**PROJECT REPORT ON DATA MINING**

**Akshaya Parthasarathy**

**Batch: PGPDSBA\_online\_July E 2020**

**Problem 1:**

**Clustering**

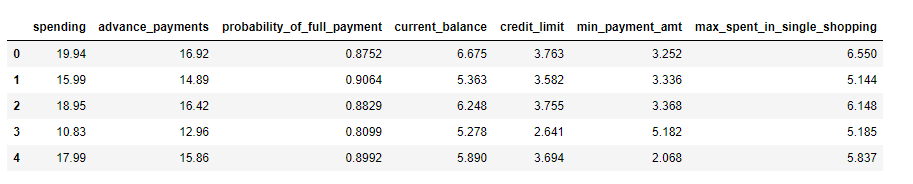
**Problem statement:**

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.

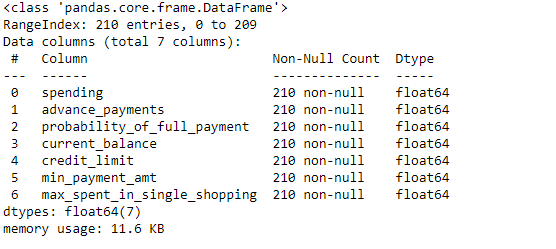
**1.1 Read the data and do exploratory data analysis. Describe the data briefly.**

**Exploratory Data Analysis:**

**Head of the dataset:** Verify whether the dataset is loaded correctly

****

**Information of the dataset:** There are seven variables in the dataset all of which are of float type.

****

**Shape of the dataset:**

There are 210 rows and 7 columns in the Fever dataset.

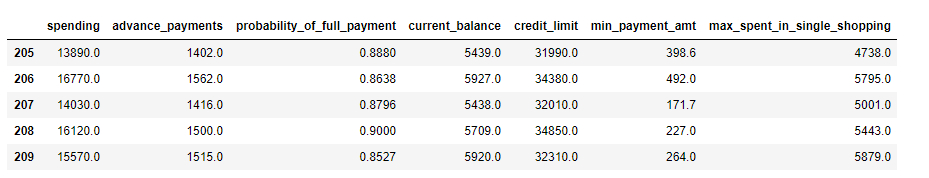
**Converting the columns to the original scale in order for better interpretation and comparison:**

* spending: Amount spent by the customer per month (in 1000s)
* advance\_payments: Amount paid by the customer in advance by cash (in

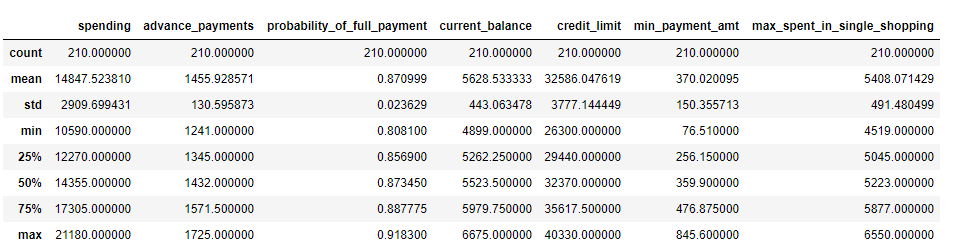
100s)

* 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)

**Tail of the dataset for verifying the columns:**

****

**Summary statistics of the dataset:**

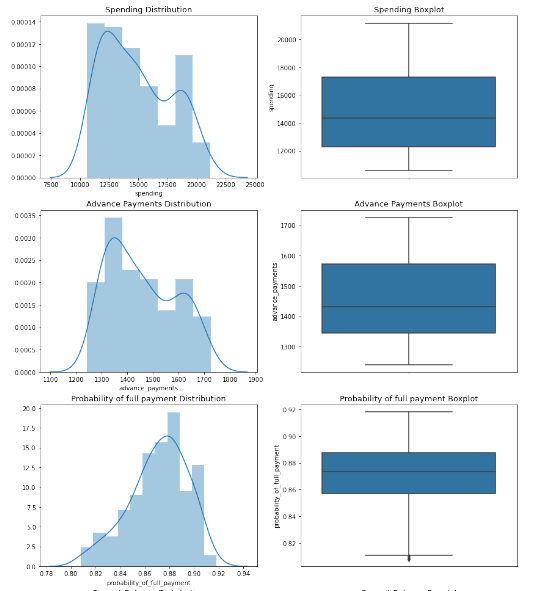


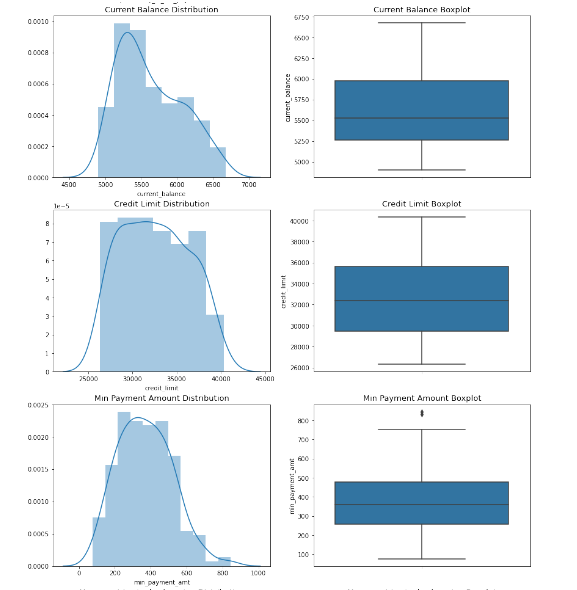
From the summary statistic, we can observe that the data does not contain any skewness since the maximum value is within the range and there are no anomalies. Scaling is required for the data since the variables are of different magnitude (100s, 1000s, 10000s)

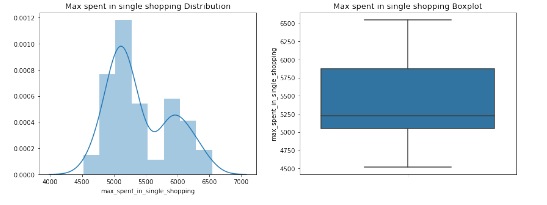
**Missing values/Duplicate in the dataset:**

There are no missing values or duplicates in the given bank dataset.

**Univariate Analysis of the columns:**

****

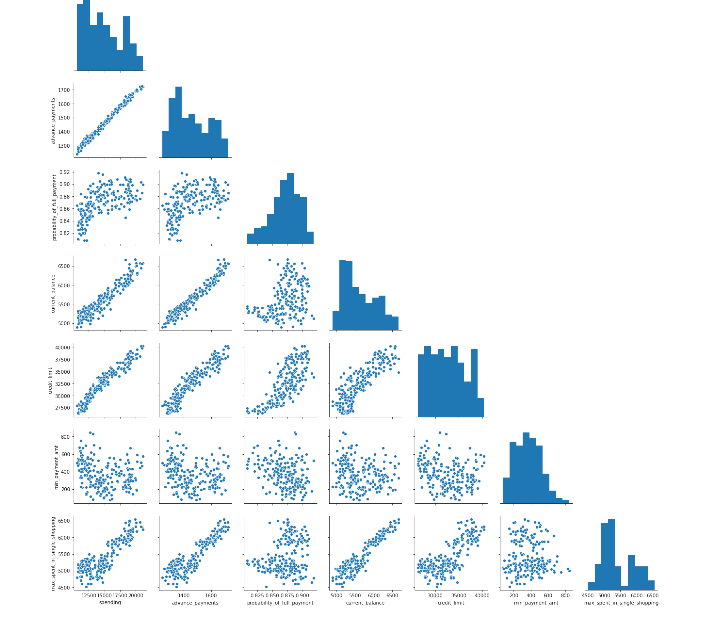
****

****

**Inference of the univariate plots:**

* From the boxplots, other than Probability of full payment and minimum payment amount, rest of the columns have no outliers.
* Credit limit variable almost follows a uniform distribution.
* Spending and Advance payments distribution follows normal distribution.
* Current balance, probability of full payment and minimum payment amount has some skewness present.

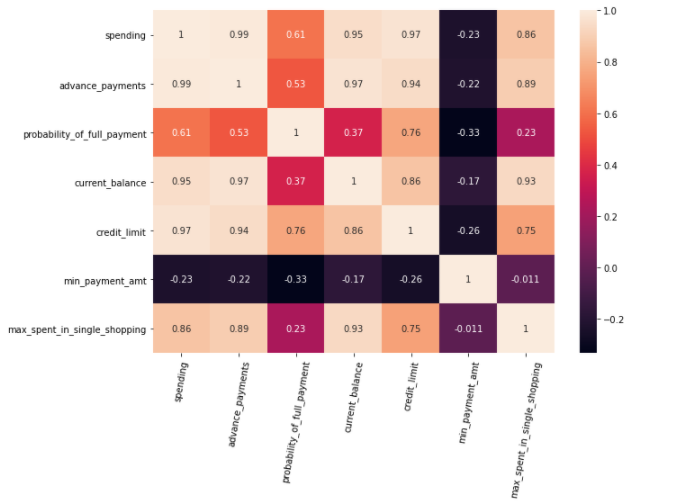
**Multivariate analysis of the columns:** Using pairplot



**Inference:** We could infer from above plot that there exists linear relationship between some of the columns like,

* Spending with advance payments, current balance and credit limit.
* Advance payments with current balance and credit limit.
* Current balance also follows some kind of linear relationship with maximum amount spent in single shopping.

**Correlation matrix:**

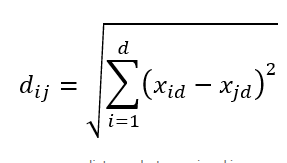
****

There exists a correlation between almost all variables either positively or negatively.

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

Standardisation/Scaling is an important step of Data pre-processing. It controls the variability in the dataset. In the given bank dataset, we have columns which are different in magnitude like 100s, 1000s, 10000s (even though they are all measured in same units except probability of full payment).

Distance-based algorithms like clustering and K-means are very likely to be affected by scaling. The algorithm needs to calculate distance between the rows. The most common distance measure used is Euclidean distance, given by the formula:



If the magnitude of variables in the dataset differ by large amounts, then the result of the distance measure is likely to be biased and move towards the variable which represents greater magnitude which in-turn suppresses other variables that might be of importance.

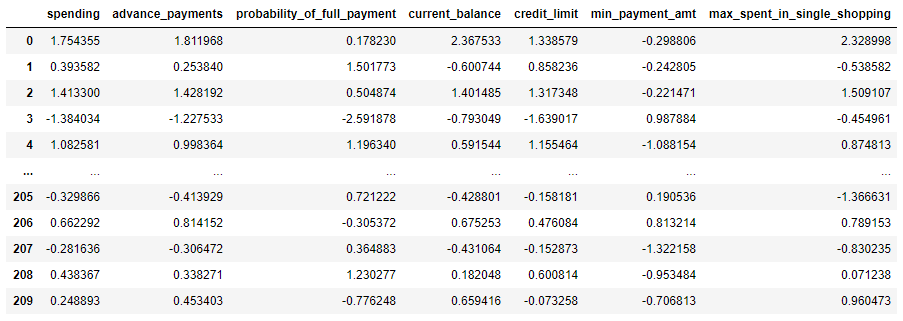
In order to offset this effect, Scaling is done to give equal considerations for each variable/feature. It converts data into specific range using a linear transformation which generates good quality clusters and improves the accuracy of clustering algorithms.

For this case study, standard z-score scaling is used which converts the group of data in our distribution such that the mean is 0 and standard deviation is 1. Z-score is expressed in terms of standard deviations from their means. It converts the dataset within a range of (-3,3) standard deviation, provided the dataset is free from outliers or skewness.

Table

Description automatically generated

**Scaled Data frame:**



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

**Hierarchical clustering** is a type of unsupervised learning technique which does not deal with target variable. Clustering groups data points such that we have high intra-cluster similarity and high inter-cluster dissimilarity. It is widely used for customer segmentation.

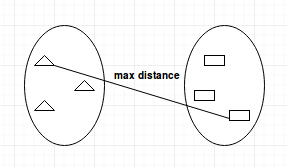
**Agglomerative hierarchical clustering** is a sequential step wise method in which initially all data points are considered as an individual cluster. At each iteration similar clusters merge with other clusters until one cluster or k clusters are formed. It can help us to identify even small size clusters present in the data.

This hierarchical clustering technique can be visualized using a **dendrogram,** a tree-like diagram that summarises the process of clustering (merges and splits).

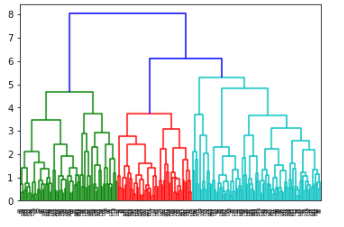
The similarity/distance between two clusters are calculated using linkage methods.

For this case study, I have tried out three linkage methods (wards, complete and average) for finding the clusters. Comparing the resulting clusters of these three, it was observed that more or less we are obtaining same groups/clusters. Hence proceeding with complete linkage method since, the similarity of two clusters is the similarity of their most dissimilar data points. This results in a compact cluster with small diameters.

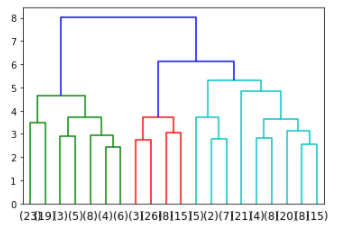
**Complete method** calculates maximum distance between two clusters before merging.



**Dendrogram for clusters formed using Complete linkage method**

****

**Truncated version:**

****

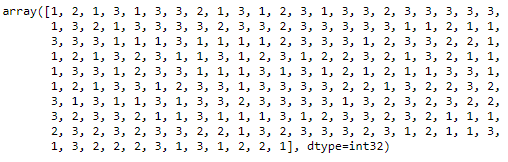
**X-axis** denotes the number of records (rows in the dataset).

**Y-axis** denotes the height/distance between the two records.

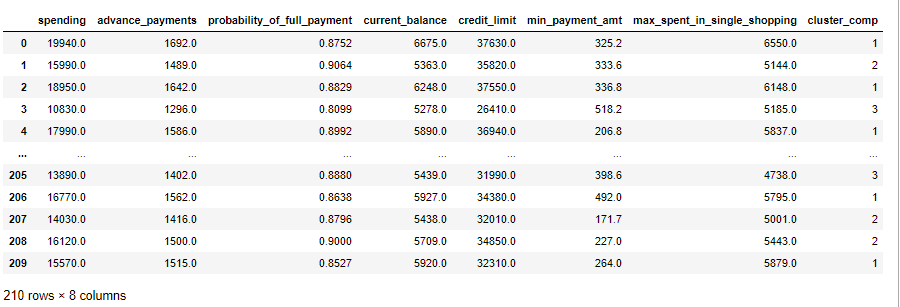
**Horizontal line** represents the distance at which two clusters merge together and **Vertical line** represents the height at which the merge has happened.

**Optimum number of clusters:** Clusters finally merge together because of agglomerative property. The cut-off point for identifying the number of clusters can be given as the highest length of the vertical line. In this case, we can go with three clusters.

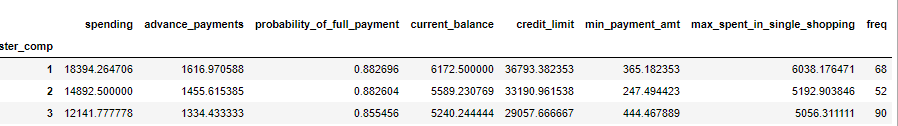
**Clusters formed:**

****

**Add this as a column to the original dataset before scaling:**

****

**Forming the cluster profiles:**

****

**Inference:**

Comparing the mean values of each clusters and categorising the values as high, low medium based on the individual column values.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Clusters** | **Spending** | **Advance payments** | **Prob of full pymt** | **Current Balance** | **Credit limit** | **Min pymt amt** | **Max spent in single shopping** | **Freq** |
| **1** | HIGH | HIGH | HIGH | HIGH | HIGH | MED. | HIGH | MED. |
| **2** | LOW | MED. | HIGH | MED. | MED | LOW | MED | LOW |
| **3** | MED. | LOW | LOW | LOW | LOW | HIGH | LOW | HIGH |

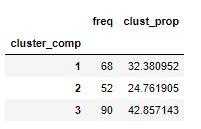
Customer behaviour in each cluster based on the above values:

**Cluster 1 – High level customers with moderate minimum payment amount:** Customers with moderate minimum payment amount and high credit limit, spending, advance payments, current balance and maximum spent in single shopping.

**Cluster 2 – Middle level customers with low minimum payment amount and spending:** Customers with low minimum payment amount and spending and moderate credit limit, advance payments, current balance and moderate maximum spent in single shopping.

**Cluster 3 – Low level customers with high minimum payment amount:** Customers with low advance payments, current balance, credit limit and maximum spent in single shopping and moderate spending.

**Cluster proportion:** Based on the number of customers in given cluster to total number.

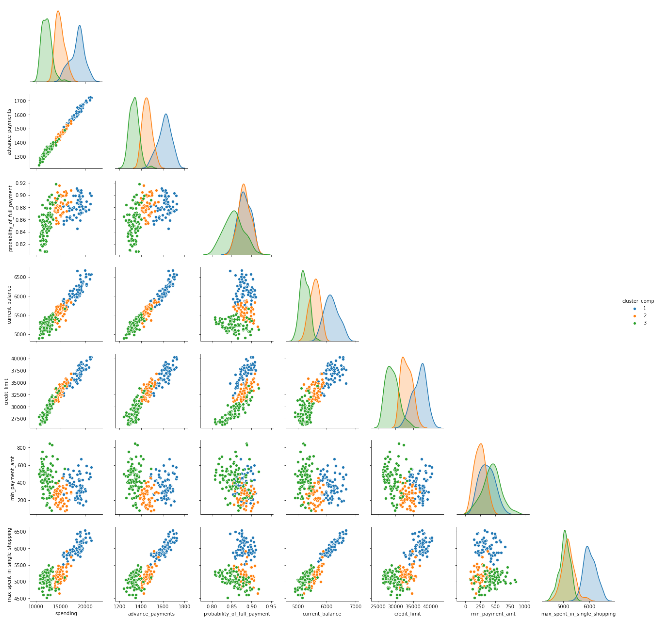


Probability of full payment has a mean of 0.87 and standard deviation of 0.02. Hence, we can say that customers in all the clusters have similar probability of full payment and it does not affect the clustering much.

**Excel file after adding the clusters:**

****

**Now, we will take a look at the pair-plot with hue as clusters so that we can identify if there is any overlap between the clusters:**

****

Advance payments, current balance and credit limit is distinctly seperated for three clusters.

Cluster 2 and 3 overlaps for column maximum spent in single shopping. Cluster 1 and 3 ovelaps for the column minimum payment amount. Cluster 1 and 2 overlaps for the column probability of full payment.

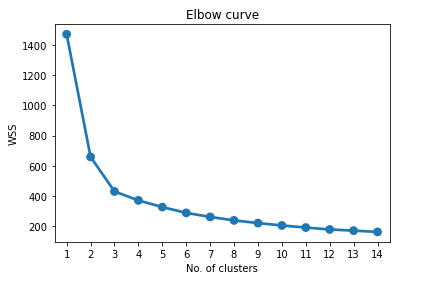
**1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Interpret the inferences from the model.**

**K-means clustering** is a non-hierarchical partitioning approach to forming good clusters by pre-specifying a desired number of clusters, k. The algorithm assigns each record to one of the k clusters, according to their distance from each cluster so as to minimize the within cluster sum of squares. It starts with k centroids and reposition the centroids until convergence is achieved, provided the clusters are stable.

**WCSS value for different number of clusters:**

****

**Elbow plot:**

****

**To determine the optimal number of clusters,** we have to select the value of k at the ‘elbow’ point after which the inertia starts decreasing in a linear fashion.

In this case, we go with 3 clusters as the optimum number.

**Silhouette score and silhouette sample –** These are indirect model evaluation technique which are used to analyse whether each and every observation mapped to clusters are actually correct.

Silhouette width is given by the formula, (b – a)/max (a, b)

a = distance of observation to its original cluster centroid.

b = distance of observation to its nearest cluster centroid.

If the minimum sil width of a particular cluster is positive, it means good mapping of data points into cluster.

Average of sil-width of all observations is called silhouette score.

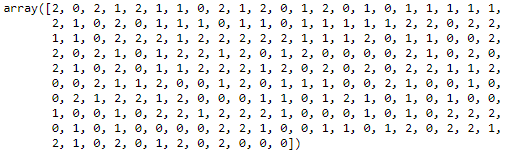
* Positive value and close to +1, clusters are well separated on an average.
* Value close to 0, clusters are not separated well enough.
* Negative value and close to -1, clustering not efficiently done.

Comparing the value of sil-width and sil-score for k = 2,3,4,5 in order to confirm on the optimum number of clusters chosen.

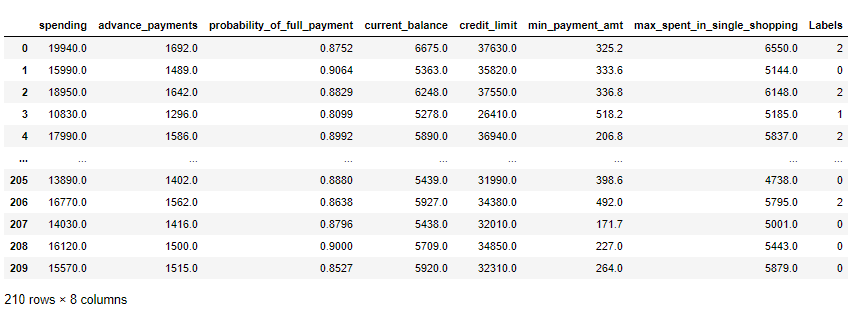
|  |  |  |
| --- | --- | --- |
| **No. of clusters** | **Min of SIL-Width** | **SIL-score** |
| 2 | -0.006 | 0.47 |
| **3** | **0.003** | **0.40** |
| 4 | -0.054 | 0.33 |
| 5 | -0.048 | 0.29 |

Minimum of sil-width of all observations for 3 clusters is positive, which means three clusters are well separated on an average compared to other clusters. Sil-score is positive and better than other clusters. Hence, for this case study we will be going with three clusters overall.

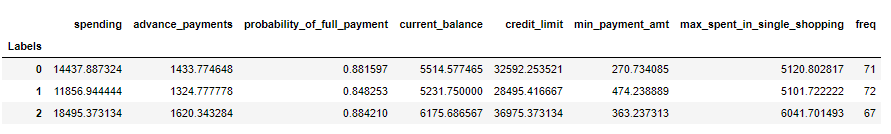
**Extracting the labels for three clusters:**

****

**Adding this label as a column to the original dataset before scaling:**

****

**Cluster profiling:**

****

**Inference:**

Comparing the mean values of each clusters and categorising the values as high, low medium based on the individual column values.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Clusters** | **Spending** | **Advance payments** | **Prob of full pymt** | **Current Balance** | **Credit limit** | **Min pymt amt** | **Max spent in single shopping** |
| **0** | MED. | MED. | HIGH | MED. | MED. | LOW | LOW |
| **1** | LOW | LOW | LOW | LOW | LOW | HIGH | LOW |
| **2** | HIGH | HIGH | HIGH | HIGH | HIGH | MED. | HIGH |

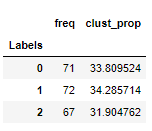
Customer behaviour in each cluster based on the above values:

**Cluster 0 – Mid level customers with low minimum payment amount:** Customers with moderate spending, advance payments, current balance, credit limit and low minimum payment amount and max spent in single shopping.

**Cluster 1 – Low level customers with high minimum payment amount:** Customers with low spending, advance payments, current balance, credit limit, max spent in single shopping and high minimum payment amount

**Cluster 2 – High level customers with moderate minimum payment amount:** Customers with high spending, advance payments, current balance, credit limit and maximum spent in single shopping and moderate minimum payment amount.

**Cluster proportions:**

****

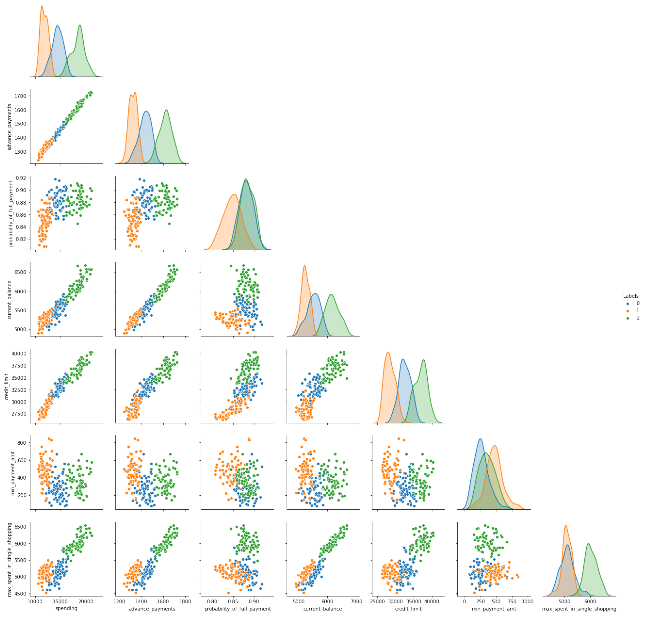
Clusters have similar proportions. There are equal number of customers in each cluster.

**Exporting the above data frame as excel**

****

From looking at the pair-plot, we can observe that clusters are distinguished. Only very few data points overlap within each cluster for different column pairs.

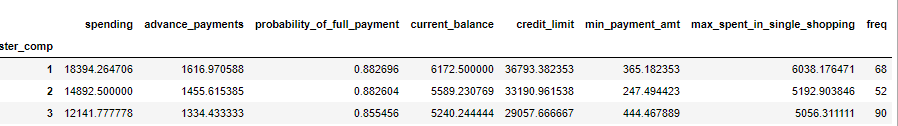
**Pair plot for different clusters**

****

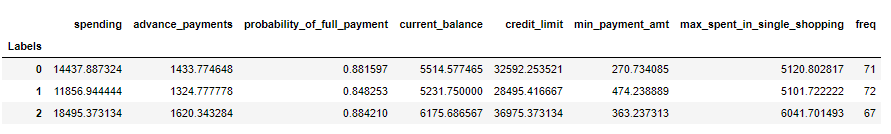
**1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters in context to the business problem in-hand.**

Credit card usage has increased rapidly in recent years owing to the online discounts, cashbacks and reward. For the given business problem, we have come up with clusters using two techniques – hierarchical and K-means clustering for segmenting the users into different categories.

Clusters by Hierarchical:

****

Clusters by K-means:

****

Both the models provide more or less the same output based on the clusters for the given dataset. Hence, we have come up with three categories of users.

**Cluster 1:** Customers with high spending, advance payments, current balance, credit limit, max spent in single shopping and medium min payment amount. These customers are typically of high value based on their spending pattern and credit limit. Probability of full payment is also 0.88 for this segment of customers, which means there is more than likely percentage of chance that these customers will pay the entire due amount. This is the wealthiest segment that has approximately 32% of customers.

**Cluster 2:** Customers with moderate spending, advance payments, current balance, credit limit, max spent in single shopping and low minimum payment amount. Customers in this cluster keep their spending in check with respect to the credit limit. They tend to not spend enough in spite of their credit limit and also, they make 10% advance payments with respect to their spending. Further analysis could be required in order in order to determine the low min payment amount. In hierarchical clustering, this group contains 24% of customers and through k-means clustering we could see that 33% of the customers belong to this cluster.

**Cluster 3:** Customers with low spending, advance payments, current balance, credit limit, max spent in single shopping and high minimum payment amount. Customers in this cluster could belong to people who might have just started their career or people who are self-aware avoiders of credit cards because of the high minimum payment amount. Through k-means clustering, we get 34% of customers in this cluster and by hierarchical clustering, this is cluster which contains maximum number of customers, approximately 42%.

**Promotional strategies for each cluster:**

|  |  |  |
| --- | --- | --- |
| **Cluster 1 – High valued customers/Wealthiest segment** | **Cluster 2 – Moderate spending customers** | **Cluster 3 – Low spending customers with high min payment amount** |
| * Rewards and cashback points. * Add-on cards at zero percent interest for family members. * Easy account management facilities. * Swipe to instalment loan since they have high spending. * Balance transfer options. * Contactless payment offers. * Zero liability on fraudulent transactions made on the card after the loss of card is reported. | * Increase in credit limit at no cost. * Simple and transparent fees, rates and terms. * Additional offers and special discounts when buying through certain stores/online. * Offering cards with waived annual fees if a specified amount is spent in a year. * Buyers protection offer and Travel amenities. * E-vouchers and gift card as reward. | * Low fees and interest. * Flat rate or percentage discount for first store purchase. * Easy EMI conversion. * Increased reward points as a joining benefit. * Complimentary benefit/reward program if a specified amount is spent in a year. * Auto assist facility 24\*7. |

**These are some of the promotional offers targeted for each cluster of customers.**

**Problem 2:**

**CART/RF/ANN**

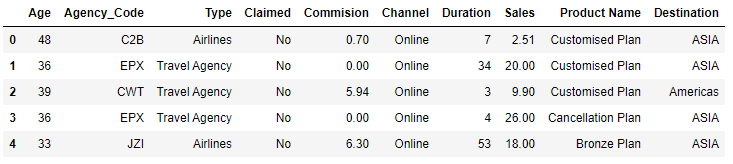
**Problem statement:**

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.

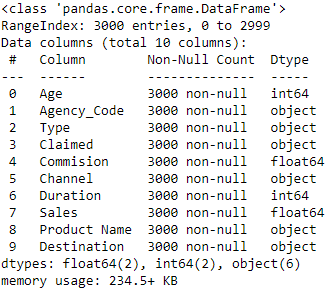
**2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, Interpret the inferences from the descriptive statistics in a detailed manner.**

**Exploratory data analysis:**

**Head of the dataset:** Verify that the dataset is loaded correctly.

****

**Information of the dataset:** There are a total of 10 variables out of which Agency code, Type, claimed, channel, product name and destination are object type and others are either float or int. We can also see that there are no missing values in the dataset.

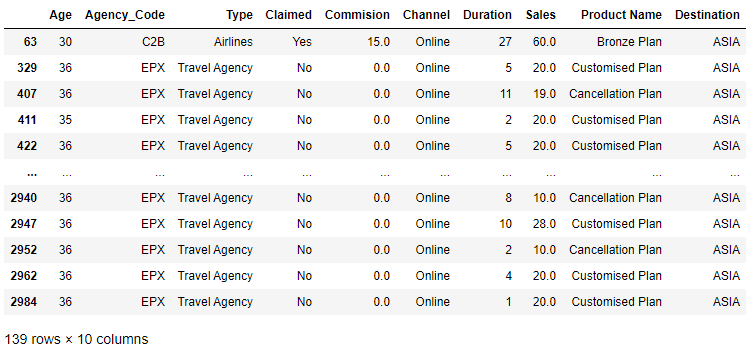
****

**Shape of the dataset:**

There are 3000 rows and 10 columns in the dataset.

**Duplicate value check**

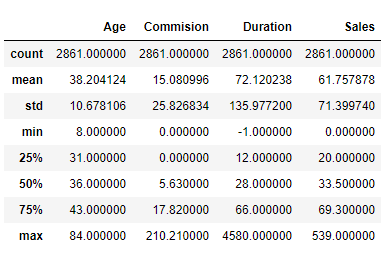
There are duplicate values in the dataset.

****

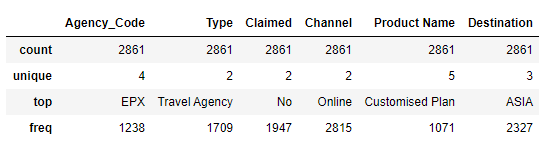
**Removing these records from the dataset since these records are of repeated values.**

Shape of the dataset after removing duplicates: (2861, 10)

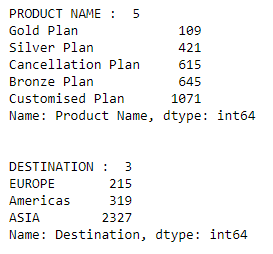
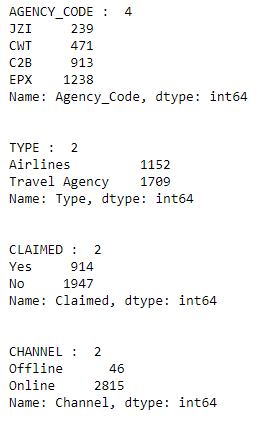
**Summary statistics of numerical columns in dataset:**

****

**Summary statistics of categorical columns:**

****

**Unique values and count of unique values in each category:**

****

**Target variable proportion:**

Proportion of Yes in claimed category: 31.95

Proportion of No in claimed category: 68.05

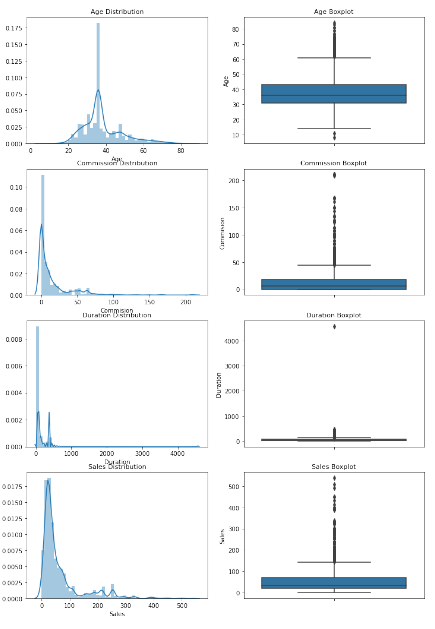
**Major proportion of customers in the given dataset, approximately 68% have not claimed for insurance.**

**Univariate analysis:** Taking into consideration single variable for analysis.

**Boxplot and Distplot of all variables:**

As observed from the below plots, out of 4 continuous variables, all variables have outliers.

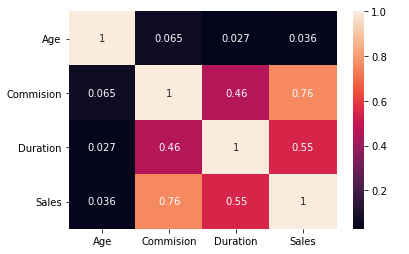
Skewness of the variables are visible in the dataset and as inferred before, all the variables are right skewed, except Age which looks like normal distribution.



**Bivariate and Multivariate Analysis of the dataset:** Taking two or more variables into consideration for analysis.

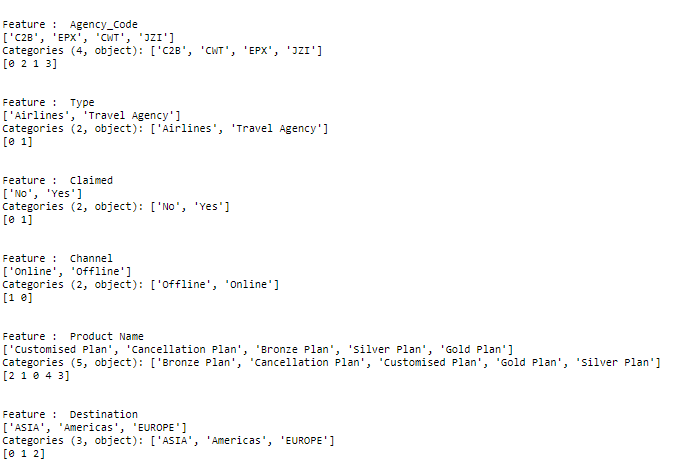
Since, all the variables are continuous, we will be going with heatmap.

**Heat map for the correlation:**

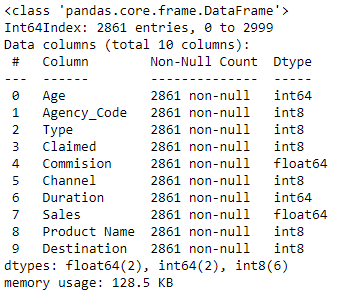


From the above heatmap, we can see that the positive linear relationship is exhibited by the variable pairs Commission & Sales (0.76), Duration & Sales (0.55). Other than that, all other variables correlation is not conclusive.

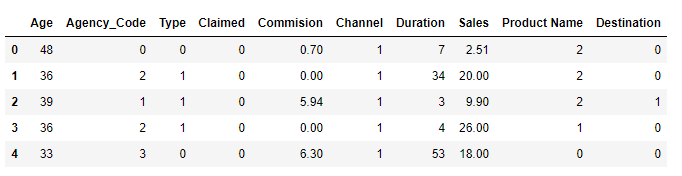
**Converting all objects to categorical column for the dataset (which is devoid of duplicates):**

****

**Information of the dataset after converting:**

****

**Head of the dataset:**

****

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

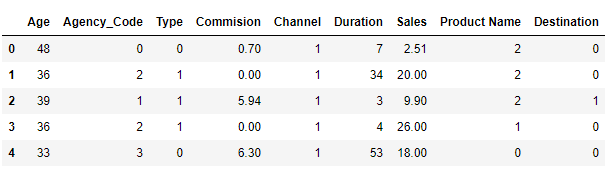
* **First step** of building a model is to **Separate the dataset into X and y variable.**

For the given business problem of tour insurance firm, ‘Claimed’ is the target variable since the problem is to come up with a model to predict the claim status

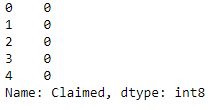
**X – Independent variable** (Removing ‘Claimed’ variable)

**Y – Dependent/ Target variable** (Having only ‘Claimed’ variable)

Head of X:



Head of y:

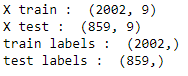


* **Second step** is to **Split the data into training and testing test.**

Splitting the data as 70% training and 30% testing.

Output of this step will be: Training independent variable (X\_train), Testing independent variable (X-test), Training dependent variable (train\_labels) and testing dependent variable (test\_labels).

Dimension of the above:



* **Third step is to build model for each CART/RF/ANN and fourth step is to predict on training and testing set**

1. **CLASSIFICATION AND REGRESSION TREE MODEL:**

Objective of CART is to split the data in such a manner that 1’s fall into particular node (pure node) and 0’s fall into another node (impure node).

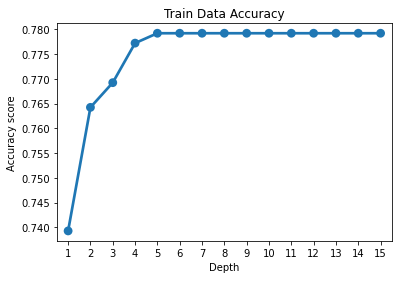
Building model using grid search cross validation method and tuning the hyper parameters in order to obtain a stable tree.

**Regularised model with pruning parameters:**

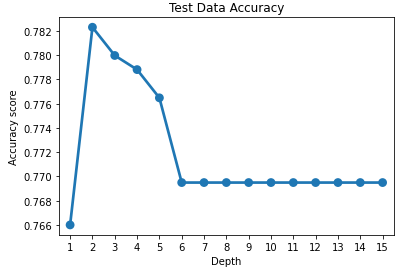


**Finalising max depth parameter by looking at the accuracy score for both training and testing set:**

**Train Data Accuracy:**

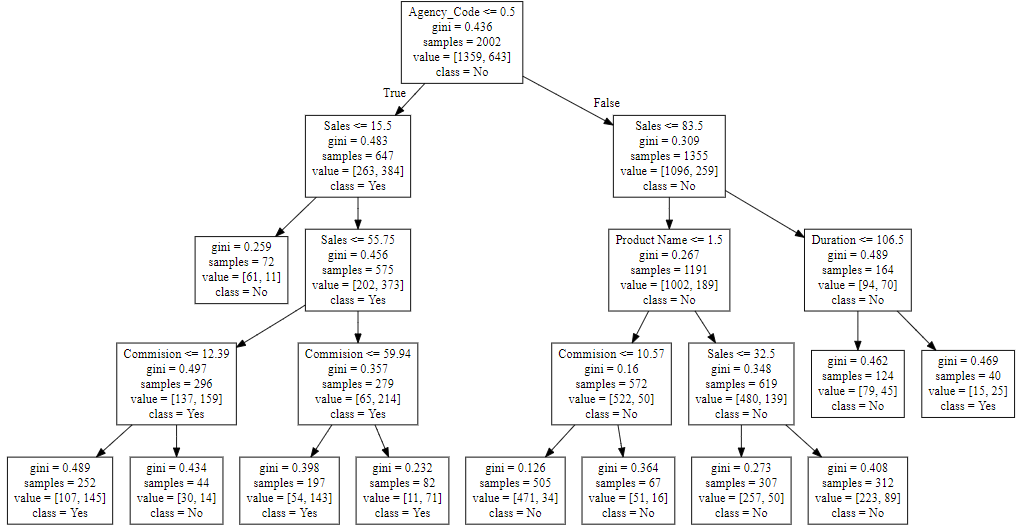
****

**Test Data Accuracy:**

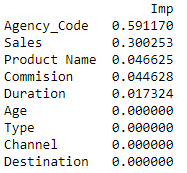
****

From the above plots we can see that for depth 4 both train and test data accuracy is stable and does not indicate over fitting or under fitting.

**Generating tree for the above model:**

****

**Variable importance:** One of the advantages of CART is feature importance.

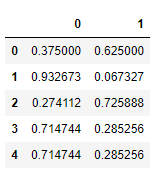


We can infer that Agency code, Sales, Product name, commission and Duration has more importance. Age, Type, Channel and Destination does not have much importance in splitting the data.

**Predicting on training and testing data:**

Using predict function on training and testing data, output will be 0’s and 1’s.

Probability of Y being 0 and Y being 1 is given by:



1. **Building Random forest classifier:**

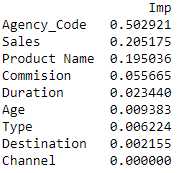
Random forest is an ensemble technique that combines several base models (decision trees) in order to produce one optimal predictive model.

Building model using grid search cross validation method and tuning the hyper parameters in order to obtain a stable random forest.

**Regularised model with pruning parameters:**



**Variable importance:**

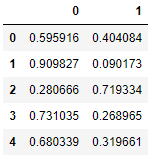
****

Random forest classifier gives high importance to Agency code comparatively. Moreover, in this model we could see that Age, Type and Destination also has some importance but minute.

**Predicting on training and testing data:**

Using predict function on training and testing data, output will be 0’s and 1’s.

Probability of Y being 0 and Y being 1 is given by:



1. **Artificial Neural network model:**

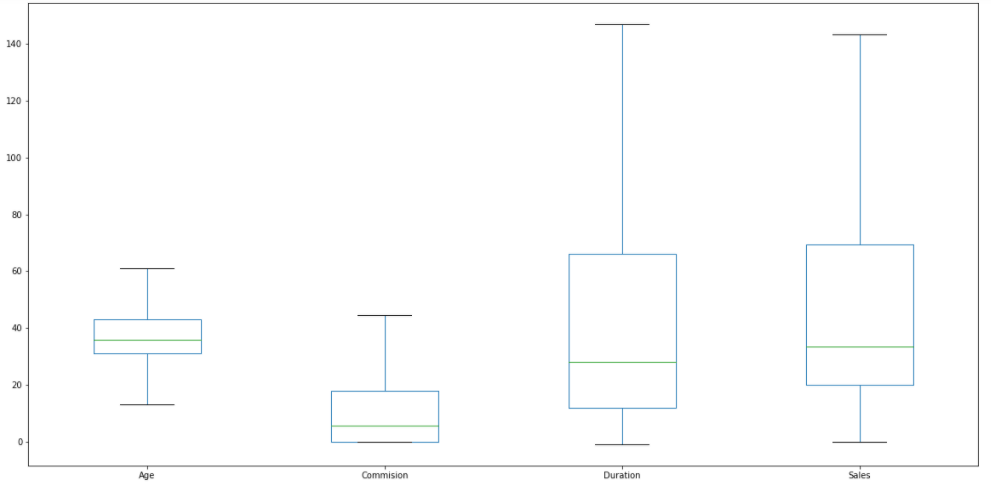
ANN is a machine learning algorithm uses a network of interconnected nodes which are connected by synaptic weights.

Dataset without outliers is being used for splitting the train and test dataset.

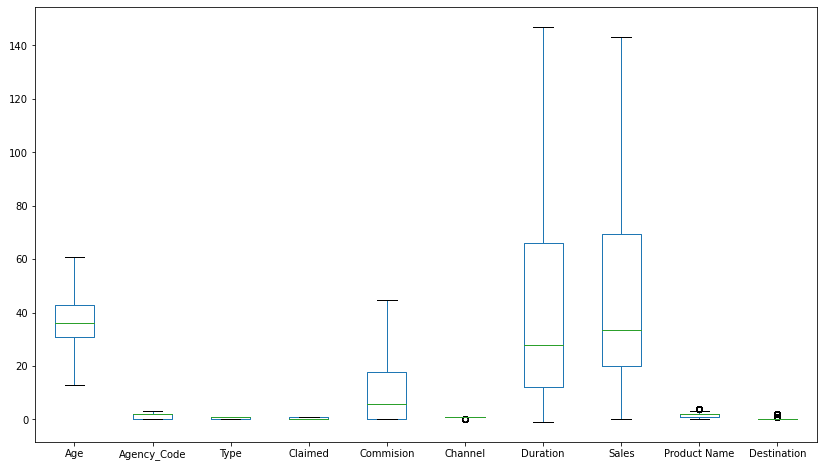
**Outlier treatment:** Before scaling the variables, since the dataset contains outliers, we have to treat the outliers in order for the output to be valid because if the scaling is done on the dataset with outliers then it would result in meaningless mean and standard deviation.

**IQR treatment for outliers:** Custom function is defined which takes column as input and returns two output for a particular column if the value is greater than maximum limit or less than minimum limit. Loop the function for all the variables such that it replaces the values greater than maximum limit by that limit and vice versa.

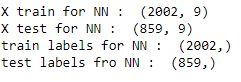
**Boxplot of all variables after the outlier treatment:**

****

Boxplot after converting the variables from object to int:



**Splitting the dataset into train and test:**

****

**Scaling:**

CART and Random forest classifier are tree-based models and does not require the variables to be scaled since it uses Gini co-efficient/ information gain which will not be affected by scaling.

However Artificial Neural Networks works different. When we observe the features, we have age, commission duration and sales, even though all are continuous variables they have different scales and units. Since neural networks uses gradient descent as an optimization algorithm to determine weights for each feature. Scaling is done after train test split for this algorithm.

For this case study (ANN model), standard z-score scaling will be used which converts the group of data in our distribution such that the mean is 0 and standard deviation is 1. Z-score is expressed in terms of standard deviations from their means. It converts the dataset within a range of (-3,3) provided the dataset is free from outliers or skewness.

Table

Description automatically generated

We will perform fit and transform for training data and transform test data with respect to the training data mean and standard deviation.

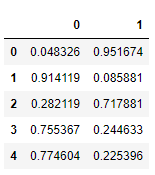
**Regularised model with hyper parameters:**



**Predicting on training and testing data:**

Using predict function on training and testing data, output will be 0’s and 1’s.

Probability of Y being 0 and Y being 1 is given by:

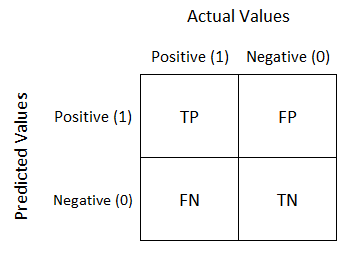


**2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model, Write inferences on each model.**

**Fifth step** of the model is to evaluate it and see how good it will perform for future records.

Some of the model evaluation techniques are:

* Accuracy – how precisely the model classifies the data points.
* Confusion Matrix – 2 \* 2 tabular structure reflecting the model performance in four blocks



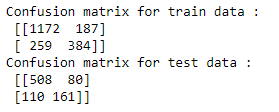
* Receiver operating characteristics (ROC) curve – A technique to visualize classifier performance
* ROC\_AUC score – Area under curve, which is by calculating the percentage area below the curve.

**CART Model evaluation:**

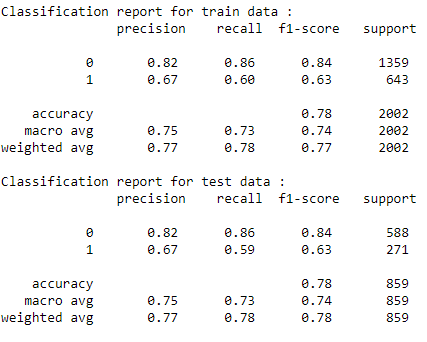
**Accuracy score:**

****

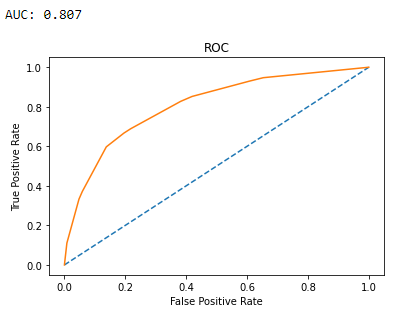
**Confusion matrix:**

****

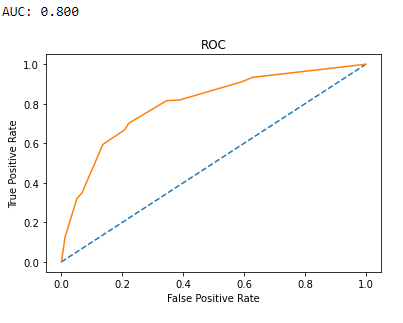
**Classification report:**

****

**AUC and ROC curve for training set:**

****

**AUC and ROC curve for testing set:**

****

**INFERENCE:**

|  |  |  |
| --- | --- | --- |
| **Metrics** | **Training set** | **Testing set** |
| **Accuracy** | 0.77 | 0.77 |
| **Precision** | 0.67 | 0.67 |
| **Recall** | 0.60 | 0.59 |
| **F1 score** | 0.63 | 0.63 |

Training and testing set results are almost similar, with overall measures on a decent side.

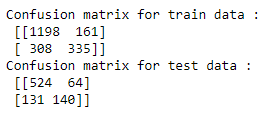
Agency is the most important variable for predicting claimed.

**Random Forest model evaluation:**

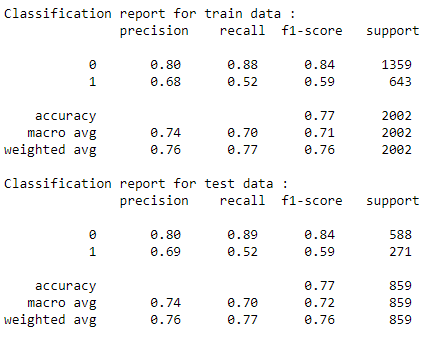
**Accuracy score:**

****

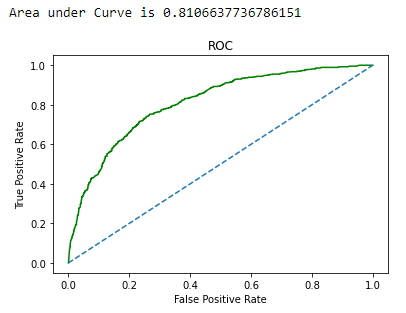
**Confusion matrix:**

****

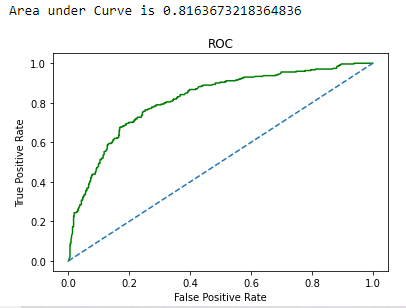
**Classification report:**

****

**AUC and ROC curve for training set:**

****

**AUC and ROC curve for testing set:**

****

**INFERENCE:**

|  |  |  |
| --- | --- | --- |
| **Metrics** | **Training set** | **Testing set** |
| **Accuracy** | 0.76 | 0.77 |
| **Precision** | 0.68 | 0.69 |
| **Recall** | 0.52 | 0.52 |
| **F1 score** | 0.59 | 0.59 |

Training and testing set results are almost similar, with overall measures on a decent side.

Agency is the most important variable for predicting claimed.

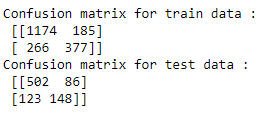
Recall has come down with respect to the CART model. Accuracy and precision have increased by 1% for training and testing set.

**Artificial neural network model evaluation:**

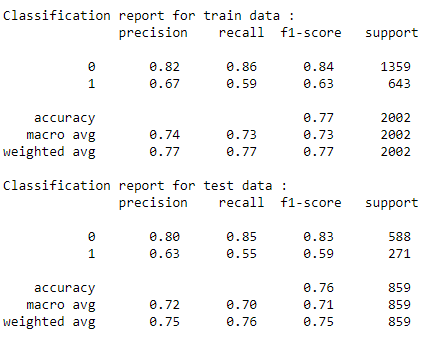
**Accuracy score:**

****

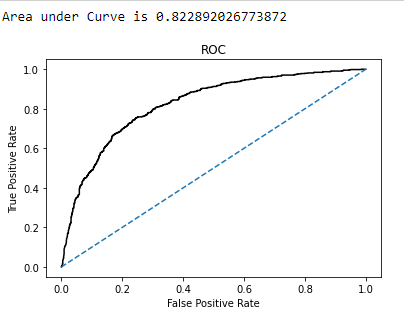
**Confusion matrix:**

****

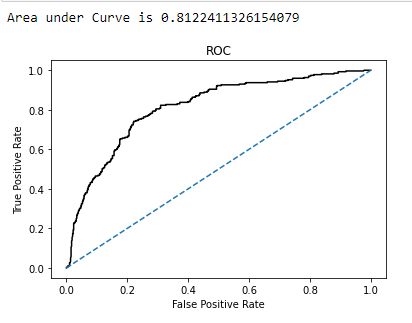
**Classification report:**

****

**AUC and ROC curve for training set:**

****

**AUC and ROC curve for testing set:**

****

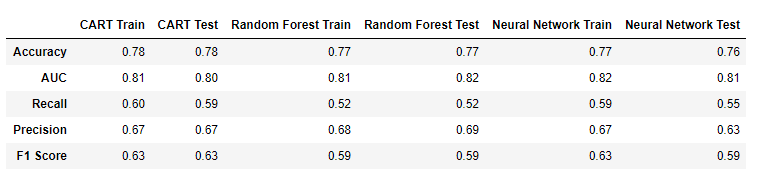
**INFERENCE:**

|  |  |  |
| --- | --- | --- |
| **Metrics** | **Training set** | **Testing set** |
| **Accuracy** | 0.77 | 0.75 |
| **Precision** | 0.67 | 0.63 |
| **Recall** | 0.59 | 0.55 |
| **F1 score** | 0.63 | 0.59 |

Training and testing set results are somewhat different, with overall measures on a decent side. Metrics have significantly reduced in testing set than training set. Recall has come down with respect to the CART model.

**2.4 Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized.**

**Performance metrics combined:**

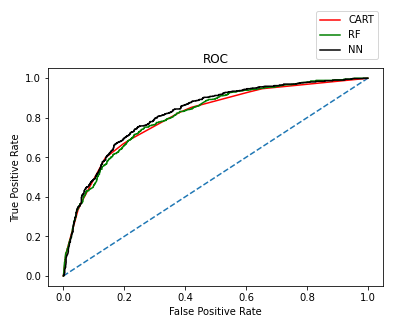


In each model, one of the metrics is better. Like in CART, recall is better than other models. AUC is more or less same for all the models. Precision for the test set is better in random forest but F1 score is very less compared to other models.

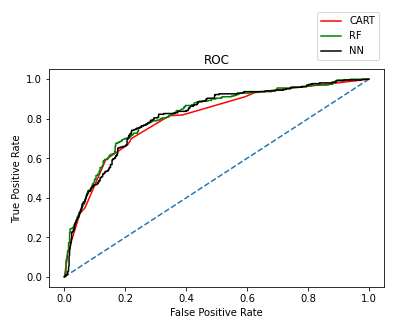
In Neural network algorithm, performance metrics have decreased significantly in testing set than training set.

No model indicates underfitting or overfitting.

**ROC AUC curve for training set**

****

**ROC AUC curve for testing set**

****

It is practically observed tree-based algorithms perform well when we have small amount of data. They prevent overfitting and can also work well with outliers and missing values. Overall, all the three models are reasonably stable enough to be used for making predictions. From CART and random forest, the variable Agency is found to be the most important feature for predicting if a person has claimed for insurance or not.

The model which is best/optimised in this particular business would be CART because of the recall which is better in CART testing set than other two models.

**2.5 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.**

The goal of the business problem is to predict the chances of customers claiming for a particular insurance. Knowing whether a particular person is going to claim for insurance or not helps the insurer to plan in advance their expenses and also to make sure that they are solvent enough to pay the claim amount. This would also help them to change their underwriting guidelines and premium amount accordingly when they are facing higher claims.

Using the dataset given by the insurance firm, models were built and evaluated how different variables like Agency code, sales, commission, age, type, channel, etc. Here the target variable to predict is ‘Claimed’.

**Insights from the models:**

* In this case study, 68% of the data points are in ‘No’ claimed category and 32% of the data points are in ‘Yes’ claimed category, the classifier might easily classify all the ‘No’ stances which could be the majority while calculating accuracy.
* CART model performed consistently with both training and testing set.

Our business problem is to identify the claim status, we are focused on determining the false positives and false negatives.

* False positive (FP) - Datapoints that are actually false but predicted as true. This is also known as type 1 error. In order to reduce the type 1 error, we have to increase the precision of the model (among the points identified as positives by the model how many are actually positive).

Type 1 error in this case study will lead to insurer charging higher premium for customers who is not going to claim. This type of error is of less priority in this business context, since tagging the actual negative as positive does not affect the insurance firm drastically.

* False negative (FN) – Datapoints that are actually true but predicted as false. This is known as type 2 error. In order to reduce to type 2 error. We have to increase recall (how many actual true data points are identified as true by the model)

Type 2 error is very important for our case study, since predicting the actual true data points are false will lead to higher insurance claims for which the firm might not be ready.

**Recommendations for the business:**

**If the claim status comes out as Yes:**

* Can charge additional premium/interest rates for the particular customer is order to compensate for the expenses incurred when they claim.
* Penalty fees can be charged if they claim within a certain period of policy issuance in order to reduce higher first year claim, if any.
* Based on the product name (Customized, cancellation, silver, gold and bronze plan) charge premium accordingly.
* Regular internal audits for all claims and this should apply to all stages in claim processing department.

**If the claim status comes out as No:**

* Underwriting guidelines to be modified in order to determine if the insurance is really necessary for the customer, if not, we can recommend some other plan which might be even more useful for the customers.

Here we have built model with 9 independent variables for predicting the ‘Claimed’ dependent variable. If we had some more factors like claim frequency, amount those of which can affect the claim status predominantly could make the model better in predicting ‘Yes’ stances.