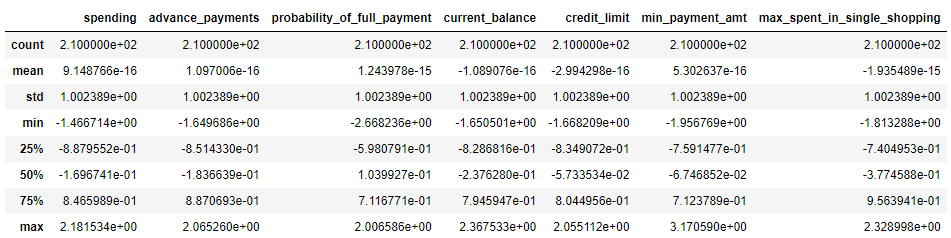
There is positive linear relationship between spending and advance payments.

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

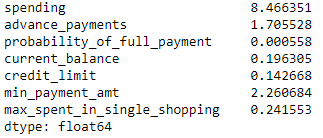
Scaling is required for clustering as it is a distance based algorithm. Besides that, standard deviation and variance of columns are different which indicates unscaled data. The data dictionary also indicates that all variables are of different dimensions. Standard scaler method in python is used to scale the given dataset which is based on z-score scaling technique. The mean will be zero and variance will be 1 post scaling the data.



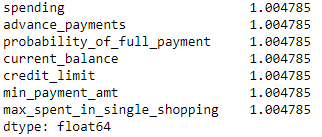
**Table 3: Dataset post scaling**



**Table 4: Description of scaled data**



**Image 19: Variance of data pre-scaling**

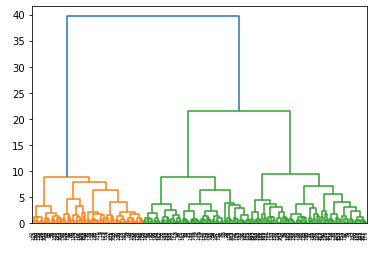


**Image 20: Variance of data post-scaling**

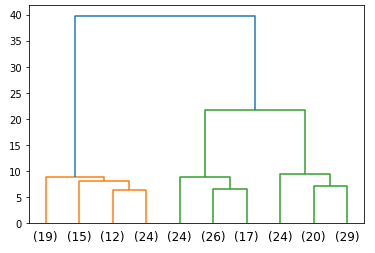
**Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them**

Hierarchical clustering is a type of clustering approach where records are sequentially grouped to create clusters based on distance between records and distance between clusters. It also produces a useful graphical display of the clustering process and results called dendrogram.

We choose ward linkage method post scaling. This method is similar to group average and centroid distance. It joins records and clusters together progressively to produce larger and larger clusters, but operates slightly differently from general approach.

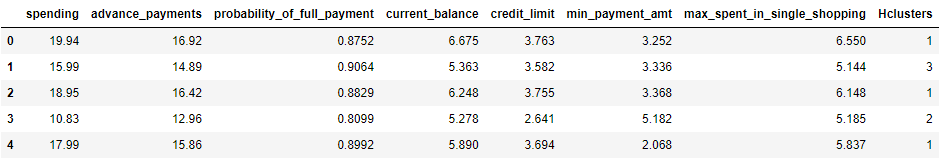


**Image 21: Dendrogram of all data points**

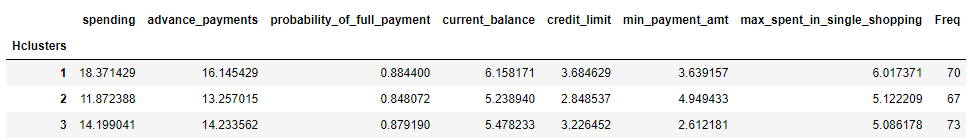


**Image 22: Dendrogram of last 10 merge data points**

We have cut the dendrogram with 2 suitable clusters. F-cluster technique is used to append the clusters to original dataset. The optimum number of clusters chosen is 3 comparing criterions maxclust and distance with values as 3 and 21 respectively and affinity by default is Euclidean which turns out to be identical values.



**Table 5: Marketing dataset segmented into clusters**



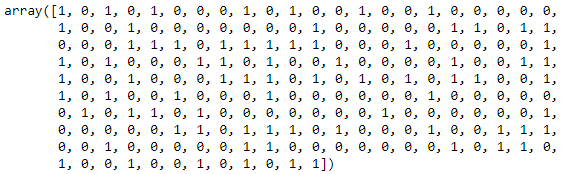
**Table 6: Hierarchical Cluster Profile**

**Inferences -**

* Cluster 1 : Customers with high credit card usage
* Cluster 2 : Customers with low credit card usage
* Cluster 3 : Customers with medium credit card usage

**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**

K-Means clustering is an unsupervised learning algorithm whose goal is to find groups or assign the data points to clusters on the basis of their similarity.



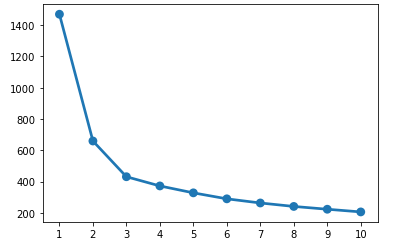
**Image 23: Cluster Output for all observations**

Elbow method is the most popular and well known method to find optimal number of clusters or value of k in process of clustering. Setting different values of number of clusters with random state assumed to be 1, we compute and compare within cluster sum of squares for different values of K.

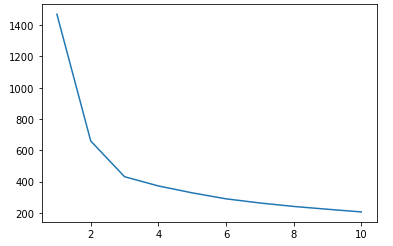
|  |  |
| --- | --- |
| K-value | Within cluster sum of squares |
| 1 | 1469.99 |
| 2 | 659.17 |
| 3 | 430.65 |
| 4 | 371.30 |
| 5 | 327.96 |
| 6 | 290.59 |

**Table 7: WSS for varied values of K**

WSS reduces as K keeps increasing. The optimal number of clusters selected is 3 using K-elbow method as there is no drastic drop in the value of WSS after K=3.



**Image 24: Point plot of 10 values of K against inertia**



**Image 25: WSS plot**

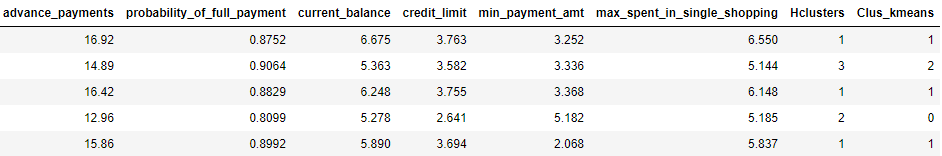
WSS plot helps to know how many clusters are needed as output in k-means clustering.

Silhouette score is an indirect model evaluation technique which helps to analyse whether each and every observation which is mapped to respective clusters based on distance criteria is correct or not.

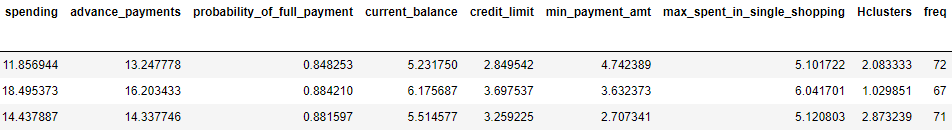
Silhouette score and width are positive for K=3. Hence, we can conclude that clusters have been mapped correctly to its centroid.

Silhouette score for K=3 is greater than K=4.Hence, final cluster is considered as 3.

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



**Table 8: Clusters appended to dataset**



**Table 9: K-means cluster profiling**

**Inferences –**

* Cluster 0: Low credit card users
* Cluster 1: High credit card users
* Cluster 2: Medium credit card users

**Some Recommendations**

1. The customers in Cluster 2 are using credit cards to make moderate purchases and paying their dues on time. Banks need to provide priority to such customers and give them promotional offers that will lure them into spending more than they usually do. Cards with discount on use at renowned websites can encourage them to prefer using cards instead of other modes of transaction.
2. Customers in Cluster 1 seems to prefer payment through credit cards as they record the highest spending. Banks can offer reward points and EMI against cards for huge transactions for this category of customers. Banks can tie up with luxury brands to increase sales on expensive products and credit card use.
3. Customers in Cluster 0 need to be given offers on early payment of bills as they are low spending group. The banks need to provide rebates on daily requirement essentials to transform these customers into high spending ones.

**Problem 2: CART-RF-ANN**

**Problem Statement:**

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.

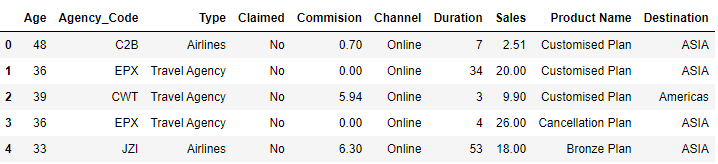
Attribute Information:

1. Target: Claim Status (Claimed)   
2. Code of tour firm (Agency\_Code)   
3. Type of tour insurance firms (Type)   
4. Distribution channel of tour insurance agencies (Channel)   
5. Name of the tour insurance products (Product)   
6. Duration of the tour (Duration in days)   
7. Destination of the tour (Destination)   
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)   
9. The commission received for tour insurance firm (Commission is in percentage of sales)   
10.Age of insured (Age)

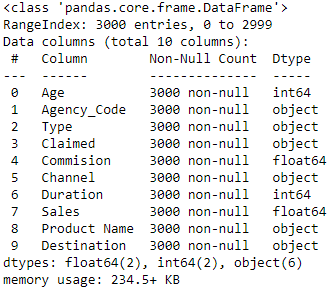
**Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).**

In this step, we will perform the below operations to check what the data set comprises of. We will check the below things:

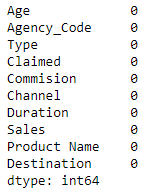
* head of the dataset
* shape of the dataset
* info of the dataset
* summary of the dataset
* presence of duplicates
* outlier proportion



**Table 10: Head of insurance dataset**



**Image 26: Information on datatypes**



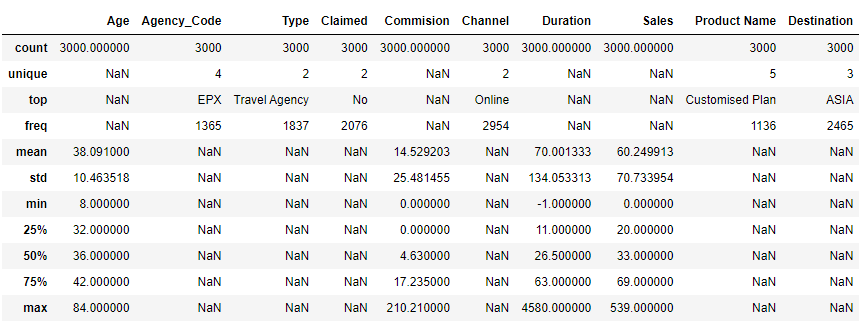
**Image 27: Count of null values in each column of dataset**

Observations-

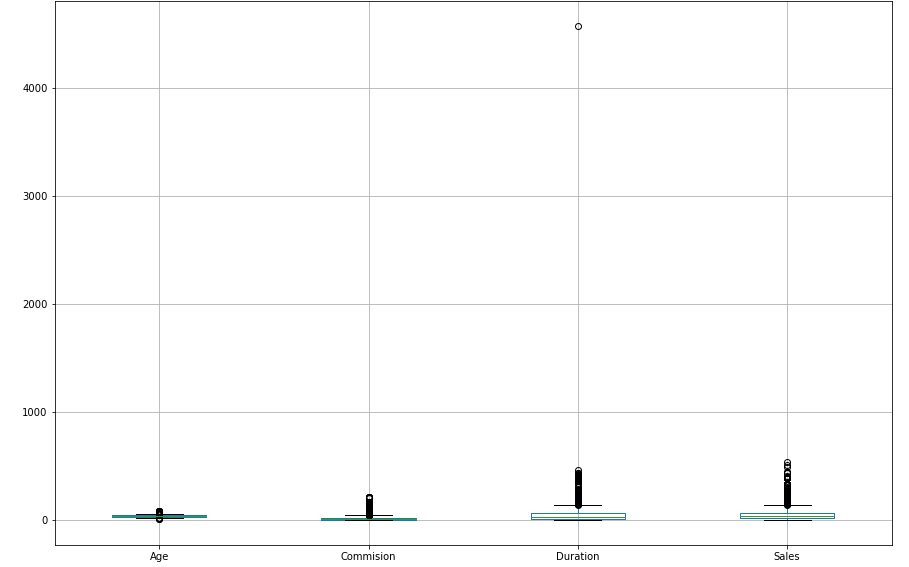
* (3000, 10)-The dataset contains 3000 observations of data and 10 variables.

### The data has 3000 instances with 10 attributes. 2 integer type, 2 float type and 6 object type (Strings in the column).

* There are no null values in the given dataset.
* There are 139 duplicate values in dataset.
* There is no unique column to identify which customer is provided with duplicate products, so duplicates are not eliminated from dataset.
* The data is not normally distributed as mean and median are not close to each other for all columns except age as shown in five point summary of dataset in figure below.
* ASIA is the most preferred destination of travelling customers.
* Customised Plan is the most frequent product offered to customers.
* Customers prioritise online channel and travel agency while selecting insurance firms.
* There is presence of outliers in the dataset as shown in boxplot below.

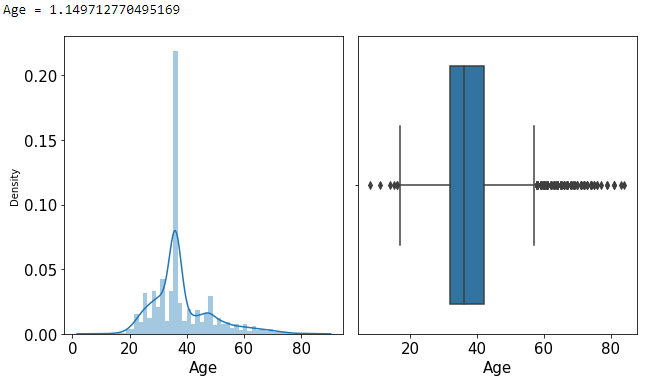


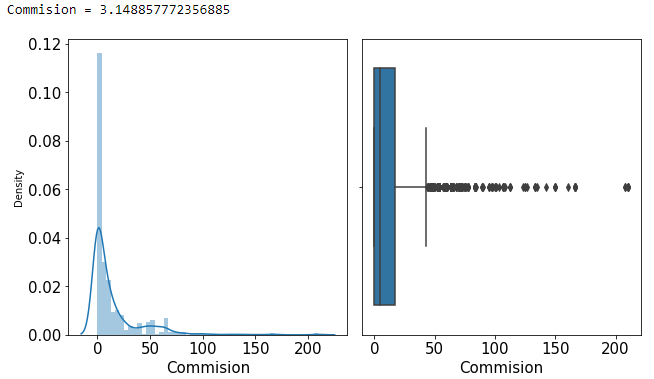
**Table 11: Descriptive summary of dataset**

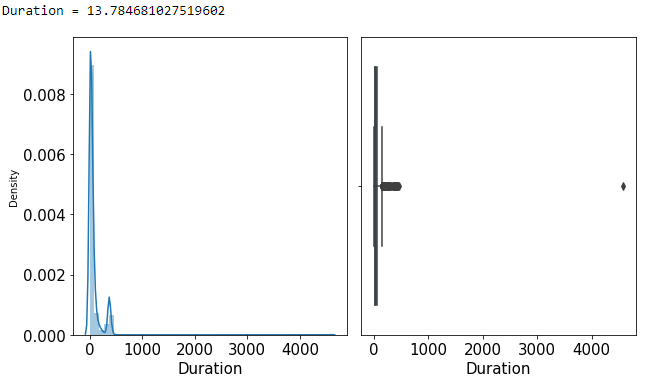


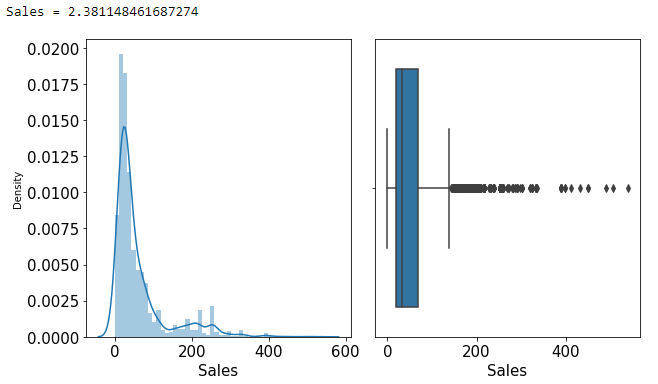
**Image 28: Boxplot of all numerical features**

**Univariate Analysis-**





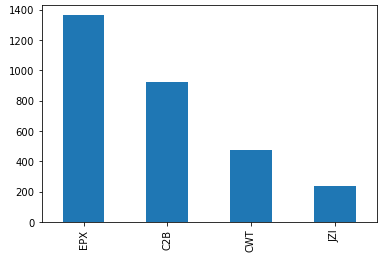




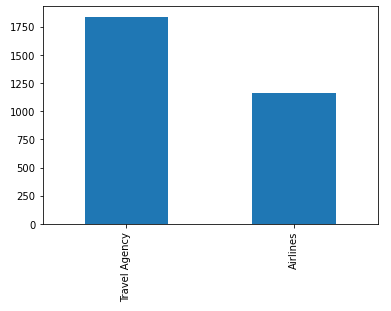
**Image 29: Histogram boxplots and skewness of mentioned columns**

#### Insights:

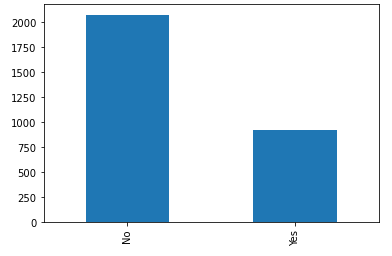
Age, commission, duration and sales are numerical columns having positive skewness and outlier presence.



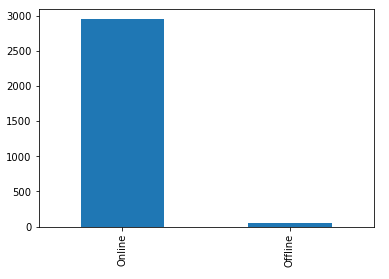
**Image 30: Bar plot on agency code**



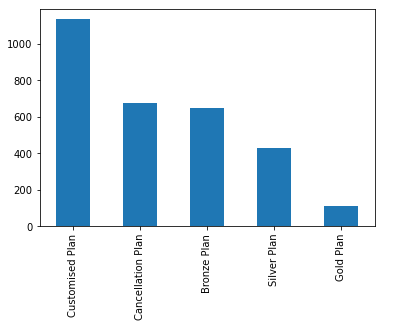
**Image 31: Bar plot on type column**



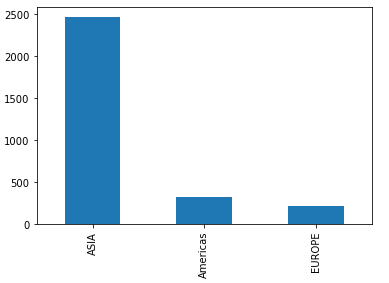
**Image 32: Bar plot on claimed column**



**Image 33: Bar plot on channel variable**



**Image 34: Bar plot on product name**

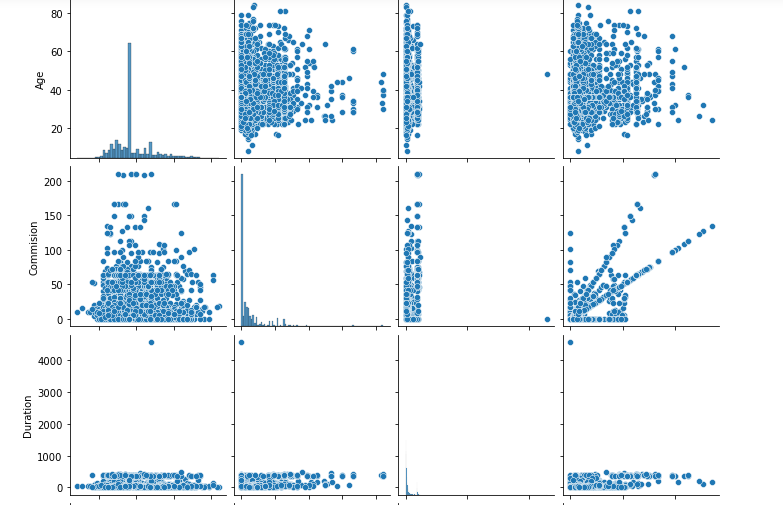


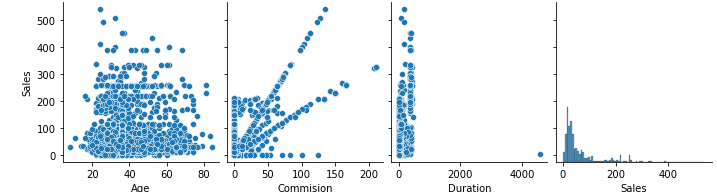
**Image 35: Bar plot on destination feature**

#### Insights:

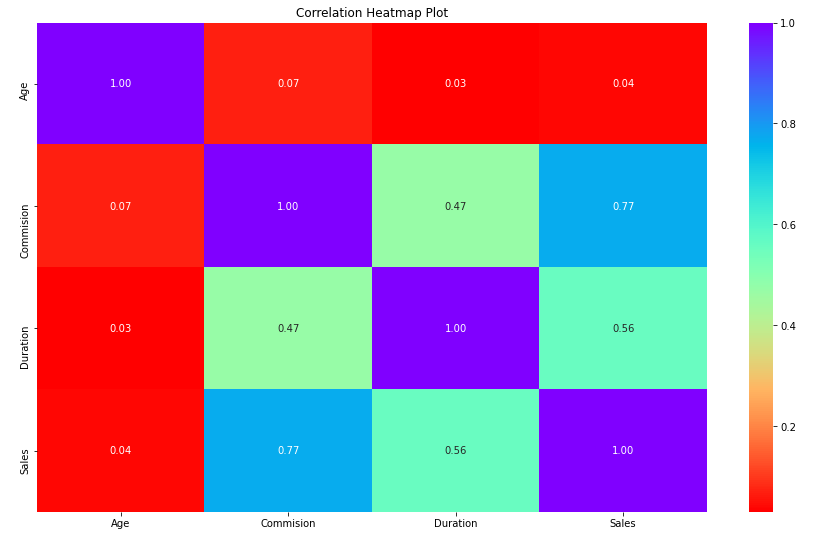
* Agency code, type, claimed, channel, product name, destination are 5 categorical columns depicts univariate analysis with help of bar plots.
* There are 4 agency codes with EPX and JZL having maximum and minimum occurrence.
* Travel agency and airlines are types of insurance firms available for customer service.
* Customers prefer to settle claims on online channels compared to offline mode.
* ASIA has the most customers visiting for travel purpose and Europe has the least travellers landing on its soil.

Bivariate Analysis

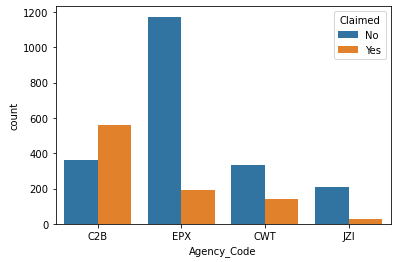


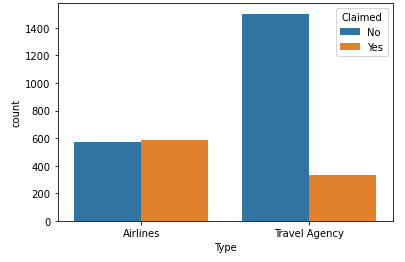


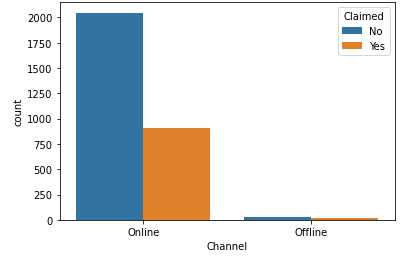
**Image 36: Pair plot of numerical column combination**

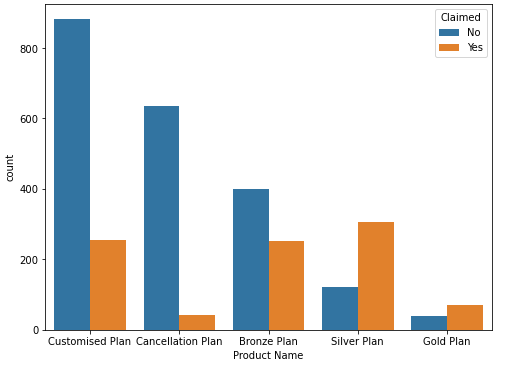


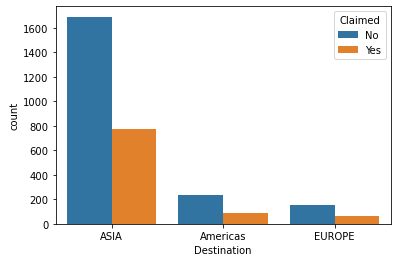
**Image 37: Correlation plot**



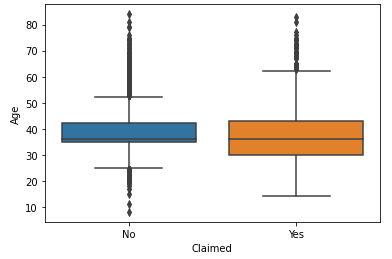


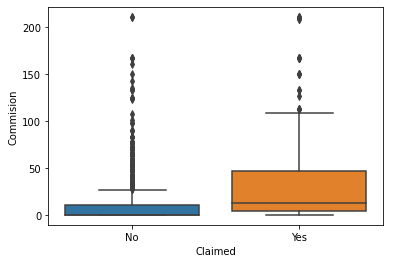


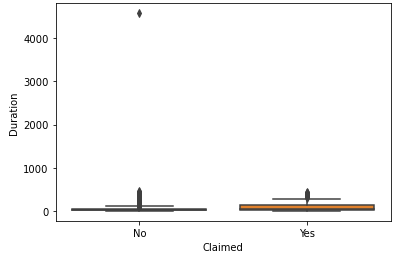


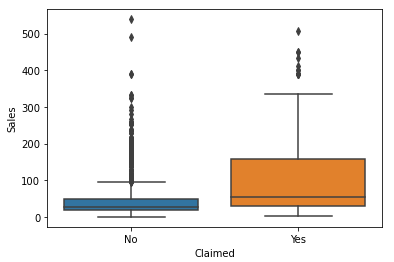


**Image 38: Categorical vs categorical count plot**







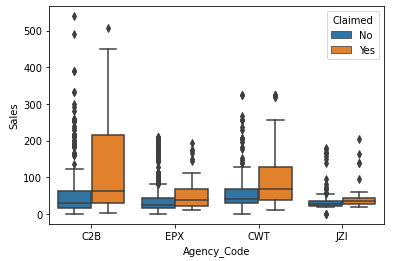


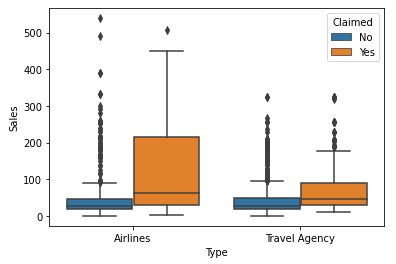
**Image 39: Categorical vs numeric box plot**

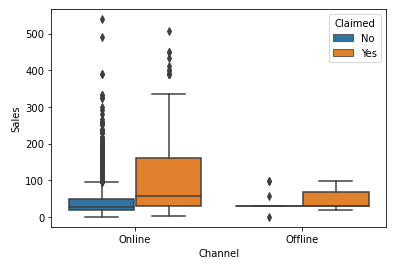
#### Insights:

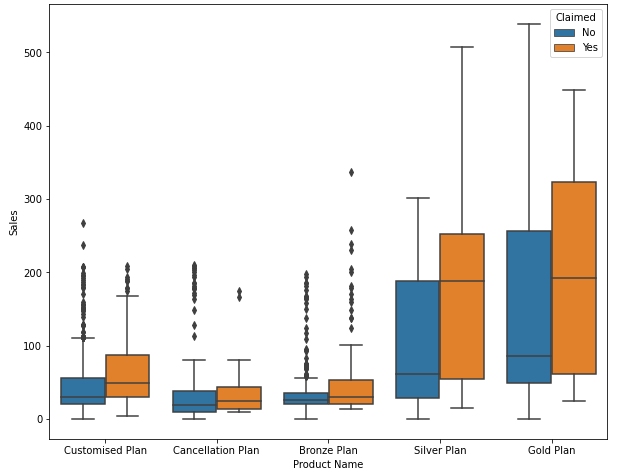
* All numerical columns have positive correlation with each other.
* C2B agency code have highest claims recorded.
* Airlines type have almost same number of claim status.
* Online channels have high number of claims with status as ‘No’.
* Silver plan product received maximum claims.
* ASIA travel destination have customers with more claim status as ‘No’ than ‘Yes’.
* Customers with/without claims have nearly identical median age.
* Commission received and amount of sales per customer have high number of claims insured.
* Duration of tour is independent of claims insured.

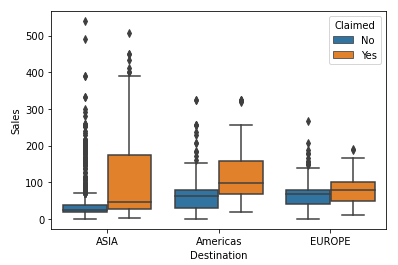
**Multivariate Analysis-**



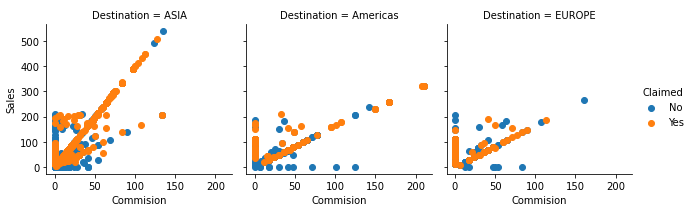








**Image 40: Box plot showing multivariate combinations**



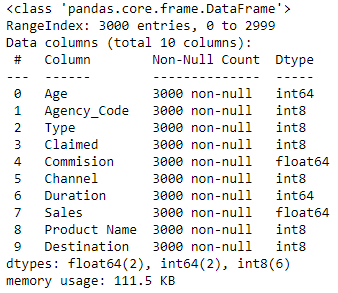
**Image 41: Facet Grid**

### Insights-

* It is observed that C2B agency code have highest median sales per customers insured as compared to other codes.
* Airline type insurance firm has maximum median sales.
* Americas has highest median sales in terms of destination travelled.
* Proportion of customers who have claimed and not claimed are very different independent of destination.

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

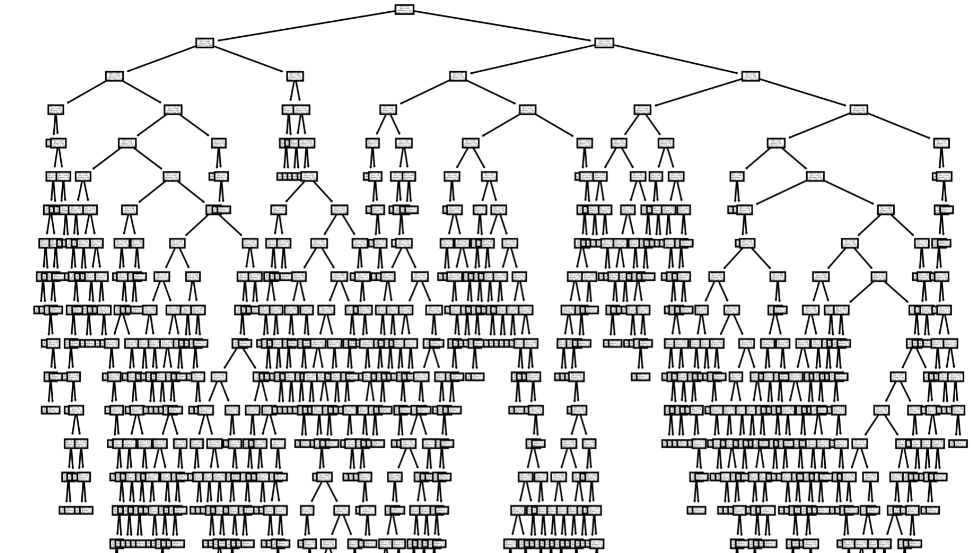
Before splitting the data into test and train, the object datatype is converted into numerical datatype to fit in the models. The dataset has been split into 30% testing and 70% training data. Random state has been considered as 1.Claimed is the target variable given upon which the model is designed.



**Image 42: Datatype changes post encoding**

**70% customers do not have claims while the remaining 30% have claims in the insurance firm.**

**CART model is built with the help of decision tree classifier assuming criterion to be gini so that target variable claim can have maximum purity.**



**Image 43: Decision Tree**

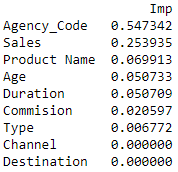
**Grid Search is applied on CART model. Hyper parameters considered for decision tree are as follows:**

* **Maximum depth is considered in an inclusive range of 7 to 10 which is the maximum number of nodes from root to leaf.**
* **Minimum samples per split is assumed to in a range of 45 to 75 with intervals of 15 between them which prevents inequality between sample count on each node and waste of resource and time that comes along with it.**
* **Minimum samples per leaf is calculated by dividing each element of samples per split by 3 which turns out to be [15, 20, 25]**

**Cross validation is assumed to be 10.Higher the value more the time it takes for the model to execute. It helps address overfitting and overcomes greedy algorithm.**

**The model is built on best parameter -** {'max\_depth': 7, 'min\_samples\_leaf': 25, 'min\_samples\_split': 60}.

According to cart model, agency code is the most important feature which impacts claimed variable.



**Image 44: Feature importance based on CART model**

**Similarly, for Random Forest model, hyper parameters are considered as below –**

* **Maximum depth = 10**
* **Maximum features = 6 as seen above there are many features to build a decision tree. When splitting, in every split, we have to check entire dataset on each of the features which can be very expensive. Limiting the number of features speeds up calculations significantly.**
* **Minimum samples per leaf = 10**
* **Minimum samples per split = 50**
* **Number of estimators = 300**

**Random forest gives the optimal solution compared to Cart model. Model is built on the below best parameters –**

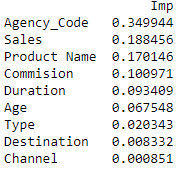
{'max\_depth': 10,

'max\_features': 6,

'min\_samples\_leaf': 10,

'min\_samples\_split': 50,

'n\_estimators': 300}



**Image 45: Feature importance based on Random Forest model**

**Agency code again is the most important feature affecting target variable in case of random forest model.**

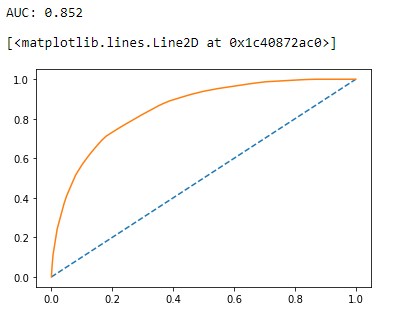
**Hyper parameters for Artificial Neural Network model is shown as below –**

* **Number of hidden layer sizes – 100**
* **Maximum iterations – 2500**
* **Solver – adam**
* **Learning Rate – 0.01**

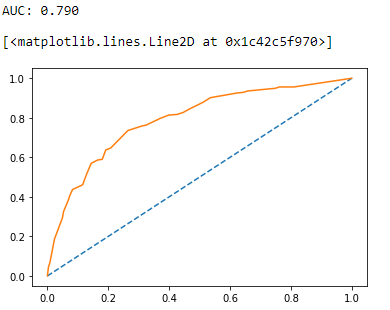
**The model is built on the above best parameters.**

**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.**

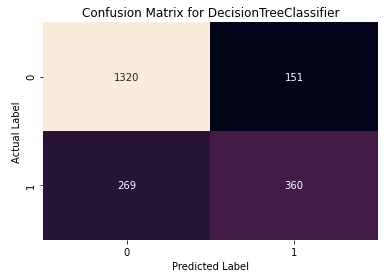
CART model –



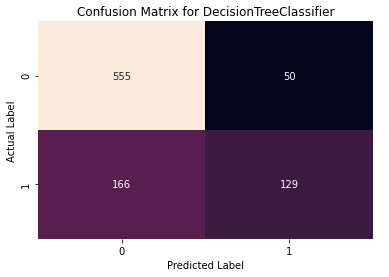
**Image 46: ROC curve for CART train data**



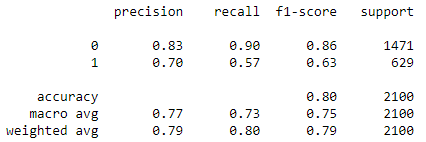
**Image 47: ROC curve for CART test data**



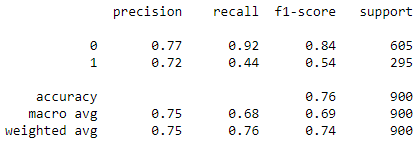
**Image 48: Confusion matrix for CART training data**



**Image 49: Confusion matrix for CART testing data**



**Image 50: Classification report for CART training data**



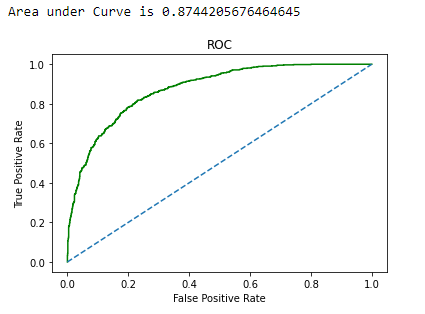
**Image 51: Classification report for CART testing data**

***Observations –***

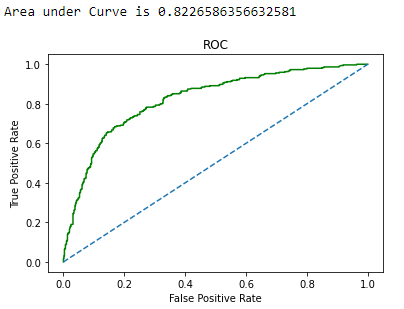
* **ROC-AUC score of train and test data are 85% and 79% respectively.**
* **There are two types of classes in hand – positive and negative as indicated by 1 and 0 respectively in confusion matrix.**
* **True positive and true negative combined is higher than false positive and false negative together for both train and test data which indicates correct predictions.**
* **Accuracy for both train and test data are close to each other. Recall has more importance than precision and f1 score.**
* **Precision is higher than recall for both train and test data as type-II error are lesser than type-I error as indicated by false negative and false positive in confusion matrix.**

**Recall is specified by type-II error and precision is signified by type-I error.**

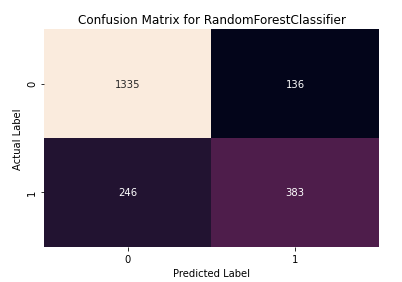
**Random Forest –**



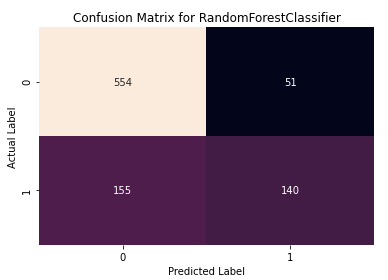
**Image 52: ROC curve for Random Forest training data**



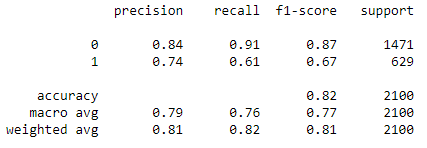
**Image 53: ROC curve for Random Forest testing data**



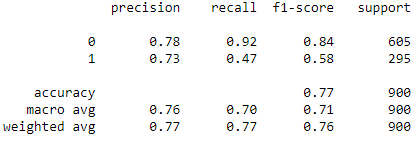
**Image 54: Confusion Matrix for Random Forest training data**



**Image 55: Confusion Matrix for Random Forest testing data**



**Image 56: Classification report for Random Forest training data**

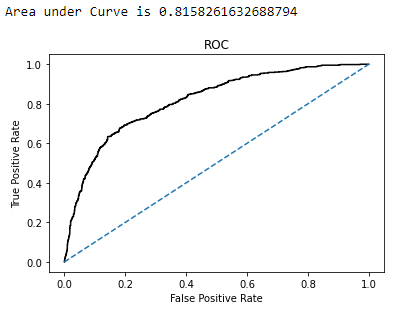


**Image 57: Classification report for Random Forest testing data**

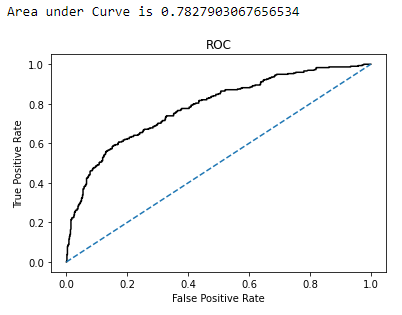
***Observations –***

* **ROC-AUC score of train and test data are 87% and 82% respectively.**
* **Accuracy for train and test data are 82% and 77% respectively which is good.**
* **Precision, recall and f1-score of trained data is near to tested data which represents optimal solution.**
* **False positive and false negatives were reduced in test data as compared to train data as shown in confusion matrix.**

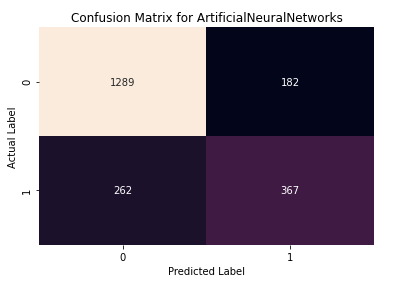
**Artificial Neural Networks –**



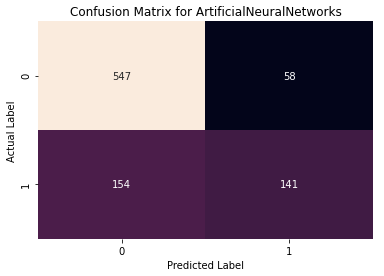
**Image 58: ROC curve for Artificial Neural Network training data**



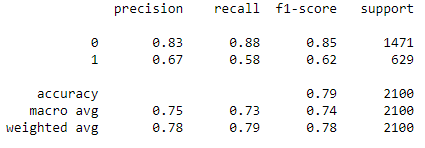
**Image 59: ROC curve for Artificial Neural Network testing data**



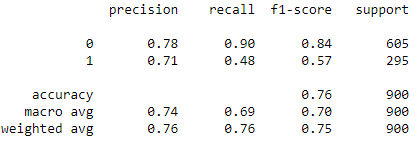
**Image 60: Confusion Matrix for Artificial Neural Network training data**



**Image 61: Confusion Matrix for Artificial Neural Network testing data**



**Image 62: Classification Report for Artificial Neural Network training data**



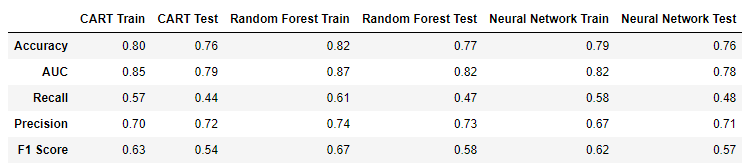
**Image 63: Classification Report for Artificial Neural Network testing data**

***Observations –***

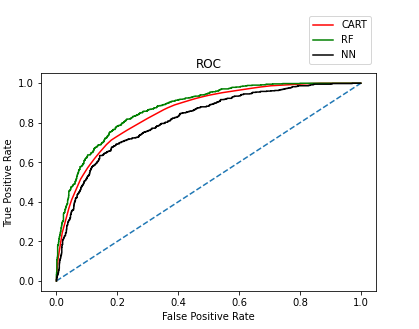
* **ROC-AUC score of train and test data are 81% and 78% respectively.**
* **Accuracy for train and test data are 79% and 76% respectively.**
* **Precision, recall and f1-score of trained data are 67%, 58% and 62% respectively.**
* **Precision, recall and f1-score of tested data are 71%, 48% and 57% respectively.**
* **Precision was better for tested data than trained one.**

**All models are under fit as the accuracy score of train and test data are less than 1.0.**

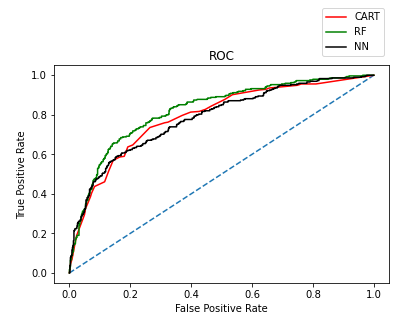
**Final Model: Compare all the models and write an inference which model is best/optimized.**



**Table 12: Comparison table of different models based on metrics**



**Image 64: ROC curve comparing models based on training data**



**Image 65: ROC curve comparing models based on testing data**

**Random forest is the final selected optimized model for the business problem in hand on the basis of performance metrics as shown in table above. It has better performance than CART and Neural Network models. The variable agency code is found to be the most useful feature amongst all other features for predicting claim status.**

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

***Insights and Recommendations* –**

* **The model performance can be improved by working around number of estimators, different combinations of grid search parameters or increasing number of layers.**
* **Insurance claims are made mostly in online mode. Offline mode should also have prevalence as it can prevent fraud claims.**
* **Airlines category have highest claims which the insurance firm needs to look into and try to resolve issues so that claims can be reduced in the future.**
* **Other agency codes need to be trained to attract more travellers to buy insurance from their firms which leads to higher profits from an organisational standpoint.**
* **Feedback needs to be taken from customers buying silver plan on the drawbacks of other plans and their suggestions on areas of improvement which will lead to further planning and action from the firm.**