PG-DSBA Dec_A 2020

Data Mining Project

Submitted by: F Maria Jasmine

Date of Submission: 02/05/2021

Table of Contents

II. Problem Statement:2

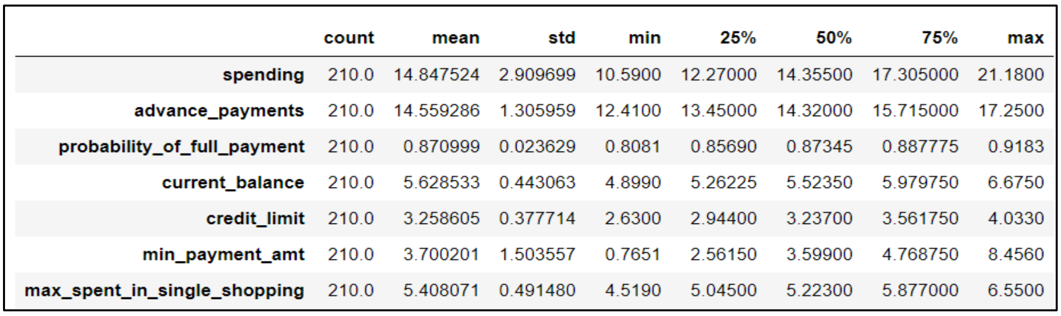
Problem 1: Clustering

Problem Objective: To get meaningful clusters from the given data and suggest appropriate promotional offers to the created clusters. Clusters are created using Hierarchical and k-means techniques.

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

Data Description:

1. The shape of the given data set is: There are 210 observations/rows and 7 columns.
2. The info function for the given data set describes:
   - There are no index changes.
   - No missing values.
   - The data types of all the columns are *float*
3. The missing values are also checked again using the 'isnull()' function. There are no null values in the data set.
4. There are no duplicates in the data set.
5. The summary statistics of the given data set:
   - The Mean and Median values for all the features are not significantly different.
   - The Minimum and Maximum spending made by the individuals are 10590.0 and 21180 respectively.
   - The probability of individuals making full payment ranges from 81% to 92% approximately.
   - The average probability of individuals making full payment is 87.1%, which is a decent number.
   - The advance payments made by the individuals ranges from 1241 and 1725. The advance payments are made not more than 11% of the total spending.
   - The credit limit given for the individuals ranges from 26300 to 40330.
   - Maximum 3% of Spending amount is paid as a Minimum Payment Amount.
   - Not more than 6550 is spent on a single shopping by any individual.

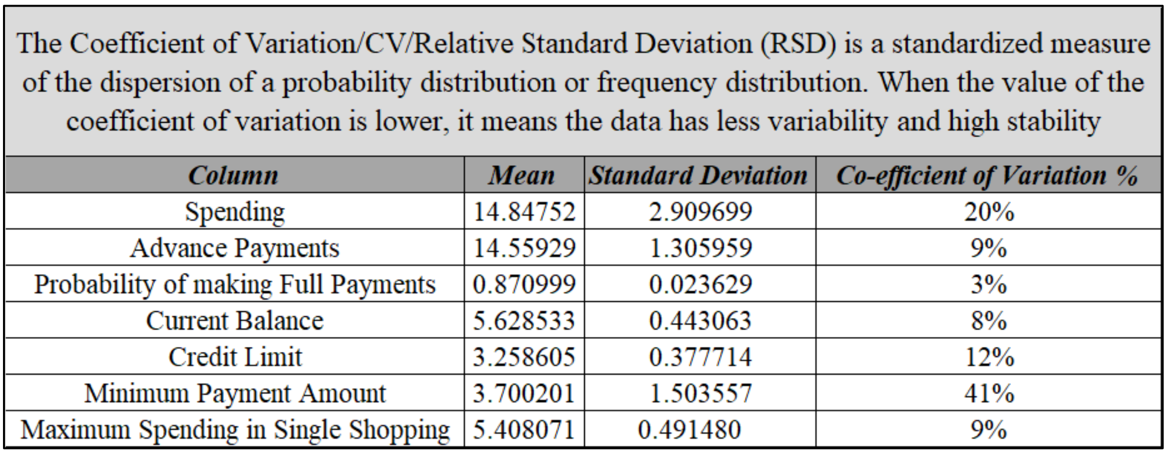


	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.5500

Coefficient of Variation:

Below is the CV% of different features. The lesser the percentage is, it means it has less variability and the data is stable. The CV% is not too high, hence the data is less variable and more stable.

Table: 1.1

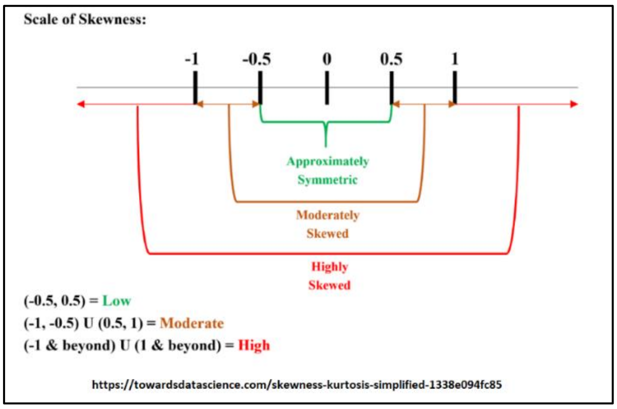


The Coefficient of Variation/CV/Relative Standard Deviation (RSD) is a standardized measure of the dispersion of a probability distribution or frequency distribution. When the value of the coefficient of variation is lower, it means the data has less variability and high stability

Column	*Mean*	*Standard Deviation*	*Co-efficient of Variation %*
Spending	14.84752	2.909699	20%
Advance Payments	14.55929	1.305959	9%
Probability of making Full Payments	0.870999	0.023629	3%
Current Balance	5.628533	0.443063	8%
Credit Limit	3.258605	0.377714	12%
Minimum Payment Amount	3.700201	1.503557	41%
Maximum Spending in Single Shopping	5.408071	0.491480	9%

Skewness and Kurtosis:

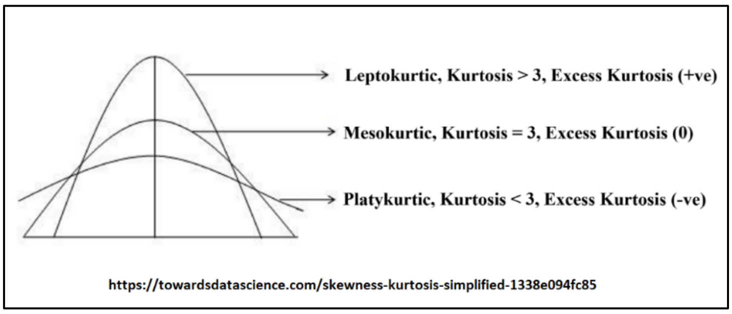
The skewness is a measure of the asymmetry of the probability distribution assuming a unimodal distribution and is given by the third standardized moment. We can say that the skewness indicates how much our underlying distribution deviates from the normal distribution since the normal distribution has skewness 0.



https://towardsdatascience.com/skewness-kurtosis-simplified-1338e094fc85

From the Table 1.1, we can see that, the mean and the median for all the columns are almost same.

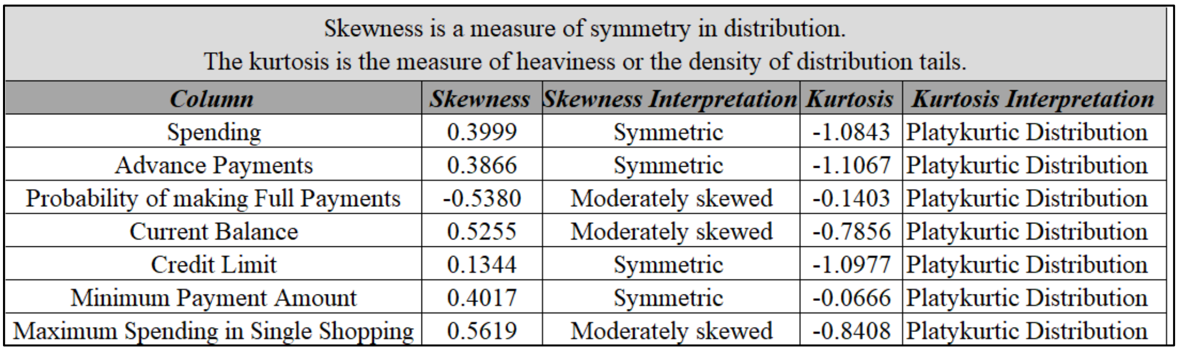
In statistics, we use the kurtosis measure to describe the “tailedness” of the distribution as it describes the shape of it. It is also a measure of the “peakedness” of the distribution. A high kurtosis distribution has a sharper peak and longer fatter tails, while a low kurtosis distribution has a more rounded pean and shorter thinner tails.


Leptokurtic, Kurtosis > 3, Excess Kurtosis (+ve)
Mesokurtic, Kurtosis = 3, Excess Kurtosis (0)
Platykurtic, Kurtosis < 3, Excess Kurtosis (-ve)

https://towardsdatascience.com/skewness-kurtosis-simplified-1338e094fc85

The Table: 1.3 shows the Skewness and Kurtosis type of all the features of the given data set.

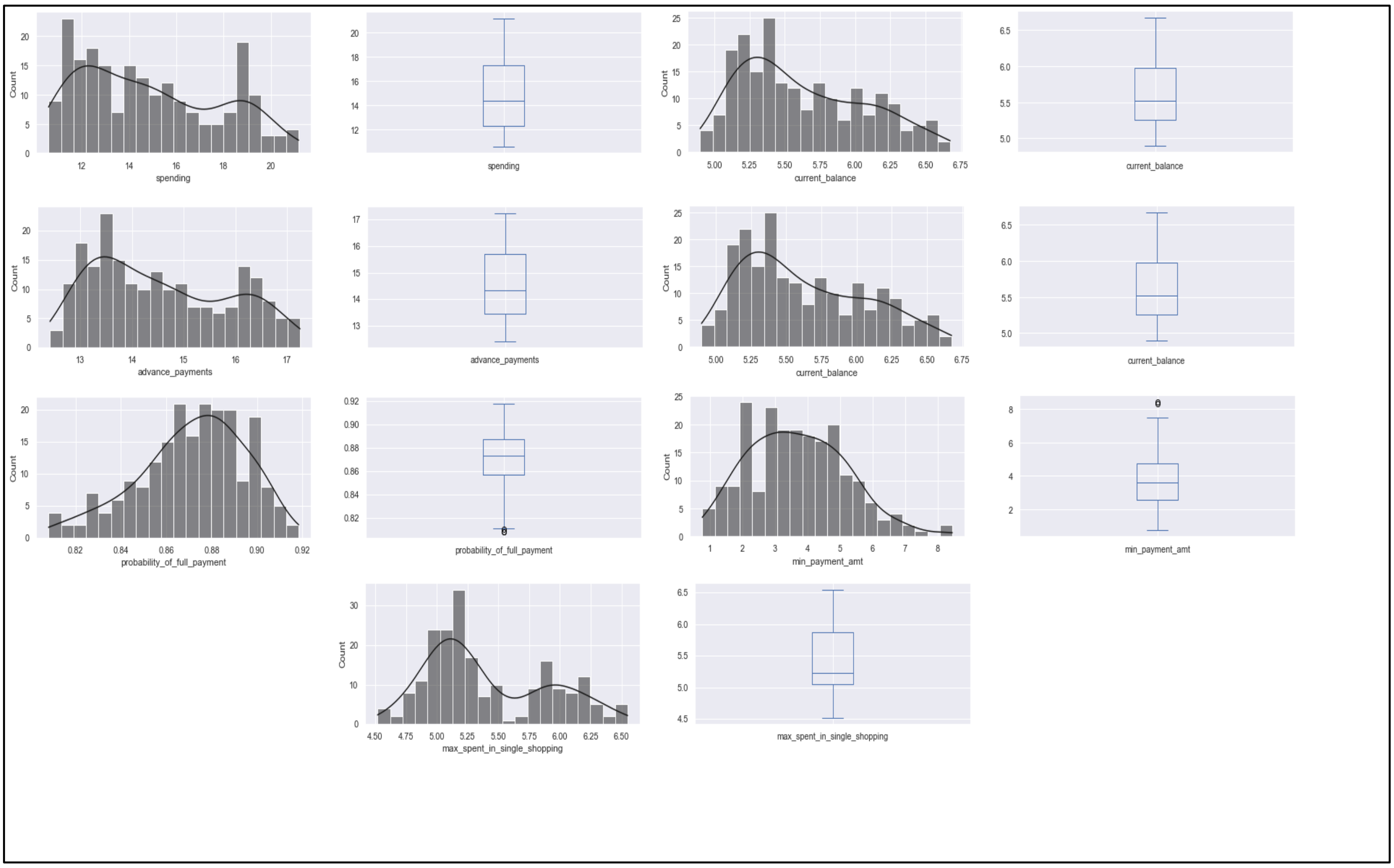
Table: 1.2



Skewness is a measure of symmetry in distribution.
The kurtosis is the measure of heaviness or the density of distribution tails.

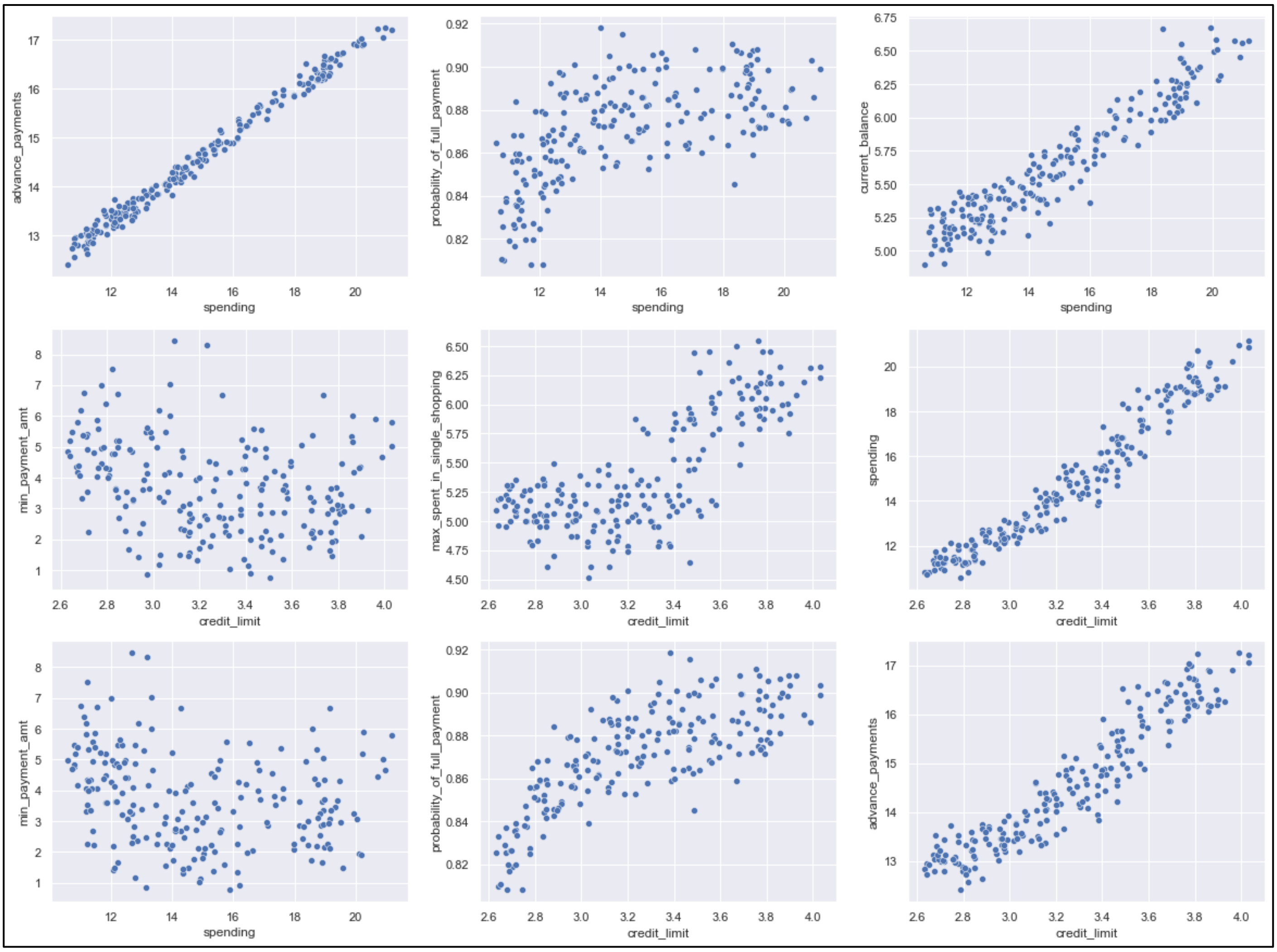
Column	***Skewness***	***Skewness Interpretation***	***Kurtosis***	***Kurtosis Interpretation***
Spending	0.3999	Symmetric	-1.0843	Platykurtic Distribution
Advance Payments	0.3866	Symmetric	-1.1067	Platykurtic Distribution
Probability of making Full Payments	-0.5380	Moderately skewed	-0.1403	Platykurtic Distribution
Current Balance	0.5255	Moderately skewed	-0.7856	Platykurtic Distribution
Credit Limit	0.1344	Symmetric	-1.0977	Platykurtic Distribution
Minimum Payment Amount	0.4017	Symmetric	-0.0666	Platykurtic Distribution
Maximum Spending in Single Shopping	0.5619	Moderately skewed	-0.8408	Platykurtic Distribution

Univariate Analysis:



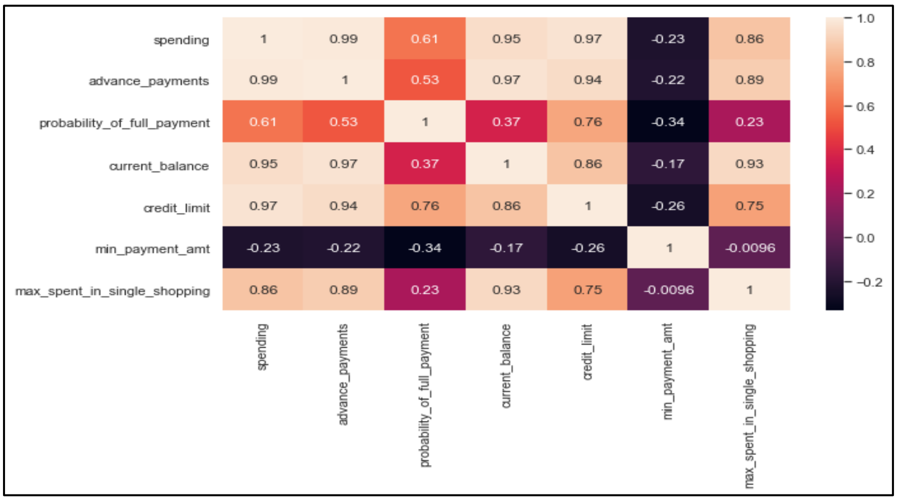
1. 6 variables in a given data set is Positively skewed and the skewness and kurtosis values and their interpretations are shown in the Table 1.3.
2. The Probability of the full payment feature alone is a negatively skewed distribution with 0.0236 as Standard deviation and 0.0006 as Variance.
3. There are outliers in Probability of the full payment and Minimum payment amount feature. Hence, needed an outlier treatment before proceeding with Clustering techniques

Bivariate Analysis:



- The Probability for full payment and Advance payments are Moderately correlated. When a customer pays his/her bill in advance, there are moderate chances of him/her in paying the full amount on time.
- Current balance and Credit limit are strongly correlated. Larger the credit limit more is the Current balance. Customers does not tend to spend more when they have large credit limit.
- When Customer has more Credit limit, they tend to spend more in a single purchase. Example: Gold purchase, Fixed Asset purchase or any Furnitures and Fitting, Electronic Appliances etc.
- The probability of the full payment and the Minimum payment amount is negatively correlated. Customers pay a Minimum sum of amount to postpone the full payment. Hence, when there is minimum payment made the Probability of making full payment is less.

Correlation Plots:



Interpretations from Correlation Plot:

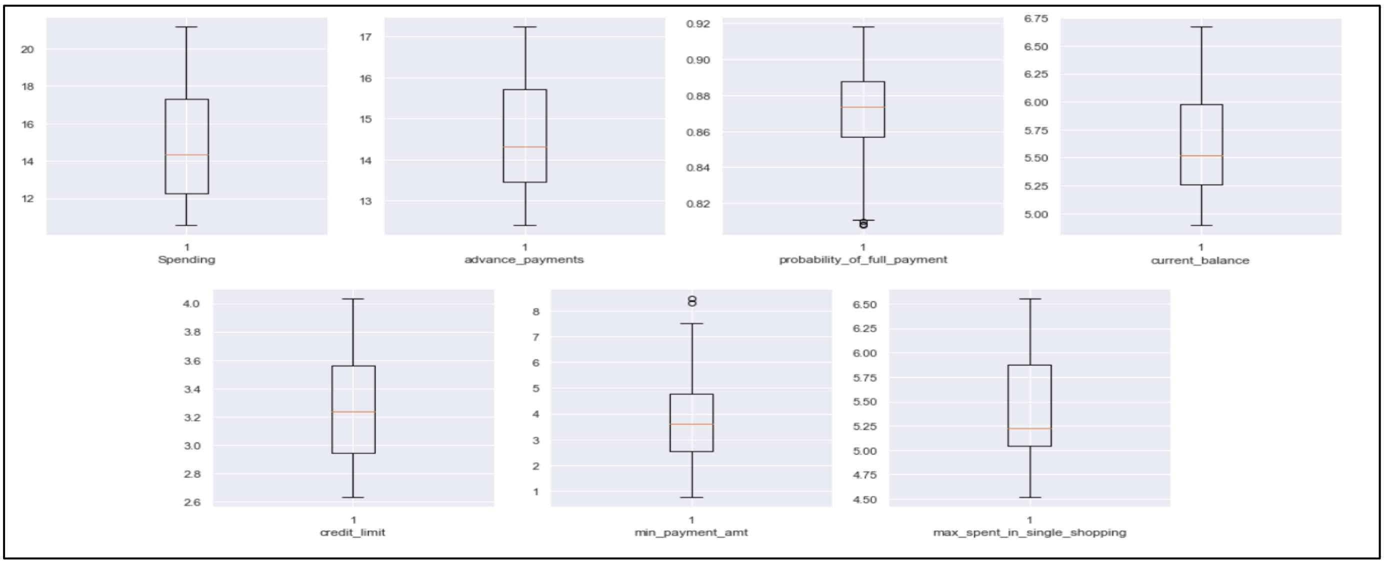
- The above correlation plot displays the correlation among the different variables in a data set
- The variable Spending is highly correlated with 4 variables, they are Advance_payments, Current balance, credit limit and Max amount spent in a single shopping.
- The Advance payments variables are also highly correlated with the above mentioned 4 variables.
- Spending is the important criteria for a Credit card mechanism.
- The most popular cards in Banking industry are the 'Spend Cards' type credit cards. The more you spend the more you save through 'Reward points.
- Hence, in our data set, the Spending and Maximum Spent in sin gle shopping is very critical for analysing the promotional offers.

Key note:

- From the heat map it is evident that there are more than 2 highly correlated variables.
- Multi collinearity does not impact the cluster output on a large scale, we can either ignore or treat the correlations.
- By performing PCA we can remove the remove the Multicollinearity among the variables as each and every Principal component are orthogonal to each other.
- To do PCA we need high dimensional data. The data available is not with high dimensional.
- For the given data set PCA is performed. Hierarchical and K-means clustering is performed on both the original data and the PCA applied data. Both the cluster outcomes are compared and selected.

Treatment for Outliers:

Graphical representation of the Boxplots before treating the Outliers:



Outliers are data points that don't fit the pattern of rest of the numbers. They can be extremely high or extremely low. Failing to treat outliers may impact the statistical procedures.

Steps followed for treating the outliers:

1. The Lower range and the Upper range of the distribution is calculated using the Inter-Quartile Range.
2. Lower Range = Q1-(1.5 * IQR)
3. Upper Range = Q3+(1.5 * IQR)

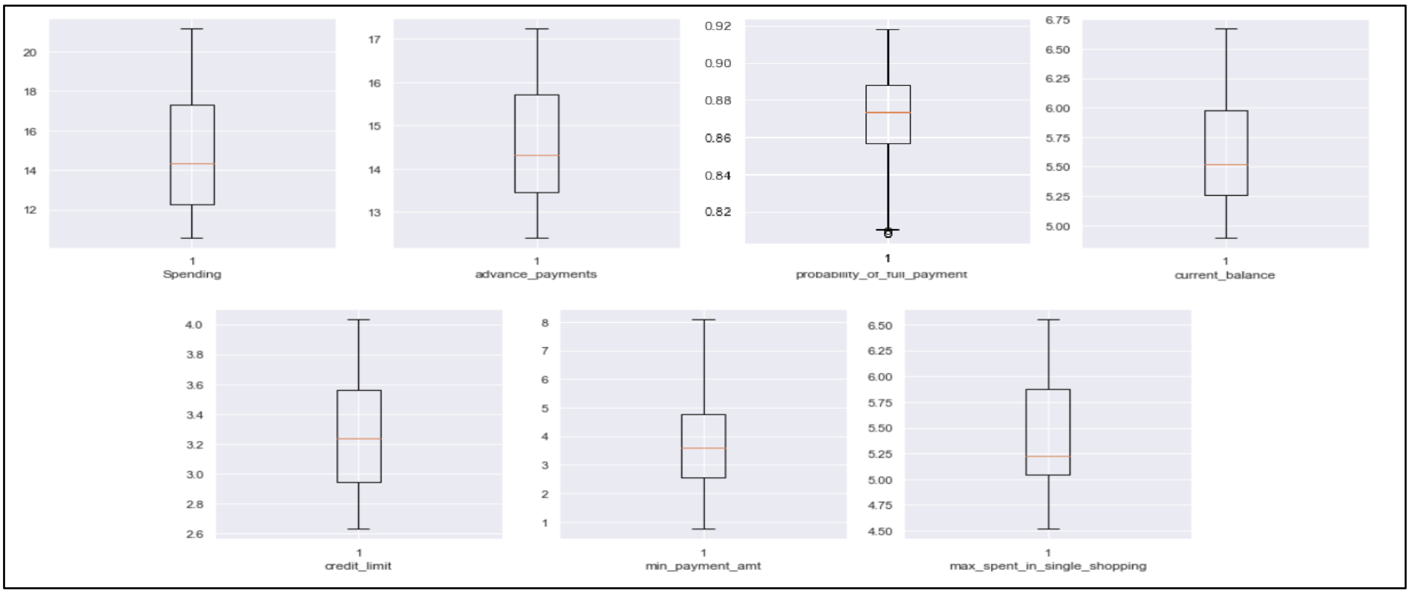
Outliers in Probability of Full payment feature:

- The Probability percentage feature has an outlier below the minimum limit.
- The least percentage is .8081 and the lower range of the quantile is 0.810.
- There is no significant difference between the values. Also, the percentages ranges between 0 to 1. As this data set is from Banking Industry-Credit Card type, payment probability is a key feature. Customers varied probability percentages should be considered before clustering.
- Hence, the outlier in the Probability percentage feature is not treated.

Outliers in Minimum Payment Amount feature:

- The Minimum payment column has outlier above the maximum limit.
- The upper range is 8.07 and the outlier amount is 8.456
- Any payment above the Upper range in a data set can be grouped, which will not have any impact.
- Hence, the outlier in this feature is imputed by the upper range value.

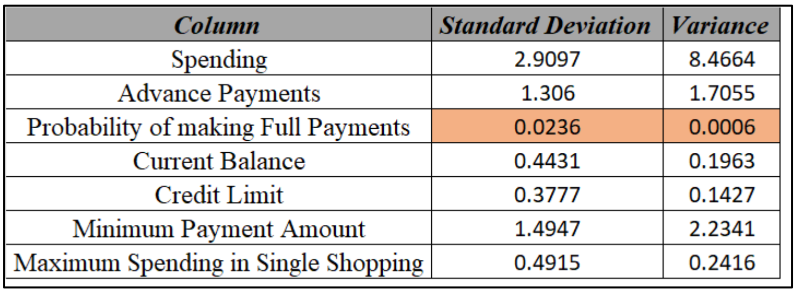
Graphical representation of the Boxplots after treating the Outliers:



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

- Machine Learning algorithm gets affected by the magnitude of the variables. The algorithm should not be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale.
- Clustering algorithms such as K-means do need feature scaling before they are fed to the algo. Since, clustering techniques use Euclidean Distance to form the cohorts, it will be wise e.g to scale the variables.
- In this data set we have 6 monetary based features with different Denomination units and 1 Percentage feature.
- **Variance** is a measure of how data points vary from the mean, whereas standard deviation is the measure of the distribution of statistical data. The basic difference between both is standard deviation is represented in the same units as the mean of data, while the variance is represented in squared units

Table: 1.3

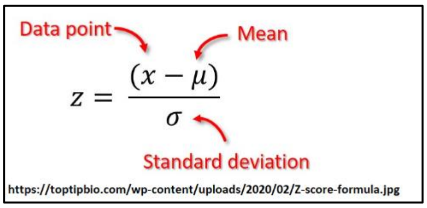


Column	*Standard Deviation*	*Variance*
Spending	2.9097	8.4664
Advance Payments	1.306	1.7055
Probability of making Full Payments	0.0236	0.0006
Current Balance	0.4431	0.1963
Credit Limit	0.3777	0.1427
Minimum Payment Amount	1.4947	2.2341
Maximum Spending in Single Shopping	0.4915	0.2416

- Standard deviation and variance for the probability_of_full_payment feature is very low. Hence, there are chances that the features with higher magnitude impacting the output.
- Therefore, the data should be scaled before performing the clustering.

Standardizing the data set using the Standard Scalar from sklearn.preprocessing:

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.



https://toptipbio.com/wp-content/uploads/2020/02/Z-score-formula.jpg

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

“Unsupervised Learning Algorithm is a machine learning technique, where you don’t have to supervise the model. Rather, you need to allow the model to work on its own to discover information, and It mainly deals with unlabelled data.”

Hierarchical clustering:

Hierarchical clustering is one of the popular clustering techniques after **K-means Clustering**. It is also known as Hierarchical Clustering Analysis (HCA). Which is used to group unlabelled datasets into a Cluster. This Hierarchical Clustering technique builds clusters based on the **similarity** between different objects in the set. It goes through the various features of the data points and looks for the similarity between them. This process will continue until the dataset has been grouped. Hierarchical Clustering deals with the data in the form of a tree or a **well-defined hierarchy**

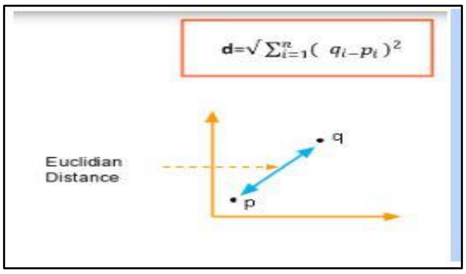
Why Hierarchical clustering:

- No need to pre-specify the number of clusters.
- An attractive tree-based representation of the observations, called a Dendrogram can be created.

Distance Metric:

Distance measure determines the similarity between two elements and it influences the shape of the clusters. Some of the types are: Euclidean, Squared Euclidean, Manhattan and Cosine.

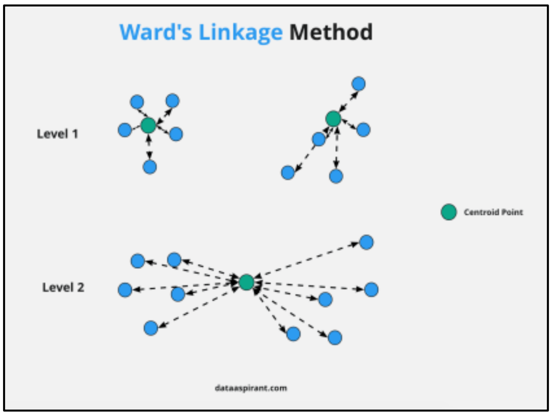
The most common method to calculate distance measures is to determine the distance between the two points. Let's say we have a point P and point Q: the Euclidean distance is the direct straight-line distance between the two points. The formula for distance between two points is shown below:



[What is Hierarchical Clustering and How Does It Work? (simplilearn.com)](#)

Linkage Method:

Ward's Linkage: This method is the similarity of two clusters. Which is based on the increase in squared error when two clusters are merged, and it is similar to the group average if the distance between points is distance squared. The objective is to *minimize the within cluster variance.*



[How the Hierarchical Clustering Algorithm Works (dataaspirant.com)](#)

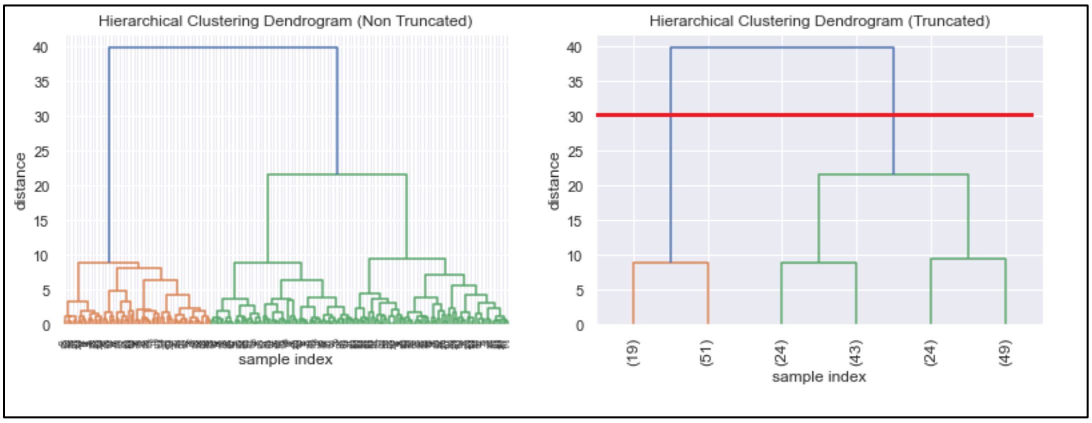
Why Ward Linkage:

1. It usually produces better cluster hierarchies.
2. This method is less susceptible to noise and outliers.

Hierarchical clustering on Original Scaled Data (Correlations not treated):

Dendrogram:

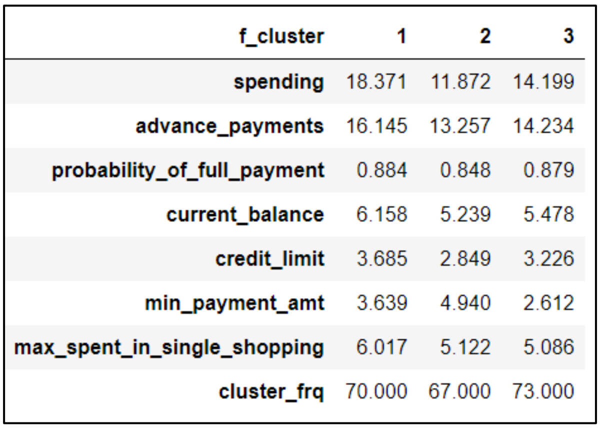
A dendrogram is a visualization in form of a tree showing the order and distances of merges during the hierarchical clustering.



Optimal Number of clusters:

- Horizontal lines are cluster merges
- Vertical lines tell which clusters/labels were part of merge forming that new cluster
- Heights of the horizontal lines show the distance that needed to be "bridged" to form the new cluster
- The 19&51, 24&43 and 24&49 forms three base clusters. 24&43 and 24&49 are merged to form the fourth cluster.
- Looking at the dendrogram we cannot decide the number of clusters.
- By entering the Maximum number of clusters required as ‘Maxclust’ in fcluster, we can get the desired number of clusters.
- For selecting the number of clusters by distance, dendrogram is cut at the distance ‘20’. By doing this we get 3 clusters.
- The different linkage methods like Ward, Complete, Average and Centroid are also tested. All the tests have given the same number of clusters.

F-Cluster analysis:



f_cluster	1	2	3
spending	18.371	11.872	14.199
advance_payments	16.145	13.257	14.234
probability_of_full_payment	0.884	0.848	0.879
current_balance	6.158	5.239	5.478
credit_limit	3.685	2.849	3.226
min_payment_amt	3.639	4.940	2.612
max_spent_in_single_shopping	6.017	5.122	5.086
cluster_frq	70.000	67.000	73.000

- 210 observations are clustered into three groups with 70,67 and 73 in each group.
- The attractive customer segmentation for the Bank is Cluster 1, where, the customers spend more and make more advance payments. The risk pertaining to these segments are very low.
- These Cluster 1 type customers, tend to make less Minimum payment and more Advance payments. When they make minimum payments, they postpone the full payment. Since, there is less Minimum payment, the Probability if full payment is also high.
- The next productive group is Cluster 3 and the least group in terms of spending and payment is Cluster 2.

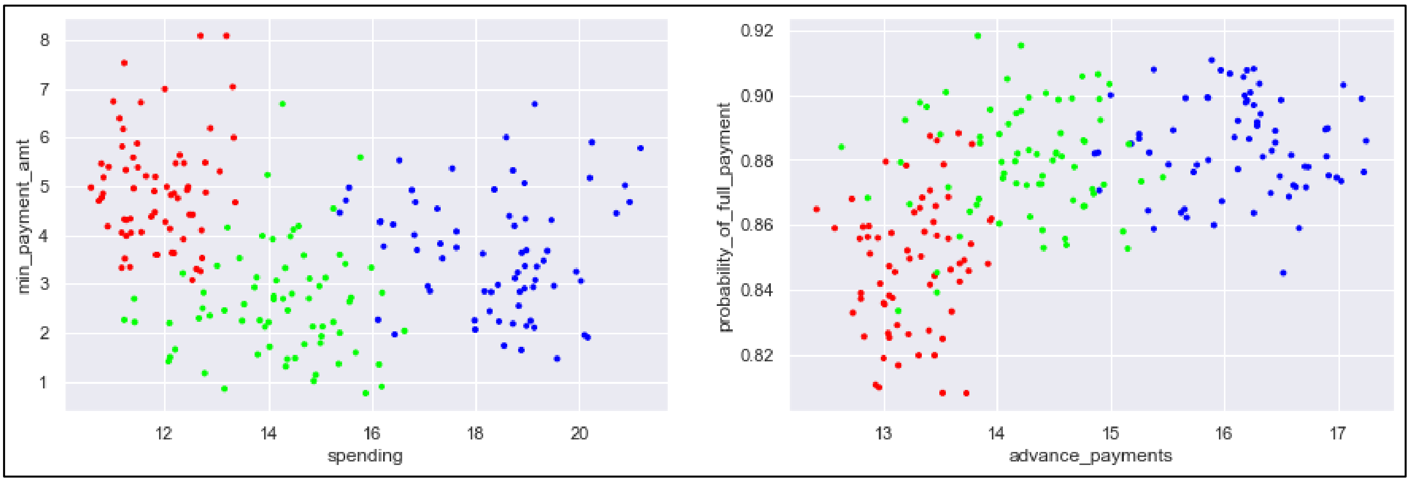
Visualisation of the Clusters:

Below plot shows the cluster segmentation in three different features, Spending, Probability of Full payment, Minimum Payment amount and Advance Payments.

Why choose these 4 features:

1. Spending: The more the customers spend more revenue for the Bank through Commission.
2. Probability of Full payment: Helps in Identification and reduction of defaulters.
3. Minimum Payment amount: Minimum amount due is the lowest amount you can pay on your credit card bill, to avoid late fees and other penalties. The minimum amount due is 5% of the total outstanding amount.

4. Advance Payments: The more the advance payments the larger is the probability of the full payment. Banks get to know their low-risk customer segments by looking into this.

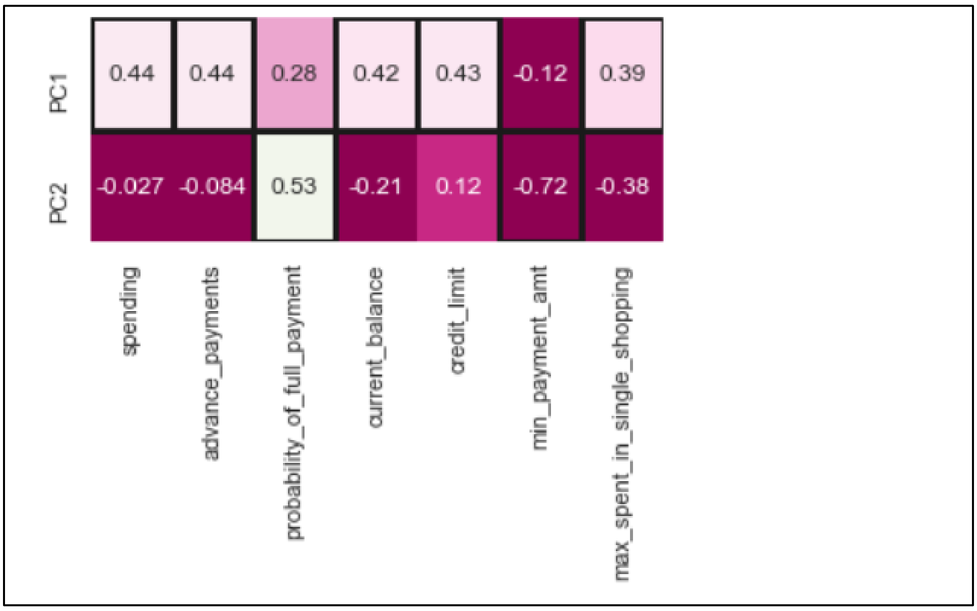


1. The clusters are clearly distinct from each other. There are few over-lapping's.
2. Blue dots represent Cluster 1 with high spending, minimum payment amount, high advance payments and high probability for full payments.
3. Green dots represent Cluster 3 with moderate range in all the presented features.
4. Red dots represent Cluster 2 which has low spending rate, high minimum payment rate, low probability of full payment and advance payments.

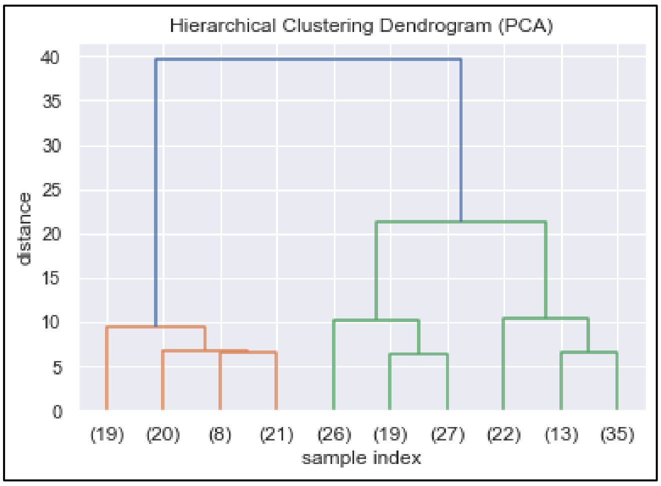
Hierarchical clustering on PCA applied Data (Correlations treated):

PCA is performed on the Scaled data. 2 components are chosen for further analysis. As 2 PCAs captures almost 100% variances in the data.

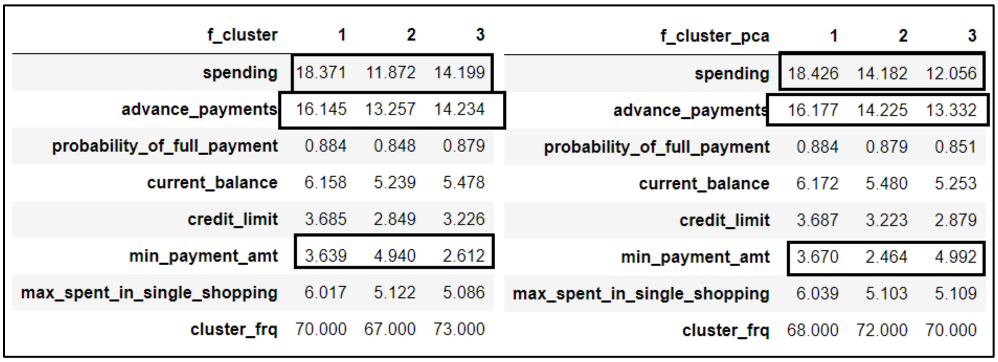
Below is the heatmap for the Captured variances of the Principal components.



<u>Dendrogram from PCA applied data:</u>



<u>Cluster Comparisons (Clusters from Original and PCA data):</u>



f_cluster	1	2	3
spending	18.371	11.872	14.199
advance_payments	16.145	13.257	14.234
probability_of_full_payment	0.884	0.848	0.879
current_balance	6.158	5.239	5.478
credit_limit	3.685	2.849	3.226
min_payment_amt	3.639	4.940	2.612
max_spent_in_single_shopping	6.017	5.122	5.086
cluster_frq	70.000	67.000	73.000

f_cluster_pca	1	2	3
spending	18.426	14.182	12.056
advance_payments	16.177	14.225	13.332
probability_of_full_payment	0.884	0.879	0.851
current_balance	6.172	5.480	5.253
credit_limit	3.687	3.223	2.879
min_payment_amt	3.670	2.464	4.992
max_spent_in_single_shopping	6.039	5.103	5.109
cluster_frq	68.000	72.000	70.000

- Almost all the features are clustered very similar to the clusters formed from the original data.
- The high correlations or the Multi-collinearity did not impact the output on a major scale.

Hence, clusters extracted from original data is selected for further cluster analysis.

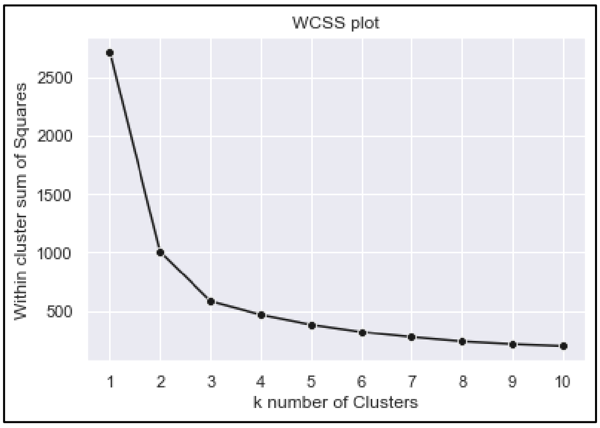
1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

<u>K-means Clustering Algorithm:</u>

It groups the data points based on their similarity or closeness to each other, in simple terms, the algorithm needs to find the data points whose values are similar to each other and therefore these points would then belong to the same cluster. the algorithm finds values between two data points by using the method of ‘Distance Measure’. Here distance measure is ‘***Euclidean Distance’***. The K-means uses the ‘***Centroid’*** method to link two clusters.

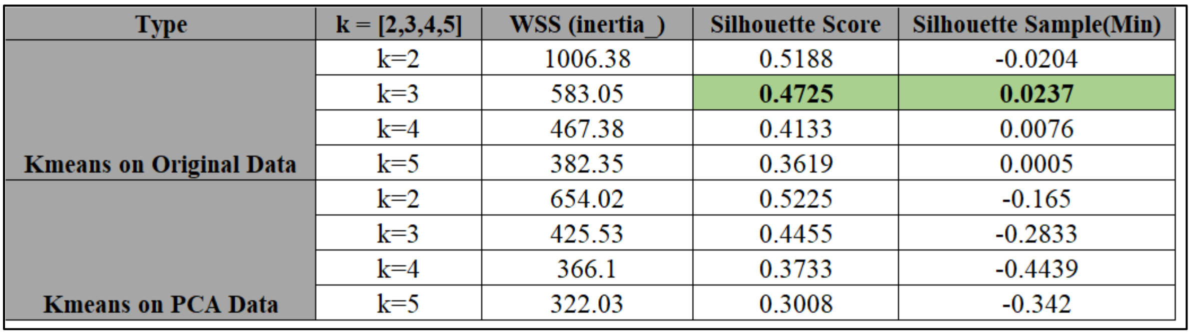
Choosing Optimal number of Clusters:

1. *Within Cluster Sum of Squares (WCSS Plot):*
    - Clustering algorithm for different values of k are performed. k = [2,3,4 and 5]
    - For each k within cluster sum of square is calculated.
    - WCSS curve is plotted according to the number of clusters.
    - The location of bend in the plot helps in identifying the optimal number of clusters.
    - In the WCSS plot attached below, the curve starts to flatten after the second bend that is formed for 3 clusters.
    - So, the optimal number of clusters from the Elbow method is 3.



2. *Silhouette Coefficients:*
    - *Silhouette Sample:* The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max (a, b).
    - The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters.
    - *Silhouette Score:* This returns the mean Silhouette Coefficient over all samples.

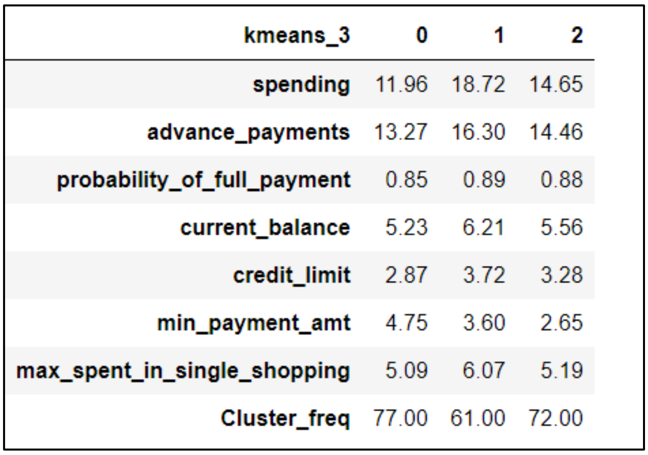
Table:1.4



Type	k = [2,3,4,5]	WSS (inertia_)	Silhouette Score	Silhouette Sample(Min)
Kmeans on Original Data	k=2	1006.38	0.5188	-0.0204
	k=3	583.05	**0.4725**	**0.0237**
	k=4	467.38	0.4133	0.0076
	k=5	382.35	0.3619	0.0005
Kmeans on PCA Data	k=2	654.02	0.5225	-0.165
	k=3	425.53	0.4455	-0.2833
	k=4	366.1	0.3733	-0.4439
	k=5	322.03	0.3008	-0.342

The silhouette score 0.518 when n cluster is 2, which is higher than the silhouette score when the n cluster is 3. But when we compare the minimum silhouette samples for both the cluster, for cluster=2 the minimum silhouette sample score is -0.0204 which is negative. Which means, there is a wrong assignment of an observation between the clusters. Whereas when we look into the silhouette sample score for cluster=3 is 0.0237. Hence, k = 3 is finalized.

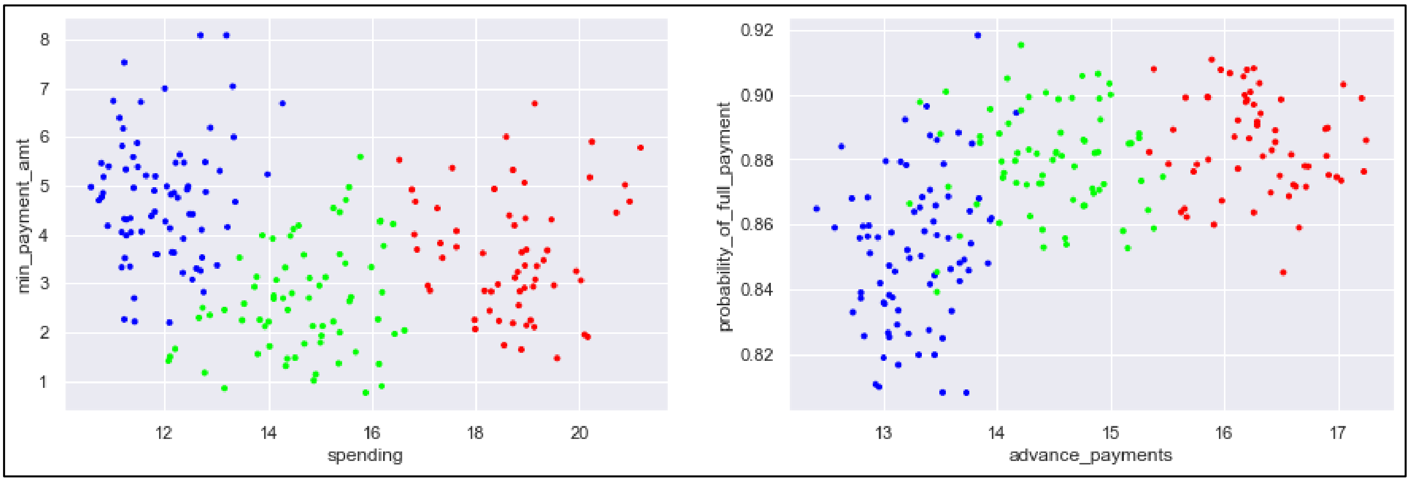
K-means cluster analysis:



kmeans_3	0	1	2
spending	11.96	18.72	14.65
advance_payments	13.27	16.30	14.46
probability_of_full_payment	0.85	0.89	0.88
current_balance	5.23	6.21	5.56
credit_limit	2.87	3.72	3.28
min_payment_amt	4.75	3.60	2.65
max_spent_in_single_shopping	5.09	6.07	5.19
Cluster_freq	77.00	61.00	72.00

- 210 observations are clustered into three groups with 77,61 and 72 in each group.
- The customers with high Spending (incl., Maximum spent in single shopping), high probability for full payment, high advance payments and with low minimum payment amount are categorized as cluster named ‘1’. They also have high credit limit and maintain high credit balance in their account.
- The Cluster type named ‘0’ are the customers who escape the monthly repayments by paying the Minimum payment amount. They also have lesser credit when compared with other two clusters.

Visualisation of K-means clusters:



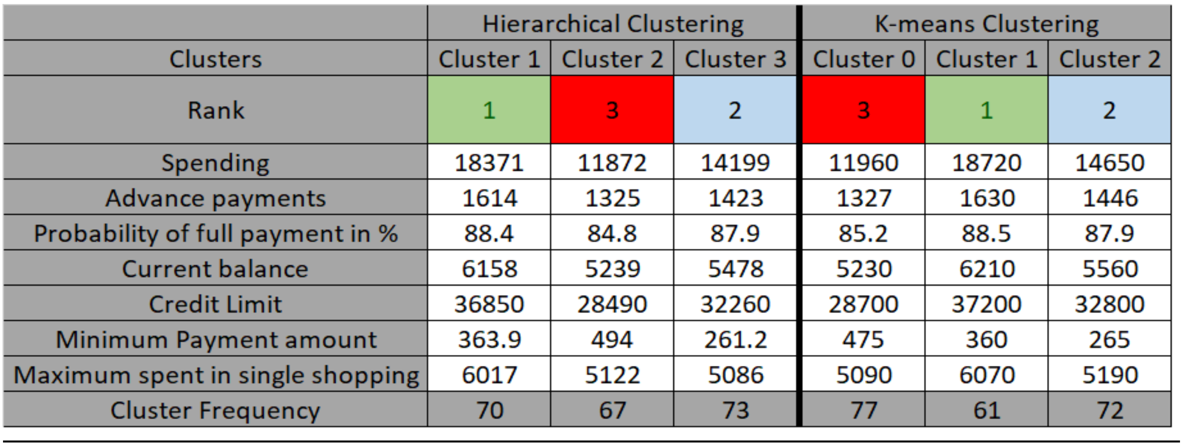
1. The clusters are significantly different from each other.
2. Red dots represent Cluster 1 with high spending, minimum payment amount, high advance payments and high probability for full payments.
3. Green dots represent Cluster 2 with moderate range in all the presented features.

4. Blue dots represent Cluster 0 which has low spending rate, high minimum payment rate, low probability of full payment and advance payments.

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

<u>Comparison of Hierarchical Clustering and K-means Clustering:</u>

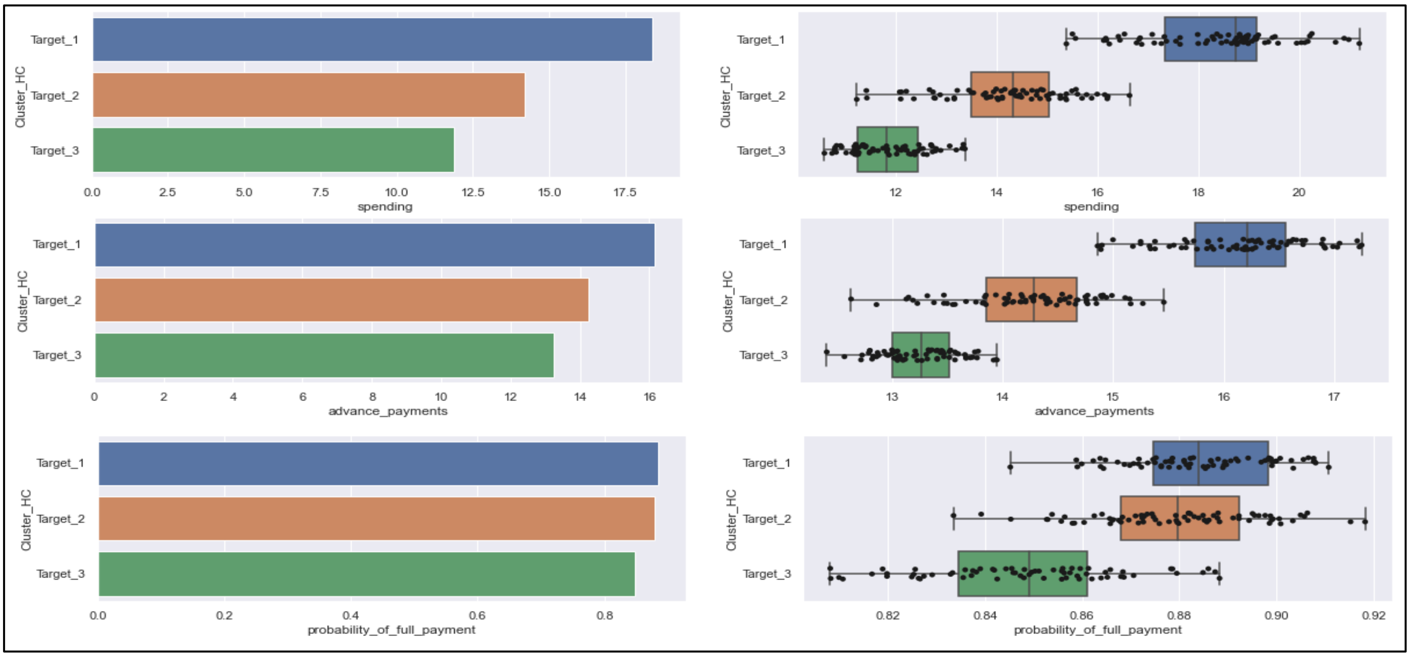
<u>Table 1.5:</u>

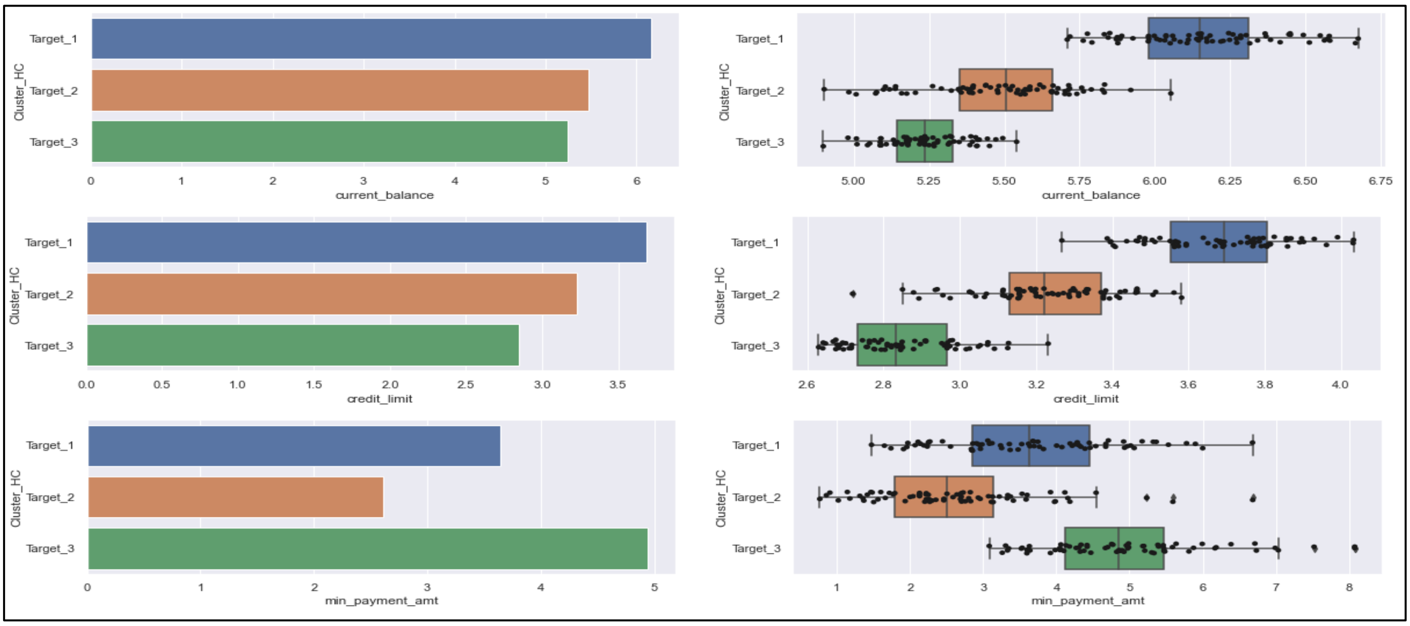


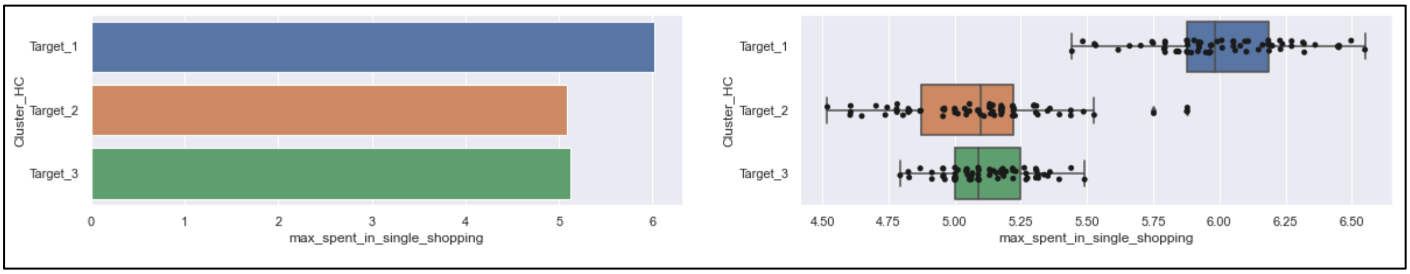
	Hierarchical Clustering			K-means Clustering		
Clusters	Cluster 1	Cluster 2	Cluster 3	Cluster 0	Cluster 1	Cluster 2
Rank	1	3	2	3	1	2
Spending	18371	11872	14199	11960	18720	14650
Advance payments	1614	1325	1423	1327	1630	1446
Probability of full payment in %	88.4	84.8	87.9	85.2	88.5	87.9
Current balance	6158	5239	5478	5230	6210	5560
Credit Limit	36850	28490	32260	28700	37200	32800
Minimum Payment amount	363.9	494	261.2	475	360	265
Maximum spent in single shopping	6017	5122	5086	5090	6070	5190
Cluster Frequency	70	67	73	77	61	72

- The above table shows the difference between the mean values of clusters formed from hierarchical and k-means clustering.
- The clusters are ranked according to the values captured by them.
- Clusters formed by using both the techniques shows very slight difference. Indeed K-means has captured slight bigger numbers.
- Looking at the cluster frequencies in hierarchical clustering, Target 1 has 70 whereas the Target 1 in K-means has only 61 customers. These customers are marked as Target 2 in K-means.
- The problem statement is about giving promotional offers. If these customers are not given promotional offers, there are lot of chances in losing them.
- They might change their Servicing bank to the bank which gives more promotional offers.
- Hence, the cluster profiling is proceeded with the clusters formed from hierarchical clustering technique where the model has captured a greater number of Max spenders and Max payers as Target 1.

The Bar-plot and Box-plot + Strip-plot analysis for all the columns among 3 clusters:



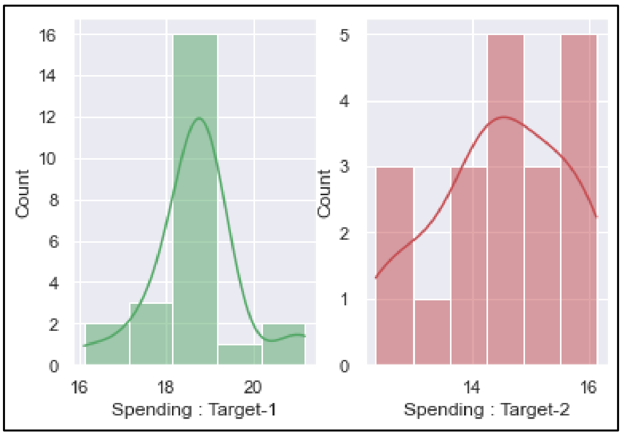


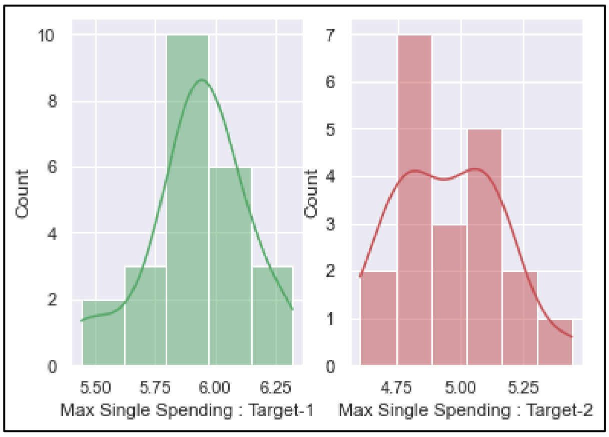


Cluster Profiling and Analysis:

The clusters formed from hierarchical clusters are taken for analysis. The clusters are merged with the original data. Clusters with all high values are classified as Target 1, moderate values are classified as Target 2 and least values are classified as Target 3.

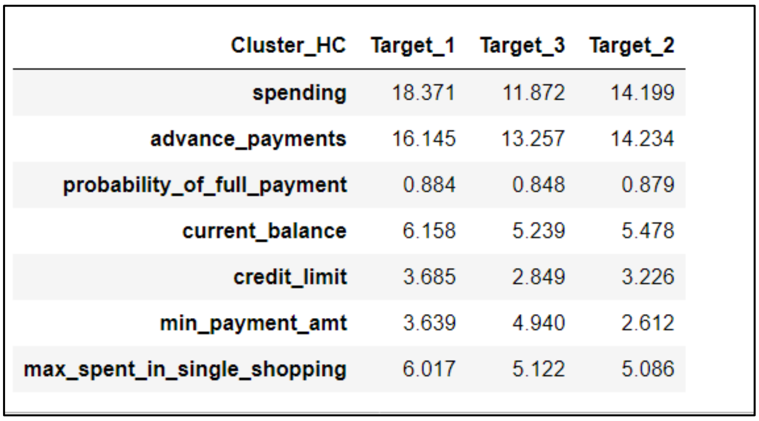
The customers who have more than an average of 89% of probability for full payment is found under both Target 1 and Target 2 profile. 24 in Target 1 and 20 in Target 2. The clusters are segmented on the basis of the 'Spending' and 'Max spend in single shopping feature'. As credit card is all about more spending and more rewards, the target types are clustered in a meaningful way. Below is the graphical representation of the Spending and the Max Single Spending feature for the customers who have more than an average of 89% probability of payment.





- Target-1 profile spends more than 18000 on an average whereas Target-2 spends ranges between 14000 to 16000 on an average.
- Target-1 profile spends more than 6250 on an average on single spend whereas Target-2 profile ranges between 5000-5250 on an average.

Clusters renamed as (Target1, Target2 and Target 3):



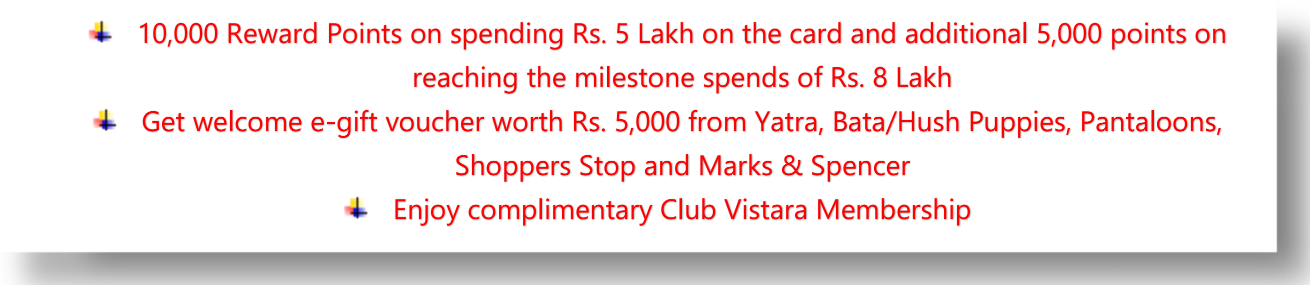
Cluster_HC	Target_1	Target_3	Target_2
spending	18.371	11.872	14.199
advance_payments	16.145	13.257	14.234
probability_of_full_payment	0.884	0.848	0.879
current_balance	6.158	5.239	5.478
credit_limit	3.685	2.849	3.226
min_payment_amt	3.639	4.940	2.612
max_spent_in_single_shopping	6.017	5.122	5.086

Target 1: Max Payers and Max Spenders:

The Banks don’t make much money from Max payers as they do from other customers. They are still profitable customers to have around as they spend more. Bank charge stores around 3% for each transaction, so even when they are paying Max Payers a 1% cash back reward, the credit card companies are still making money. Max Payers are important to Bank because it allows them to lend out more money. If they don’t pay back their dues on time, banks will not have sufficient funds to lend. It is very important to retain such low-risk profile customers by offering:

1. Reward points on every spend.
2. Bonus Reward points on renewal of cards.
3. Excessive travel benefits like air miles on every spend, lounge access, discounts and other travel deals.
4. As they spend more using credit cards, deals like Amazon purchase offer, Zomato membership offer, fuel surcharge waiver on all transactions, movie ticket vouchers., etc.
5. Exclusive shopping vouchers on Premium brands.
6. Lifetime free add-on credit cards for family (spouse, children above 18 years of age, siblings and parents)
7. Increasing their credit limit and upgrading to advanced or premium cards.
8. Retention of Target -1 customers are important and tough too. As they are good spenders and payers, they are expected to look for other credit cards with a greater number of offers and deals. By offering Free annual maintenance and renewal fee, banks can retain and gain profit on such customers.

Below are the credit card offers for Max Spenders from various Banks. Source: Top 10 Best Credit Cards in India 2021 - SBI, HDFC, ICICI, Axis, HSBC - 25 April 2021 (paisabazaar.com)



- 10,000 Reward Points on spending Rs. 5 Lakh on the card and additional 5,000 points on reaching the milestone spends of Rs. 8 Lakh
- Get welcome e-gift voucher worth Rs. 5,000 from Yatra, Bata/Hush Puppies, Pantaloons, Shoppers Stop and Marks & Spencer
- Enjoy complimentary Club Vistara Membership

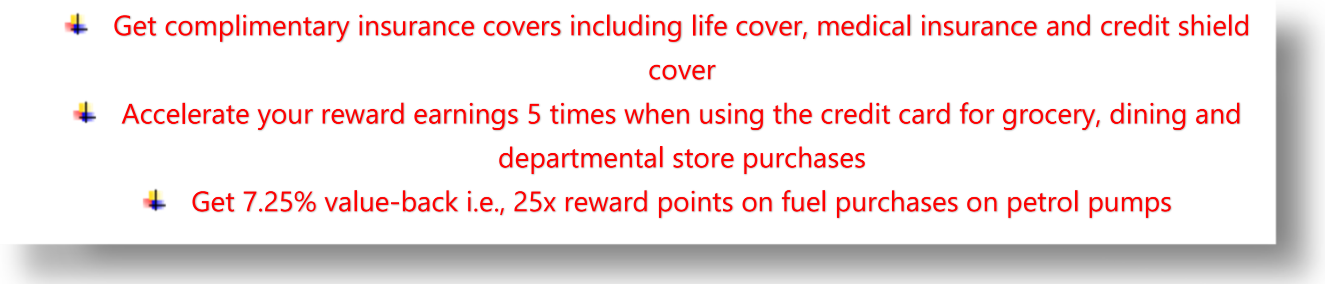
<u>Target 2: Max Payers and Less Spenders:</u>

The target 2 type customers are similar to Target 1 type. They make more advance payments and have high probability of making full payments. Like Target 1, Banks don’t gain any interest income over these types of customers. But again, they gain from the commission from third party when the customers spend using the credit cards. From the given data set we can understand that they spend less than Target 1. Targeting this type of customers with the following offers banks may boost their spending and earn their loyalty too:

1. To attract the customers who are concerned about the rising fuel costs banks can offer more reward points on every fuel purchase.
2. As they are low spenders, providing attractive offers on spending on groceries, Utility bills and departmental stores can boost their spending rate.
3. Cashbacks for Equated Monthly Instalments (EMI)
4. Complimentary Insurance and Accidental cover may help to win their loyalty.

The customers of Target 2 fall under low-risk profile and maintaining such customers enables bank to lend money as these customers pay on or before the due.

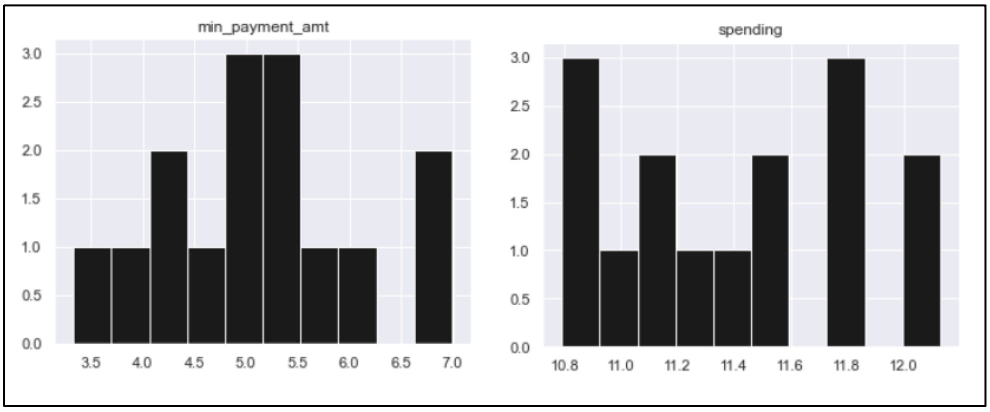
Below are some of the credit card offers for Target 2 customers: Source: [Top 10 Best Credit Cards in India 2021 - SBI, HDFC, ICICI, Axis, HSBC - 25 April 2021 (paisabazaar.com)]



- Get complimentary insurance covers including life cover, medical insurance and credit shield cover
- Accelerate your reward earnings 5 times when using the credit card for grocery, dining and departmental store purchases
- Get 7.25% value-back i.e., 25x reward points on fuel purchases on petrol pumps

Target 3: Revolvers / Always Carry debt:

These are the favourite customers for banks. These customers only make the minimum payment or slightly more, and often stretch their payments out for years. Their debt may be large or small, but it is rarely paid off in full. The interest rates and fees these users pay are the bread and butter for the bank. Below is the graphical representation which shows customers who have less than 83% probability for full payment have record of high Minimum payment amount. These customers spend less, make very few advance payments. Their credit limit is also very less.



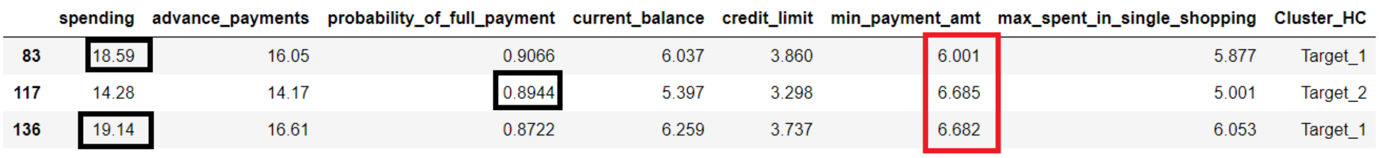
1. As banks benefits more from Target 3 profile customers through interest income, they can generally offer all the promotional offers and discounts that are specified for Target 1 and 2.
2. Customers who have had very bad experience with credit card debt will tend to stay out of the credit card by stop using it. Banks cannot miss such customers; it should offer special incentives to bring them back.
3. Credit card Balance transfer: When a customer has outstanding balance on a credit card and are not able to pay in full, he/she may opt to balance transfer the outstanding balance to other bank's credit card at a lower interest rate. By this way, Revolvers or Minimum amount payers can be targeted by offering less interest rates even when they are using other bank credit cards.

Below are some of the credit card offers for Target 3 customers: Source: Top 10 Best Credit Cards in India 2021 - SBI, HDFC, ICICI, Axis, HSBC - 25 April 2021 (paisabazaar.com)

- 0% intro APR for 15 months on qualifying balance transfers
- 2,500 Reward Points on payment of the joining fee and renewal fee every year
- Get additional 3,000 Air Miles on card renewal
- Credit card balance transfer up to Rs. 5 Lakh from other cards at lower interest rate

Unique Customer Profiles Observed:

The Target 3 profile has high Minimum payment amount whereas Target 1 and 2 has less Minimum payment amount. The below 3 observations has high Minimum payment amount yet falls under Target 1 and 2. Reasons below:



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Cluster_HC
83	18.59	16.05	0.9066	6.037	3.860	6.001	5.877	Target_1
117	14.28	14.17	0.8944	5.397	3.298	6.685	5.001	Target_2
136	19.14	16.61	0.8722	6.259	3.737	6.682	6.053	Target_1

- Customers under Target 1 has high spending rate and their probability of full payment is also above 87%. They also make high advance payments.
- Customer under Target 2 has high rate of probability of full payment.
- This type of customers spends more, pays on time and also utilises the time extension facility by paying the minimum payment amount.
- From Bank point of view, they are the most profitable profiles. Banks gets more commission from third parties as they spend more and they get interest income as the customers postpone the payments also. By making full payment before the next month bill due they fall under ‘low-risk profile’ too.
- They are *‘Max Spenders – Max payers – Revolvers – Low risk type’* customers.

Conclusion:

Revolvers and Max Payers are the most profitable customers to have. In fact, all the three target profiles are between these two categories. As the chance of defaulting is low as the least probability of full payment is 80%. In addition to the specific offers and discounts listed for target types, there are some more offers which should be given to all the customers to attract and retain them. They are,

1. Cashback on bill payments and mobile recharges.
2. No maximum limit on the cashback earned
3. Welcome gift vouchers.
4. Zero joining fee
5. Zero lost card liability, emergency card replacement, global card access, cash advance.

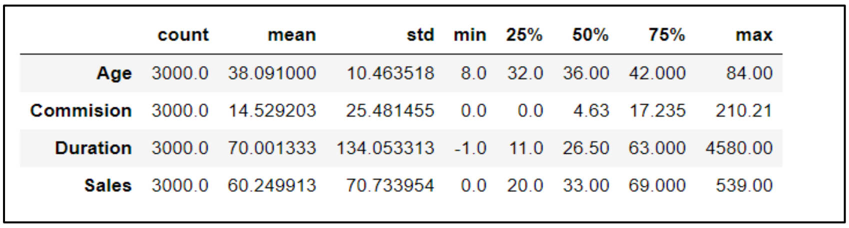
Problem 2: CART-RF-ANN

Problem Statement: A tour travel insurance is facing high claims. By executing the model like Decision Trees, Random forest and Neural network the number of claimed and non-claimed cases are predicted. The predicted numbers and actual number are analysed and the reason for the high claims are analysed further.

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

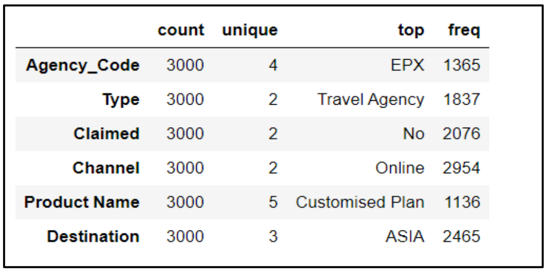
Data Description:

1. There are 3000 observations and 10 features in the data.
2. No missing values in the data set.
3. There are 6 Discrete columns and 4 Continuous columns.
4. The total number of duplicates arrived is 139. As there is no unique ID or Customer name, the duplicates found cannot be considered as duplicates. Hence, the duplicates are not treated.
5. Description of Numerical columns:
   - 75% of observations fall under the age group below 42 years. 4 observations are above the 80 years.
   - The difference between the mean and the median for commission and sales are very high due to existence of outliers.
   - The duration has a minimum value of -1. Duration cannot be in negative. The value is replaced as 1.
   - The maximum value in Duration feature is 4580 which does not correspond with the Sales and Commission rate. Hence, it is imputed with the upper range of the outliers.



	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00

6. Description of Nominal columns:
   - The unique values and the top most category is displayed in the below description.

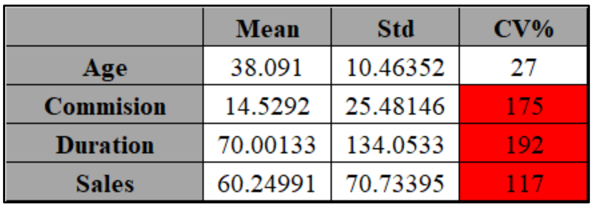


	count	unique	top	freq
Agency_Code	3000	4	EPX	1365
Type	3000	2	Travel Agency	1837
Claimed	3000	2	No	2076
Channel	3000	2	Online	2954
Product Name	3000	5	Customised Plan	1136
Destination	3000	3	ASIA	2465

Coefficient of Variation:

Below is the CV% of different features. The lesser the percentage is, it means it has less variability and the data is stable. The CV% are very high except for the Age column, hence the data has high variations and less consistent.

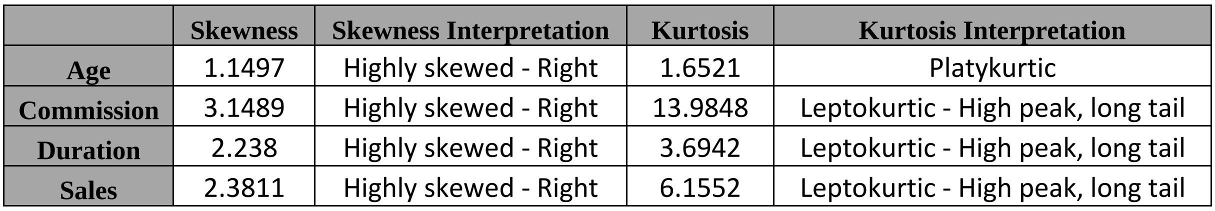
Table:1.1



	Mean	Std	CV%
Age	38.091	10.46352	27
Commision	14.5292	25.48146	175
Duration	70.00133	134.0533	192
Sales	60.24991	70.73395	117

Skewness and Kurtosis:

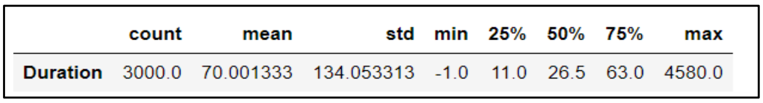
Table:1.2



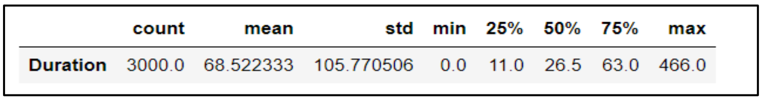
	Skewness	Skewness Interpretation	Kurtosis	Kurtosis Interpretation
Age	1.1497	Highly skewed - Right	1.6521	Platykurtic
Commission	3.1489	Highly skewed - Right	13.9848	Leptokurtic - High peak, long tail
Duration	2.238	Highly skewed - Right	3.6942	Leptokurtic - High peak, long tail
Sales	2.3811	Highly skewed - Right	6.1552	Leptokurtic - High peak, long tail

Anomalies Treatment:

1. The negative value in duration column is replaced with 1.
2. An extreme value in duration column which does not correspond with the Sales and Commission rate is replaced with the upper range of the outliers.
3. Before and after the anomaly's treatment is shown below.



	count	mean	std	min	25%	50%	75%	max
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.5	63.0	4580.0



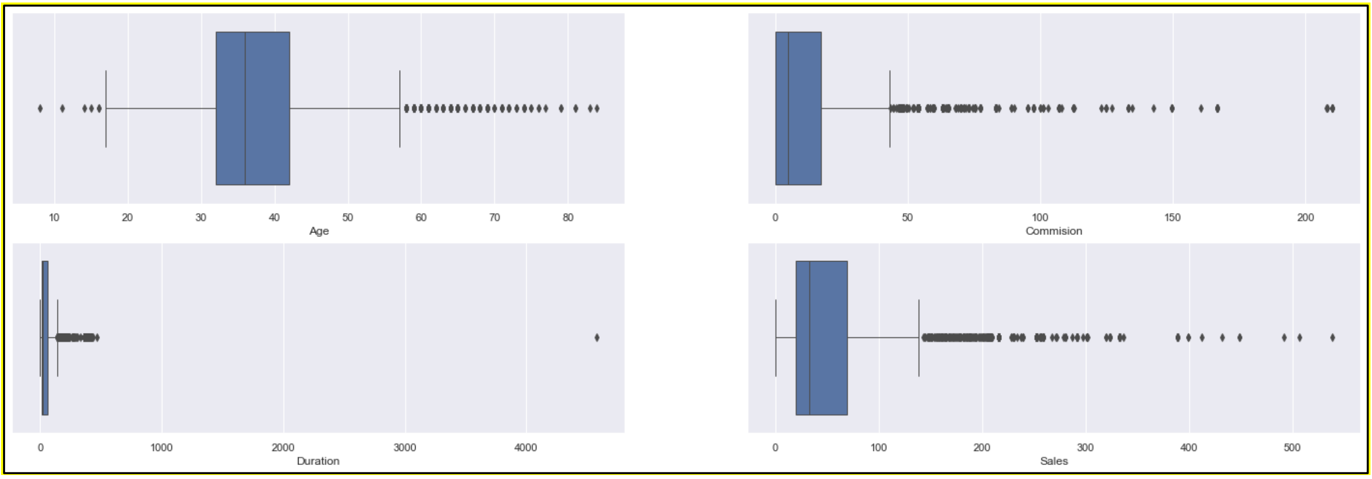
	count	mean	std	min	25%	50%	75%	max
Duration	3000.0	68.522333	105.770506	0.0	11.0	26.5	63.0	466.0

Outliers Treatment and Proportions:

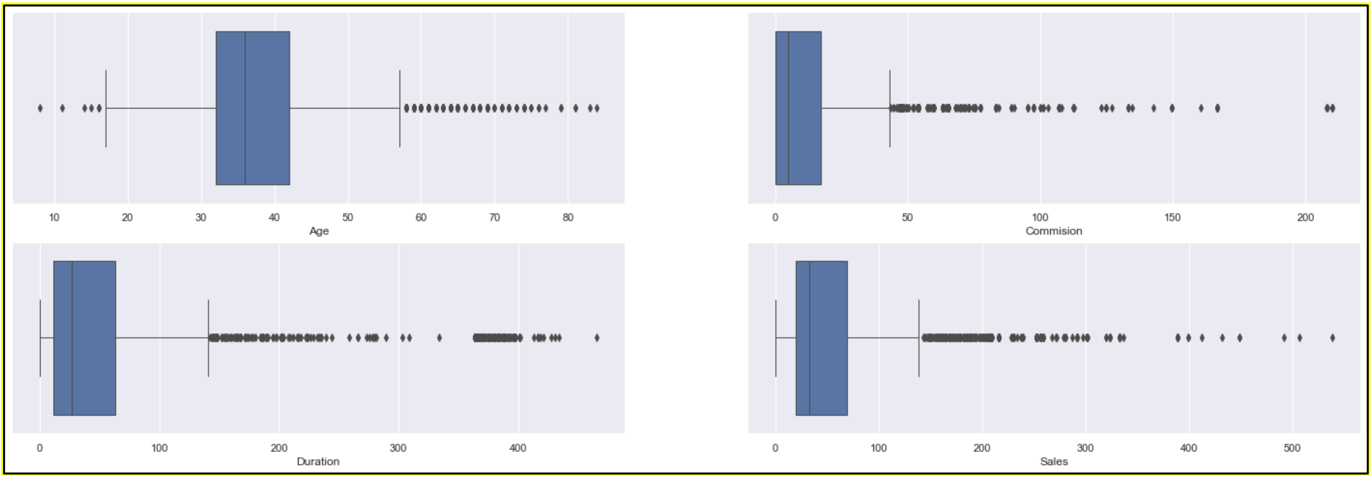
1. Only Age column has outliers below the lower range.
2. The other all columns have outliers above their upper range. The percentage of the outliers is shown in the below table.
3. Percentage of Outliers in all other numeric columns are below 13%.
4. As the given data is about an Insurance claim, we may have many unique scenarios. Treating the outliers may generalize them.
5. Also, CART Model and Random models are capable of handling Outliers.

6. The given data is also analysed by treating outliers. There is no strong significant difference between both the data sets. The comparison will be shown at the end of the report.
7. Hence, Original data without any treatment on outliers is used for further modelling.

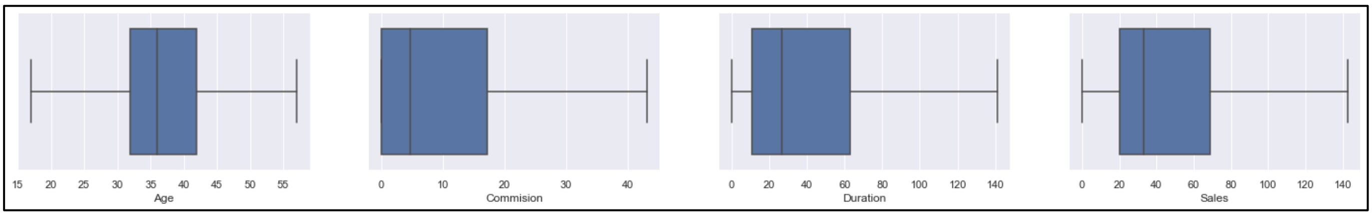
Before Treating the Extreme outlier in Duration feature:



After Treating the Extreme outlier in Duration feature:

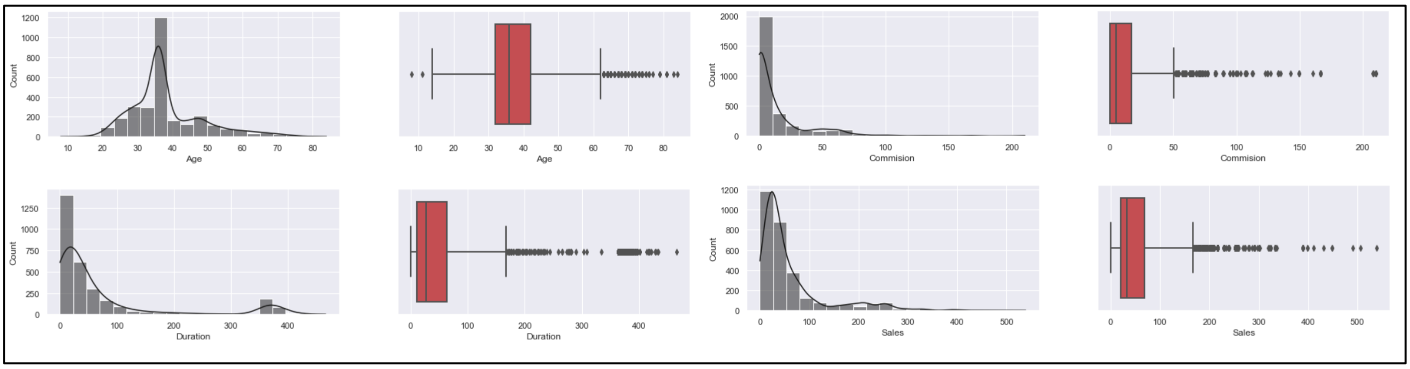


Graphical representation of the Boxplot from the Outliers treated data set:



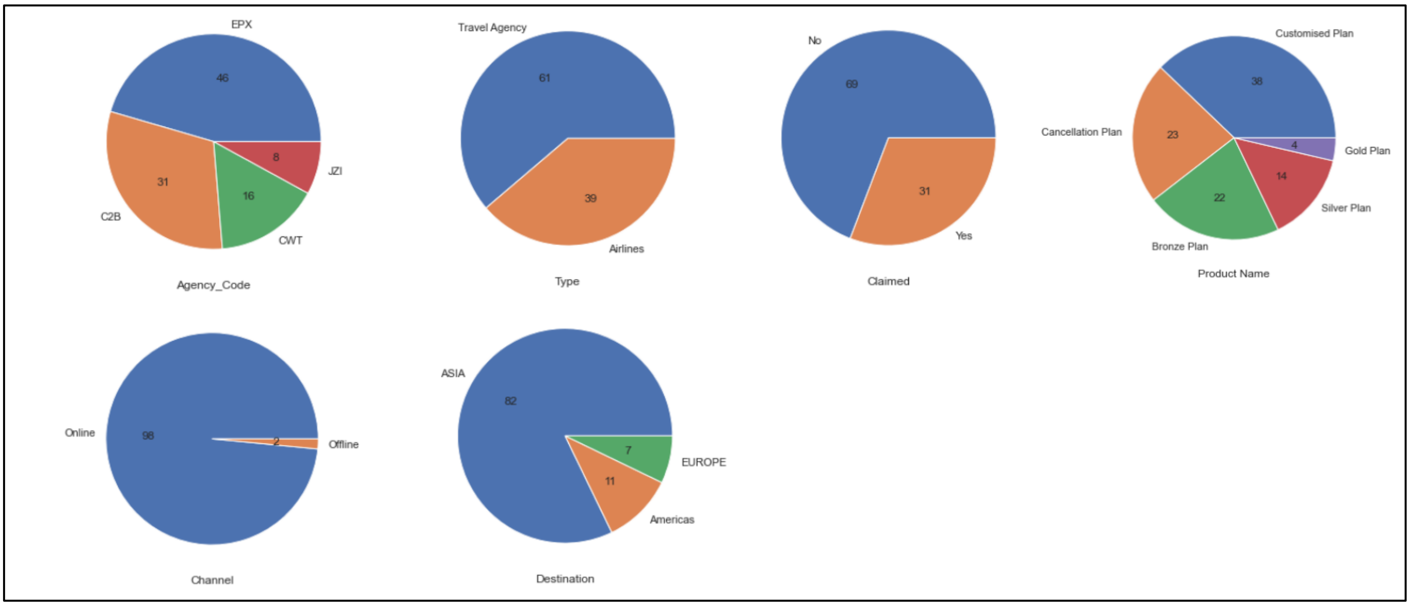
<u>Univariate Analysis:</u>

<u>On Numeric Columns: [Histogram and Boxplots representation for the Numeric Columns]</u>



- All the features are rightly skewed and have outliers above the upper range.
- The average duration for travel is less than 70 days.

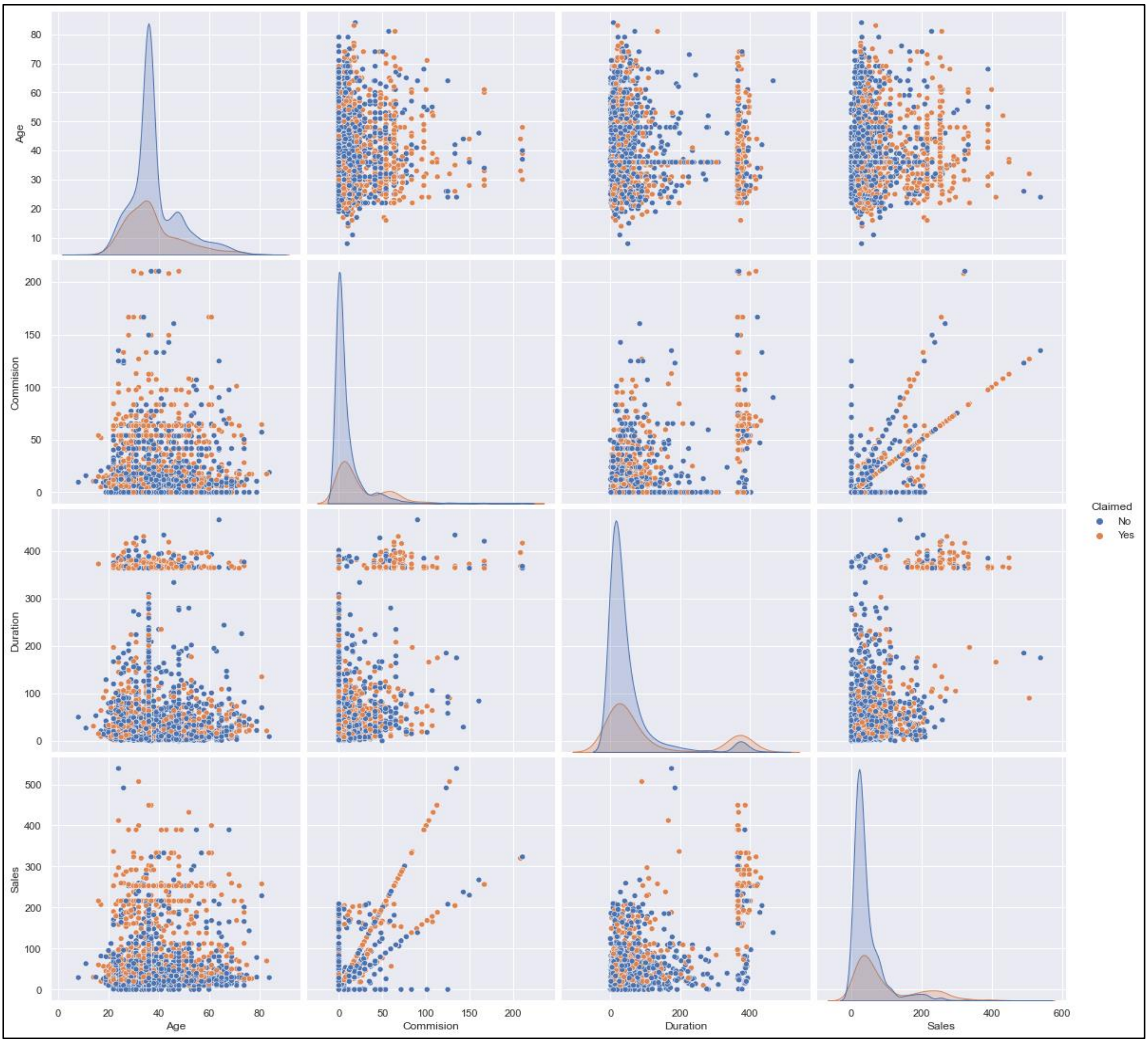
<u>On Nominal Columns: [Pie plot representation for the Nominal Columns]</u>



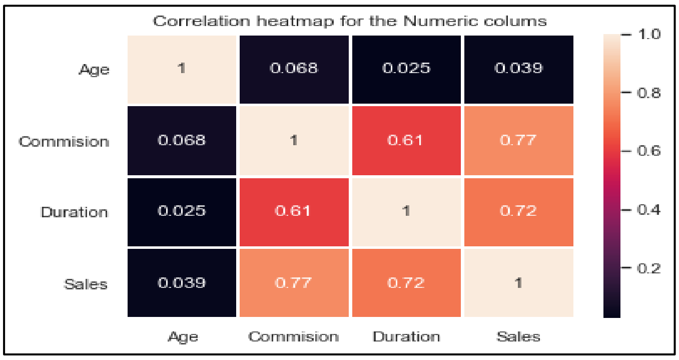
1. EPX and CWT are Travel agencies and JZI and CWT are Airlines.
2. 61% of the people has used Travel agency where as only 39% used Airlines.
3. Total number of claims in the data is 31% and the non-claimed cases accounts to 69%.
4. The Customized, Cancellation plan and Bronze plan are the commonly used plans.
5. Almost 98% are comfortable with Online distribution channel. The travel agencies operate majorly through online channel. The CWT airline has a bit higher offline channel.
6. Most of the people have travelled to ASIA. The second most travelled destination is America.

Bi-Variate Analysis:

Pair-plot representation using Scatter for all Continuous Variables:

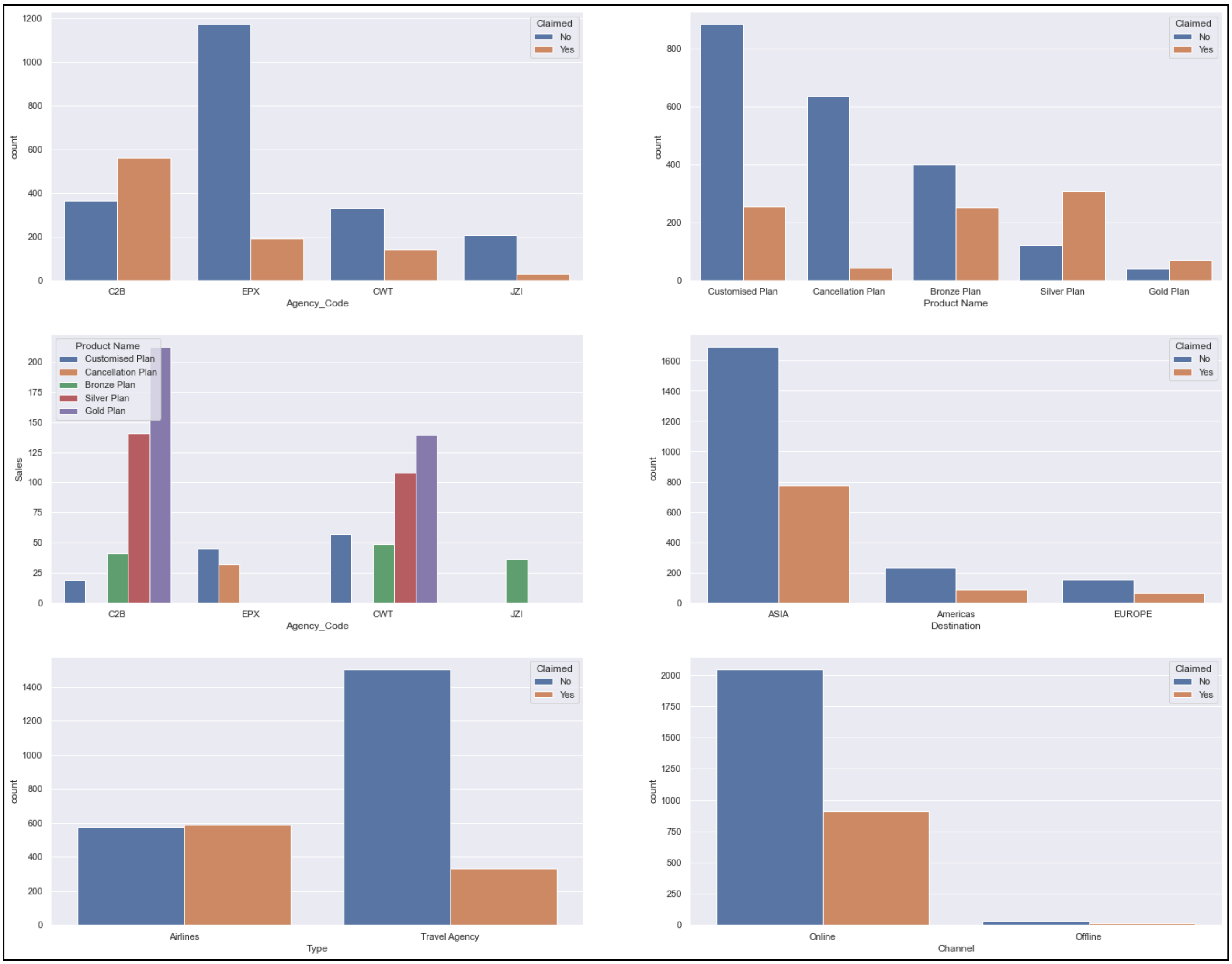


Correlation - Heatmap using Seaborn:



- The Duration, Commission and sales are highly correlated. The higher the duration of travel, the Sales and Commission are also high.
- Commission and Sales are strongly correlated. The Commission for tour insurance firms is based on the sales of the tour insurance policies.
- The Age variable is negatively correlated with all the features. Which means, age does not affect the duration, sales and commission.
- Generally, random forest model has accurate predictions on low correlated data. Here we have highly correlated features. For this particular project, the correlations are not treated.

__Multi-variate Analysis:__



- The travel insurance from agency EPX is high. Whereas a high number of cases are claimed from the C2B agency.
- The most commonly used plans are Customized plan and Cancellation plan. Silver plan has high number of Claimed cases.
- On the Sales perspective, C2B and CWT are with high sales numbers.

- The most visited destination is Asia and hence, there are high number of claimed cases from the travellers, travelling to Asia.
- It is through Travel agencies travellers prefer to book tickets. The percentage of the claimed cases in Travel agencies is 18% whereas for Airlines it is 50%.
- Out of 46 insurances through offline channel, 17 are claimed. 30% of Online channel insurances are claimed.

For Multivariate analysis, Cluster analysis and PCA are challenging to apply to this data set. As we have both continuous and categorical variable. Both the techniques work good on continuous variables.

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

Preparing the data for Modelling:

1. The categorical or discrete features' labels are encoded. For Gender, Claimed and Channel the features are encoded using labels.
2. For Agency code, Type, Product name and Destination One hot encoding is used.
3. The train and test data are not scaled for CART and RF model. For Neural network modelling the data is scaled.
4. Outliers are not treated. A cross check without outliers is simultaneously performed and the table of scores is shared.
5. The train and test data are separated using the ideal ratio 80:20. The more data in training, it understands different patterns and helps in predicting the test data.
6. Random state is 123 throughout the model.
7. The dimensions of the train and test data (2400 and 600)
8. The models are compared on the basis of different evaluation techniques, like
    - Accuracy Score: This score is the measure of how accurate the model is.
    - Confusion Matrix: A specific table layout that allows visualization of the performance of an algorithm. It includes True Negatives, True Positives, False Negatives and False positives.
    - Classification report: Classification report is used to evaluate a model's predictive power. It gives the accuracy, Precision (Ratio between True positives and (True Positive + False Positive), Recall (Ratio between True positives and (True Positives + False Negatives).
    - AUC Score: ROC-AUC score is one of the major metrics to assess the performance of a classification model.

- ROC Curve: A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

CART / Decision Tree Model:

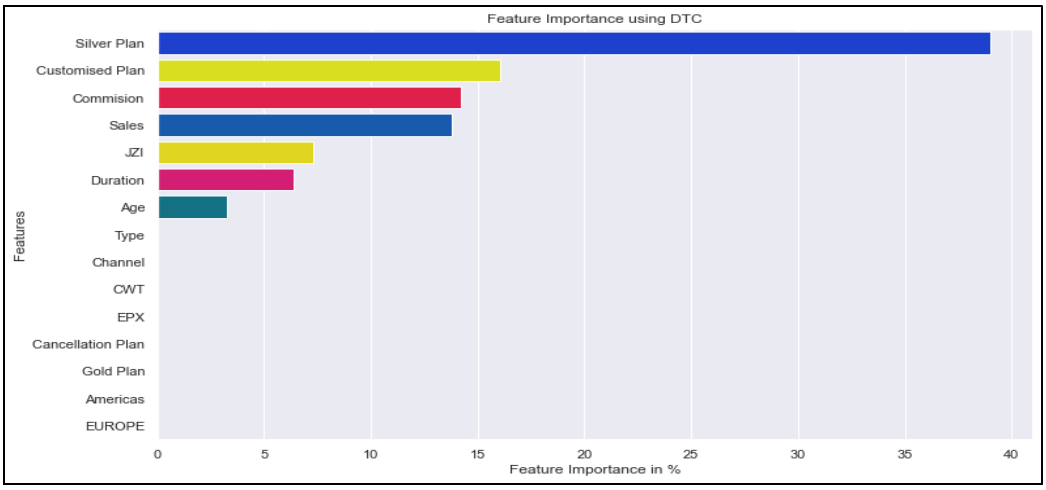
- The model is created using a Grid search CV function.
- Gini: It gives the probability of incorrectly labelling a randomly chosen element from the dataset if we label it according to the distribution of labels in the subset. Feature with lowest Gini index is treated as decision node.
- Min_Sample_Leaf: The minimum number of samples required to be at leaf node.
- Min_Sample_Split: specifies the minimum number of samples required to split an internal node.
- Max depth: The number of splits that each decision tree is allowed to make
- The Parameter grid are: 'criterion': ['gini'], 'max_depth': [7, 8],'min_samples_leaf': [10, 20, 30],'min_samples_split': [50, 80, 100]
- Best Parameters for CART model are***: Criterion: 'gini', max_depth: 7, min_sample_leaf: 30, min_sample_split:60.***

Logical reasons for choosing the Parameters:

1. Gini: It aims to decrease the level of entropy from the root nodes to the leaf nodes of the decision tree. In this way, the Gini Index is used by the CART algorithms to optimise the decision trees and create decision points for classification trees.
2. Min sample Leaf: 1% of the size of the data. (3000*1%)
3. Min Sample split: 3 times larger than the minimum sample leaf.
4. Max depth: The un pruned tree had 15 splits. This led to overfitting. Hence, half of the depth from the original model is considered for pruned decision tree.

Feature Importance from CART model:

Decision trees make split that maximize the decrease in impurity. By calculating the mean decrease in impurity for each feature across all trees we can know the feature importance. The top three features that CART model finds significant are Silver plan, Customised plan and Commission. Below is the graphical representation for the feature importance from CART model.



Random Forest Model:

Random Forest grows multiple decision trees which are merged together for a more accurate prediction. The key here lies in the fact that there is low (or no) correlation between the individual models—that is, between the decision trees that make up the larger Random Forest model. While individual decision trees may produce errors, the majority of the group will be correct, thus moving the overall outcome in the right direction.

The given data has high correlations between the variables. A separate analysis on data for removing the correlations using the PCA (n_components = 2) is done and cross checked with the original values. There is no significant difference between the models. As the scores are slightly higher than the model performed on PCA. The model performed on the original data is taken for further analysis.

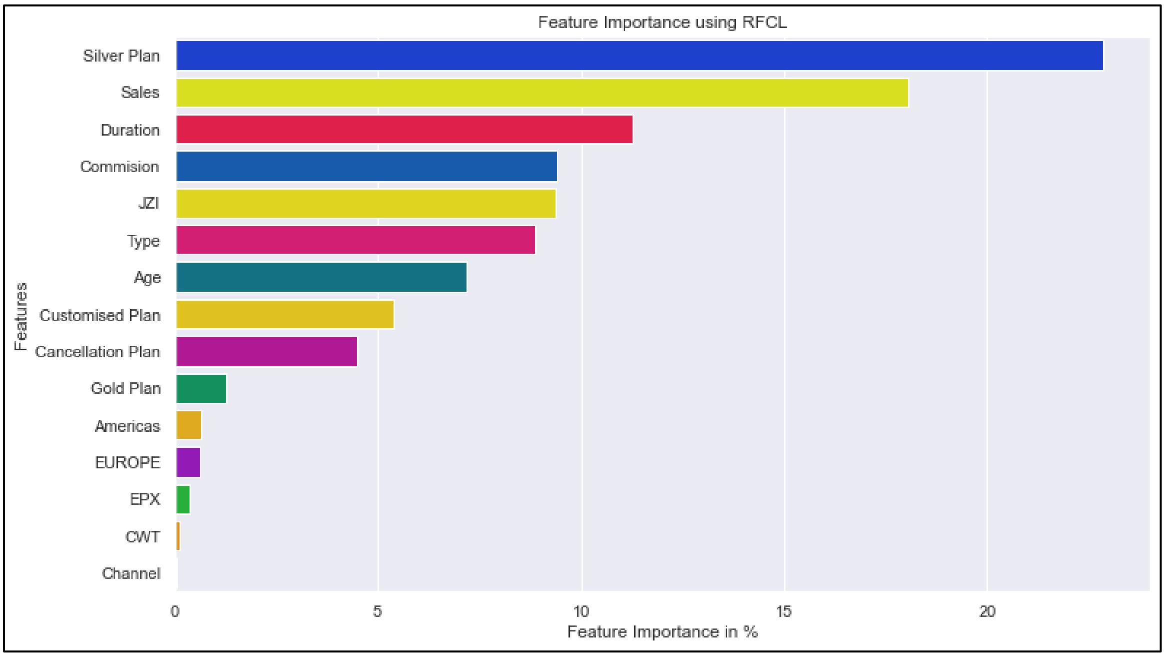
- Parameter’s grid: 'max_depth': [15],'max_features': [15],'min_samples_leaf': [10, 20, 30],'min_samples_split': [40, 60, 80],'n_estimators': [100,300]
- Best Parameters: 'max_depth': 15, 'max_features': 15, 'min_samples_leaf': 10, 'min_samples_split': 40,'n_estimators': 300
- n_estimator: Parameter controls the number of trees inside the classifier. Adding more trees just stabilizes the results. Hence, 300 is preferred.

- Max_features: Number of features to take into account in order to make the best split. Since the data has only 15 features, all the 15 features are taken for analysis.

- The accuracy score improved with a smaller number of samples required for leaf nodes and sample split. Hence, lower values are preferred.

Feature Importance from RF model:

The top three features from RF model feature importance are Silver plan, duration and Sales. The RF model has captured the importance for many features unlike Decision tree. The least important features are EPX, CWT, Channel. The least importance features are removed and a revised random forest classifier is performed. The accuracy scores and other parameters did not show a significant difference between the models.



Building Neural Network model:

1. Data is scaled using the Standard scalar before starting the model. Neural network works best on a scaled data.
2. Neural Networks are complex structures made of artificial neurons that can take in multiple inputs to produce a single output**.**
3. The Parameter's grid: 'hidden_layer_sizes': [(100, 100, 100)],'max_iter': [5000], 'solver': ['adam']'activation': ['relu'],'alpha': [0.05],'learning_rate': [adaptive],'batch_size': [500], 'tol': [0.0001].
4. ***Best Parameters: 'activation': 'relu', 'alpha': 0.05,'batch_size': 500, 'hidden_layer_sizes': (100, 100, 100),'learning_rate': 'adaptive', 'max_iter': 1000, 'solver': 'adam', 'tol': 0.0001***

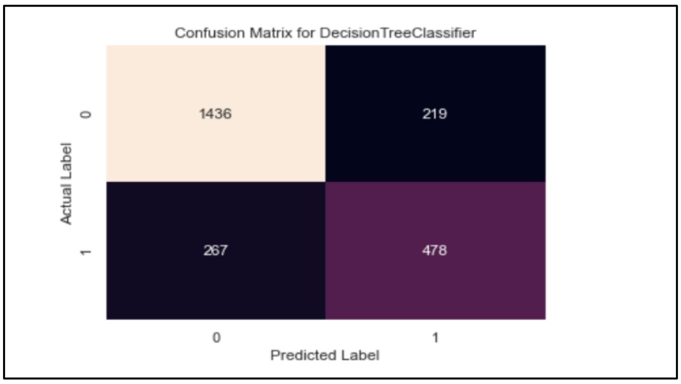
5. Max_iterations: The maximum number of iterations that you want to allow during network training.
6. Solver: Why Adam? The Adam Solver uses an extension to stochastic gradient descent. It uses the squared gradients to scale the learning rate and it takes advantage of momentum by using moving average of the gradient instead of gradient. This allows the solver to work quickly by seeing less data and can work well with larger data sets.
7. Activation: Why Relu? Its main advantage is that it avoids and rectifies vanishing gradient problem and less computationally expensive than tanh and sigmoid.
8. Learning Rate: Learning Rate is an important hyper-parameter that has to be tuned optimally for each feature in the input space for better convergence. By adopting Adaptive Learning Rate, we let these optimizers tune the learning rate by learning the characteristics of the underlying data
9. Batch size: The total size of the data is 3000. Batch size of 500 will divide the data into 6. 6 iterations for 1 epoch.
10. Tolerance: Lower the value more the accuracy.

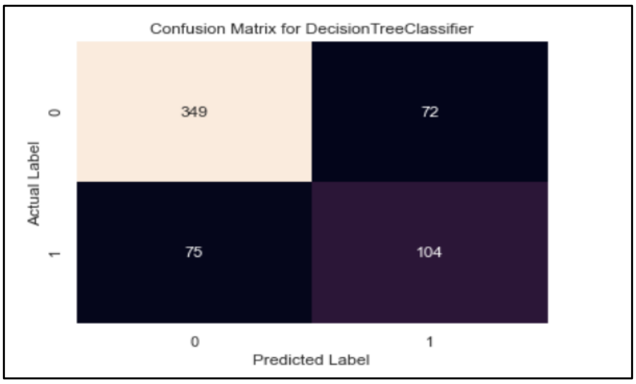
2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model.

Building a Decision Tree Classifier / CART Model:

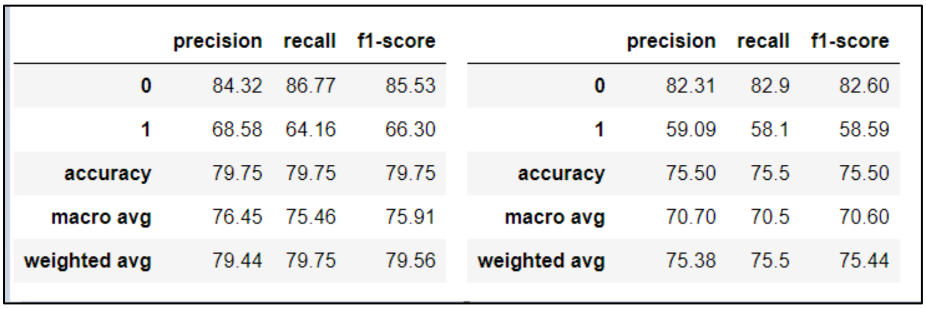
1. Accuracy score for CART model is: ***Train: 79.75% and Test: 75.50%***

2. The Confusion matrix for the CART model: Train and Test





3. The classification report for the CART model:

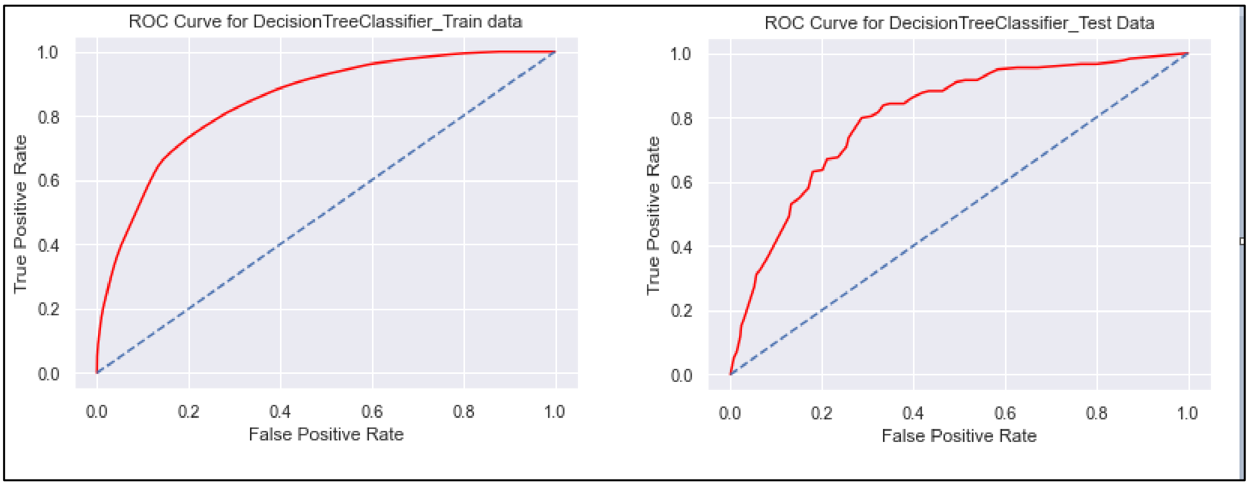


	precision	recall	f1-score
0	84.32	86.77	85.53
1	68.58	64.16	66.30
accuracy	79.75	79.75	79.75
macro avg	76.45	75.46	75.91
weighted avg	79.44	79.75	79.56

	precision	recall	f1-score
0	82.31	82.9	82.60
1	59.09	58.1	58.59
accuracy	75.50	75.5	75.50
macro avg	70.70	70.5	70.60
weighted avg	75.38	75.5	75.44

4. AUC Score: Train: and Test: ***Train: 84.73% and Test: 80.47%***

5. ROC Curve: Train and Test

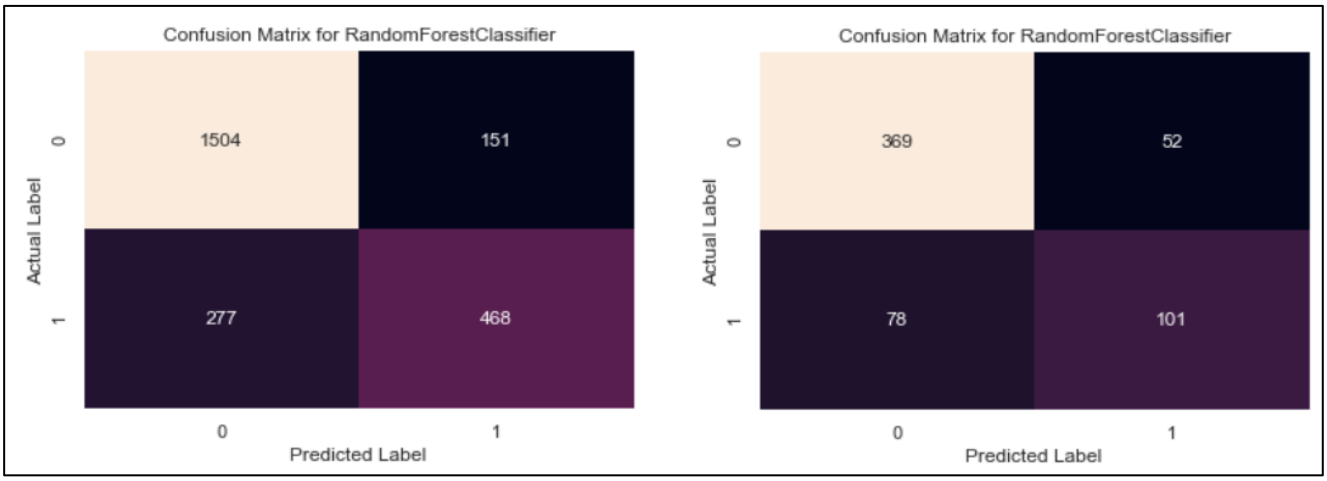


CART Model Conclusion: The performance metric is reduced for Test data. Model is not overfitted. The CART model has performed well on train and test data.

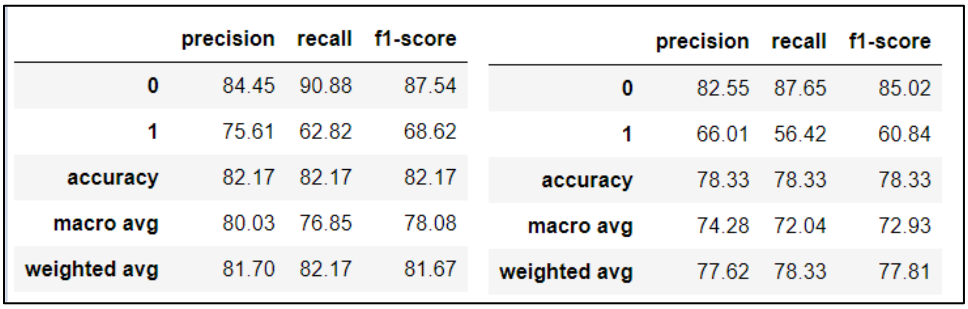
Building a Random Forest Classifier:

1. Accuracy score for RF model is: ***Train: 82.17% and Test: 78.33%***

2. The Confusion matrix for the RF model: Train and Test



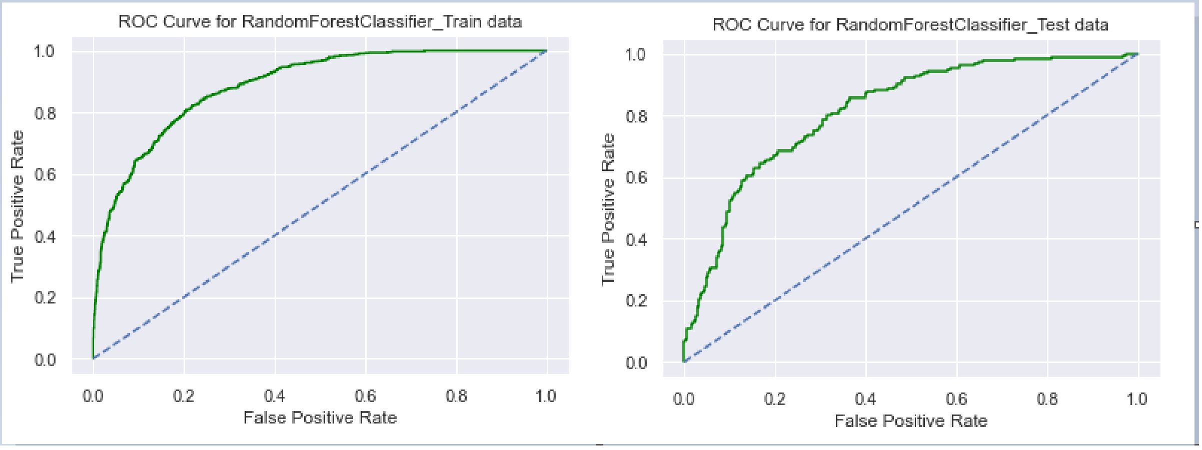
3. The classification report for the RF model:



	precision	recall	f1-score
0	84.45	90.88	87.54
1	75.61	62.82	68.62
accuracy	82.17	82.17	82.17
macro avg	80.03	76.85	78.08
weighted avg	81.70	82.17	81.67

	precision	recall	f1-score
0	82.55	87.65	85.02
1	66.01	56.42	60.84
accuracy	78.33	78.33	78.33
macro avg	74.28	72.04	72.93
weighted avg	77.62	78.33	77.81

4. AUC Score: Train: and Test: ***Train: 88.72% and Test: 81.67%***
5. ROC Curve: Train and Test

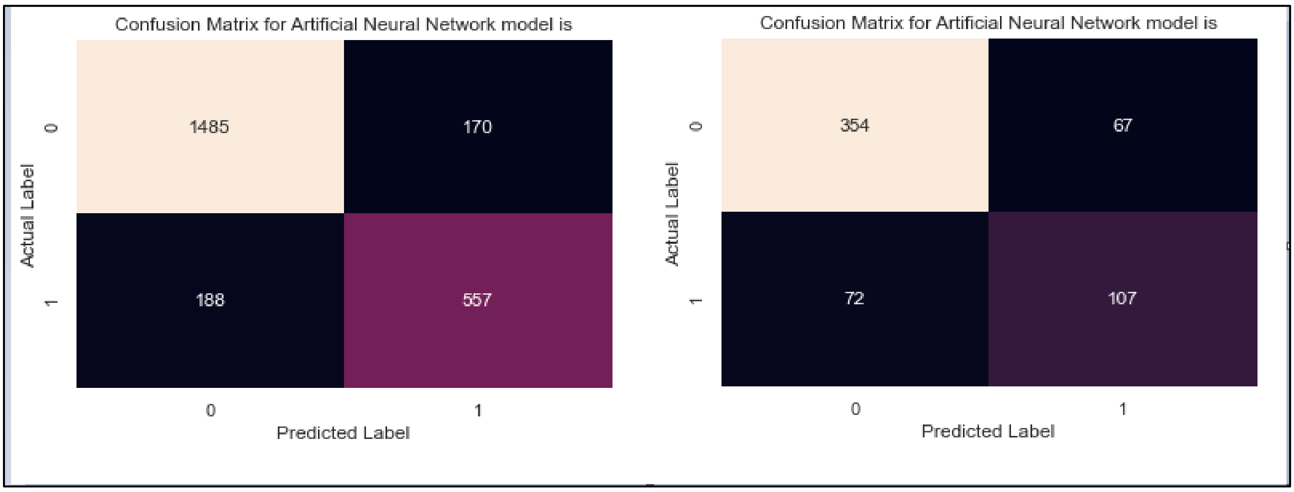


RF Model Conclusion: 2. The difference in accuracy rate between Train and Test is low. The AUC score is good when compared with all other models. RF model has performed well both on Train and test data.

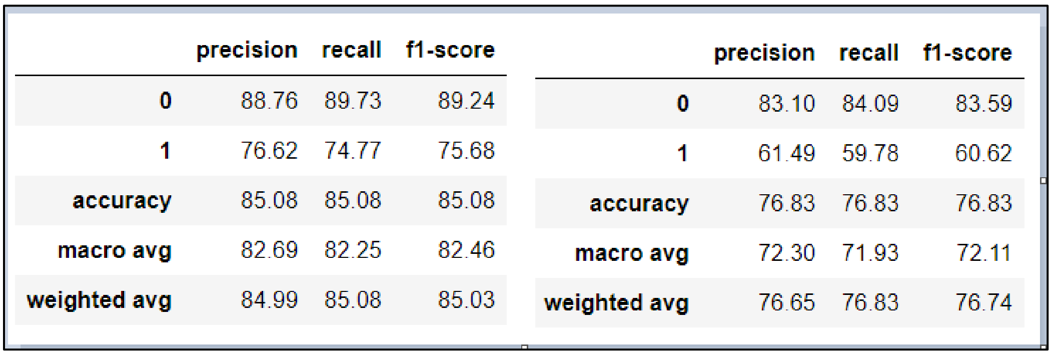
Building an Artificial Neural Network Classifier:

6. Accuracy score for ANN model is: ***Train: 85.08% and Test: 76.83%***

7. The Confusion matrix for the ANN model: Train and Test



8. The classification report for the ANN model:

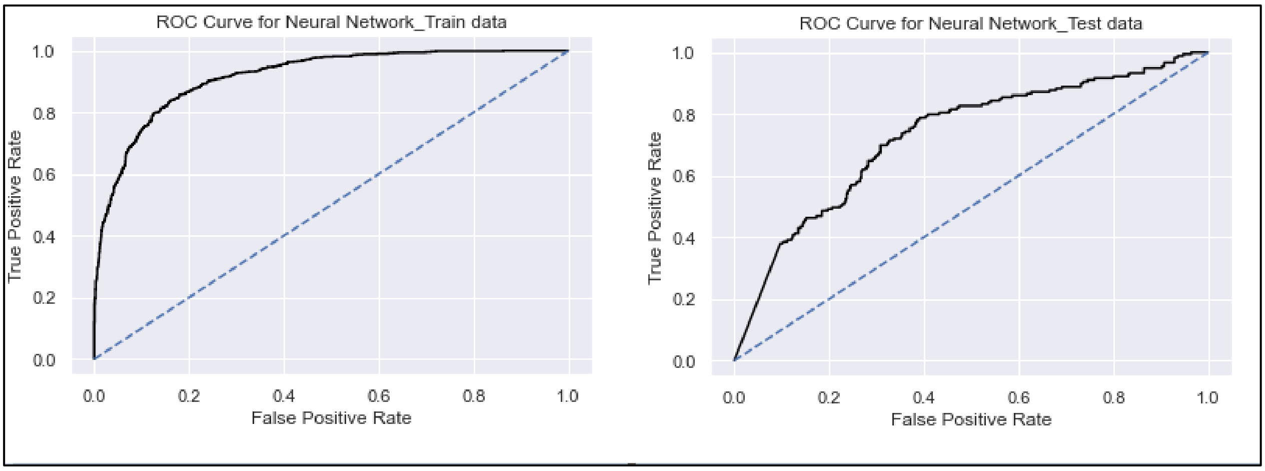


	precision	recall	f1-score
0	88.76	89.73	89.24
1	76.62	74.77	75.68
accuracy	85.08	85.08	85.08
macro avg	82.69	82.25	82.46
weighted avg	84.99	85.08	85.03

	precision	recall	f1-score
0	83.10	84.09	83.59
1	61.49	59.78	60.62
accuracy	76.83	76.83	76.83
macro avg	72.30	71.93	72.11
weighted avg	76.65	76.83	76.74

9. AUC Score: Train: and Test: ***Train: 91.61% and Test: 77.63%***

10. ROC Curve: Train and Test

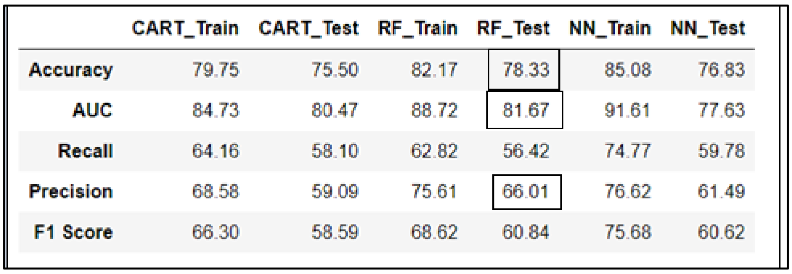


NN Model Conclusion:

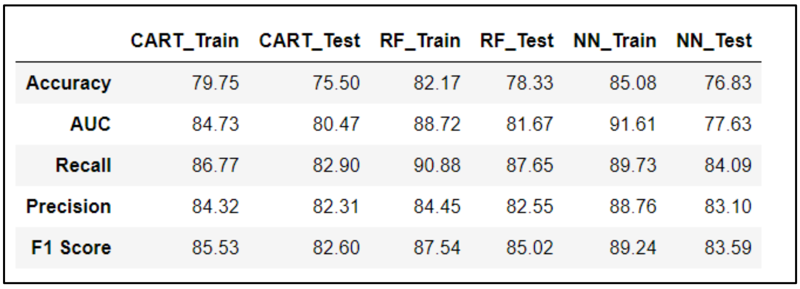
The difference in accuracy rate and AUC score between the train and test is high. NN model is overfitted. It has high difference between the train and test accuracy score.

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

- The given data has a target variable called Claimed and Not claimed. Claimed is 1 and Not claimed is 0.
- It is very important to analyse the number of claims and its predictions and accuracies.
- The below table is the consolidation of all models' Train and test evaluation parameters on Claimed and Not Claimed class.

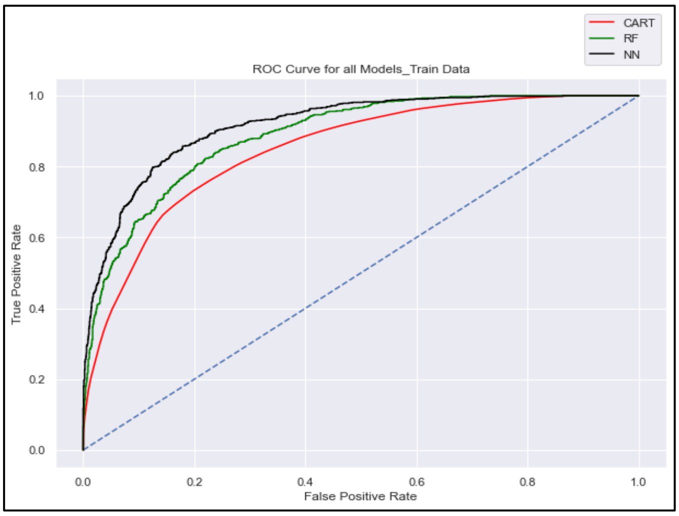


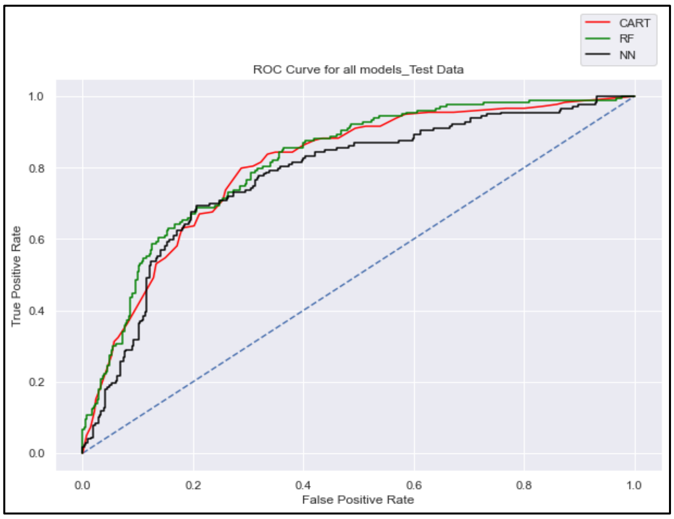
	CART_Train	CART_Test	RF_Train	RF_Test	NN_Train	NN_Test
Accuracy	79.75	75.50	82.17	78.33	85.08	76.83
AUC	84.73	80.47	88.72	81.67	91.61	77.63
Recall	64.16	58.10	62.82	56.42	74.77	59.78
Precision	68.58	59.09	75.61	66.01	76.62	61.49
F1 Score	66.30	58.59	68.62	60.84	75.68	60.62



	CART_Train	CART_Test	RF_Train	RF_Test	NN_Train	NN_Test
Accuracy	79.75	75.50	82.17	78.33	85.08	76.83
AUC	84.73	80.47	88.72	81.67	91.61	77.63
Recall	86.77	82.90	90.88	87.65	89.73	84.09
Precision	84.32	82.31	84.45	82.55	88.76	83.10
F1 Score	85.53	82.60	87.54	85.02	89.24	83.59

The ROC curve for all the models on Train and Test:



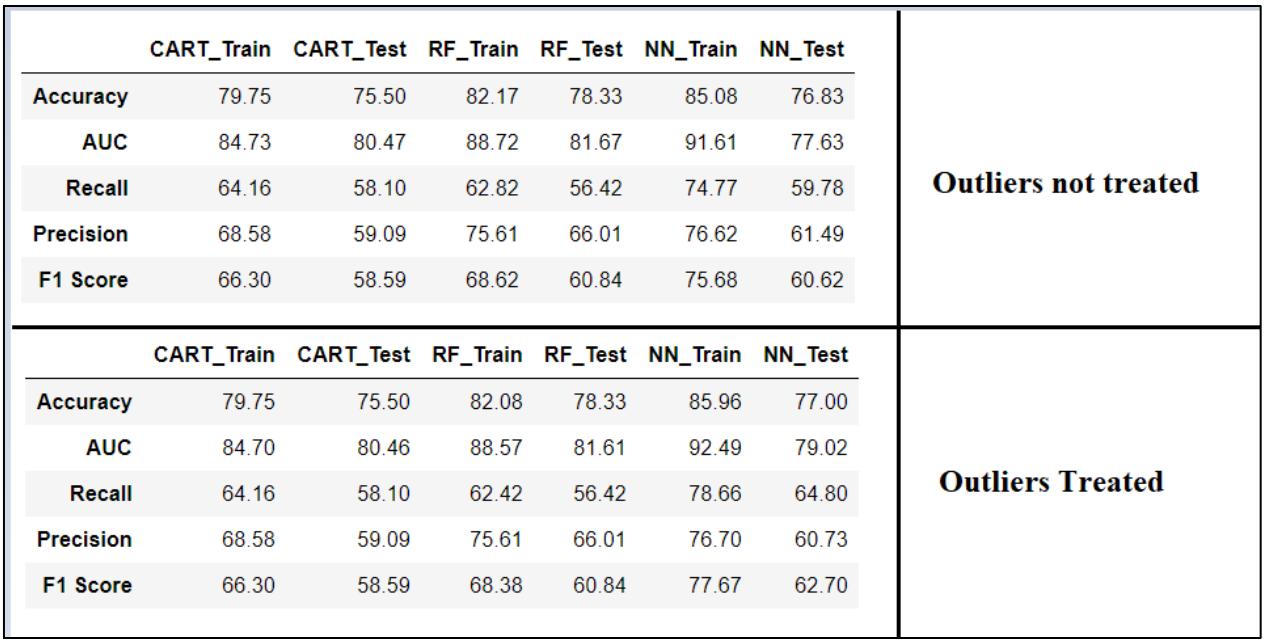


1. CART: The performance metric is reduced for Test data. The train data has good metrics but does not hold good on test data.

2. RF: The difference in accuracy rate between Train and Test is low. The AUC score is good over all other models. RF model is good on train and test data.

3. ANN: The difference in accuracy rate and AUC score between the train and test is high. Though the recall, precision and F1 score are similar to RF model, considering the steep fall between the accuracy and AUC score, ANN model is not preferred.

Model's performance Evaluation metrics on Original data (Outliers not treated) and Outliers treated data:



	CART_Train	CART_Test	RF_Train	RF_Test	NN_Train	NN_Test
Accuracy	79.75	75.50	82.17	78.33	85.08	76.83
AUC	84.73	80.47	88.72	81.67	91.61	77.63
Recall	64.16	58.10	62.82	56.42	74.77	59.78
Precision	68.58	59.09	75.61	66.01	76.62	61.49
F1 Score	66.30	58.59	68.62	60.84	75.68	60.62

Outliers not treated

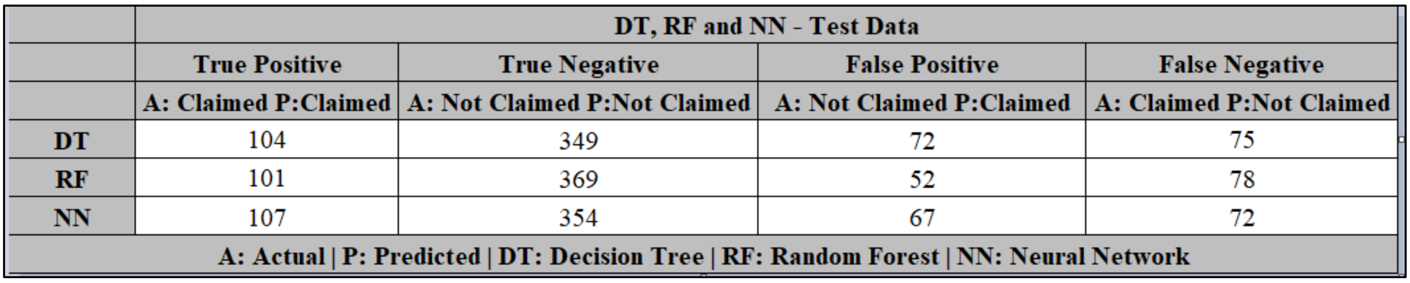
	CART_Train	CART_Test	RF_Train	RF_Test	NN_Train	NN_Test
Accuracy	79.75	75.50	82.08	78.33	85.96	77.00
AUC	84.70	80.46	88.57	81.61	92.49	79.02
Recall	64.16	58.10	62.42	56.42	78.66	64.80
Precision	68.58	59.09	75.61	66.01	76.70	60.73
F1 Score	66.30	58.59	68.38	60.84	77.67	62.70

Outliers Treated

- The scores of CART model and RF shows no significant improvement or changes after treating the outliers. Hence, CART and RF can handle outliers and unscaled data.
- The score of NN model shows significant difference after treating the outliers.
- Though the score for AUC, accuracy and Recall improved for Neural network model after treating outliers, the precision score is having a slight fall. Considering the precision score also as an important metric to decide the reason for high claims, the NN model is not selected as the best model for this case.
- The codebook the outliers treated data set is attached as a separate Jupyter file.

Comparison of Confusion Matrix (all models):

1. The test data (600 observations) have 179 Claimed cases and 421 Not claimed cases.
2. As per the business problem, the identification of correct number of claimed cases, identification of correct number of not claimed cases and identification of falsely claimed cases are very important. Because these counts will increase the claim counts.
3. On analysing the table, RF captures relatively equal True positives, captures high true negative and less False positives.



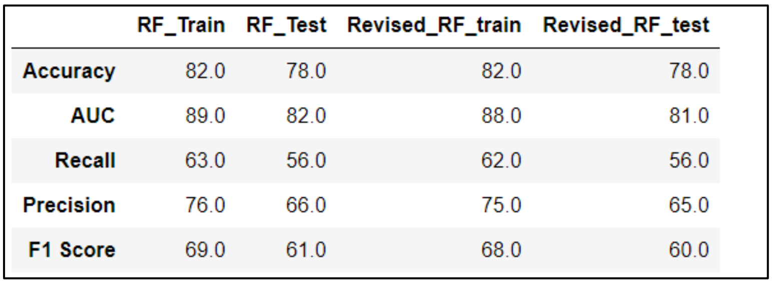
	DT, RF and NN - Test Data			
	True Positive	True Negative	False Positive	False Negative
	A: Claimed P:Claimed	A: Not Claimed P:Not Claimed	A: Not Claimed P:Claimed	A: Claimed P:Not Claimed
DT	104	349	72	75
RF	101	369	52	78
NN	107	354	67	72
A: Actual | P: Predicted | DT: Decision Tree | RF: Random Forest | NN: Neural Network				

Model comparison and Selection of best model:

- The given data is classification type.
- The Target or the dependent variable is Binary in nature.
- The accuracy score, AUC score and the ROC curve is favourable to Random forest.
- The fall in accuracy score and AUC score between the Train and Test is high for NN model.
- Precision: Considering Non-Claimed as Claimed. This will lead to the increase in frequency of the claims without actually claiming.
- Recall: Classifying Claimed observation as Claimed.
- In this case, Precision is equally important to Recall.
- The model, Random forest on both the data (Outliers not treated and Outliers treated) have high Precision and comparatively equal recall values. Hence, Random Forest is used for further analysis.

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

The final best selected model is the Random forest model with 300 decision trees. For further understanding, the least important features are removed and compared with the RF model with all features. Below is the comparison table. The scores of Random forest model and the scores of revised random forest model after removing the least important features like ['Americas','EUROPE','EPX','CWT','Channel'] are not significantly different.

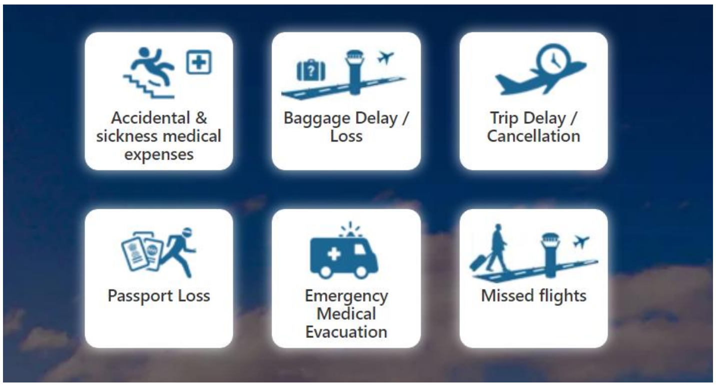


	RF_Train	RF_Test	Revised_RF_train	Revised_RF_test
Accuracy	82.0	78.0	82.0	78.0
AUC	89.0	82.0	88.0	81.0
Recall	63.0	56.0	62.0	56.0
Precision	76.0	66.0	75.0	65.0
F1 Score	69.0	61.0	68.0	60.0

Understanding Travel/Tour Insurance:

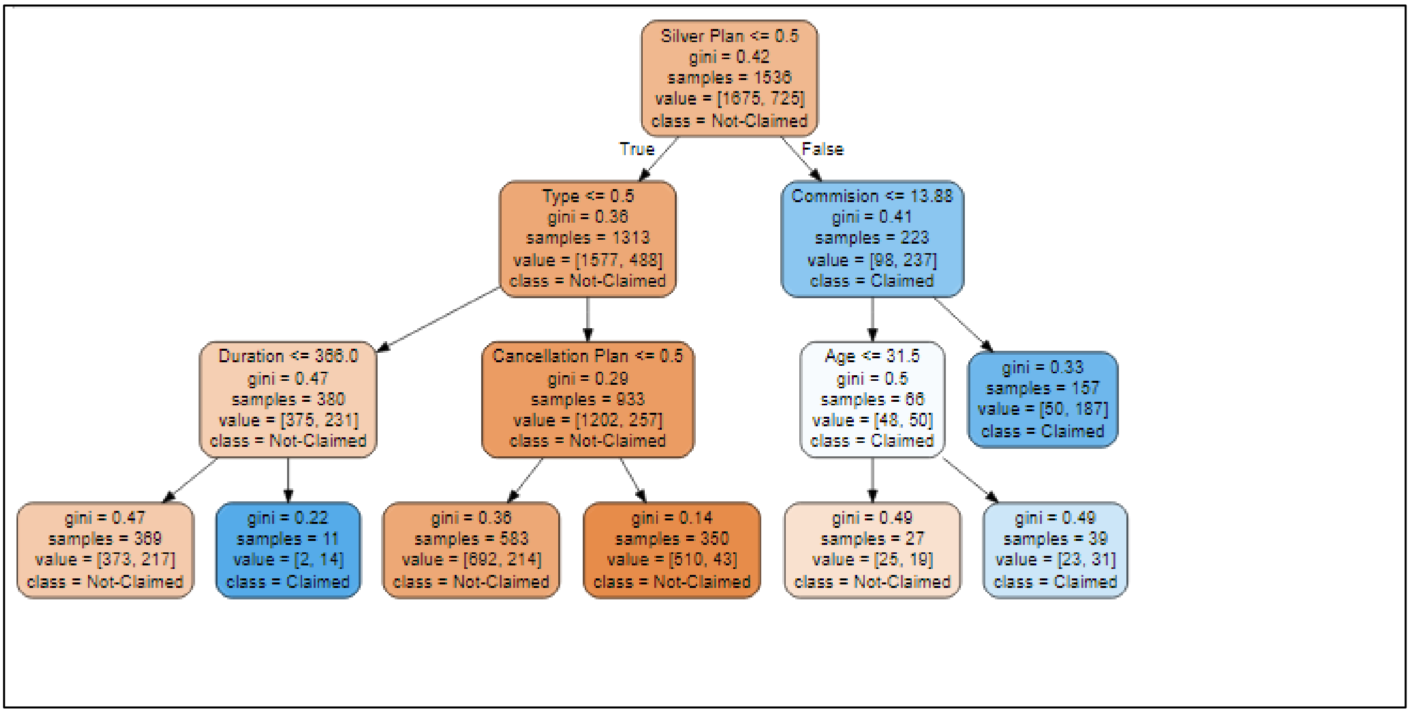
A travel insurance would help you tackle all the travel and medical contingencies while you travel abroad. Buying a travel insurance would safeguard one from all the possible unforeseen situations like flight delay, loss of baggage, loss of passport as well as medical emergencies. The main aspects of the travel insurance policies:

1. Effective online distribution channel.
2. Effective medical coverage
3. Coverage Benefits:

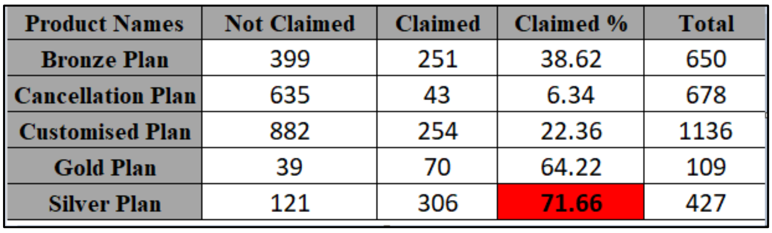

Accidental & sickness medical expenses
Baggage Delay / Loss
Trip Delay / Cancellation
Passport Loss
Emergency Medical Evacuation
Missed flights

4. Claim rejection ratio: Claim rejection is actually the percentage of the claims rejected to the claim settled by the insurance company. Higher the claim rejection ratio, higher are the chances of claim getting rejected. Always opt for an insurance company with a lower claim rejection ratio.

A tree from the random forest model is randomly extracted. The tree extracted is pruned with Maximum depth:3.

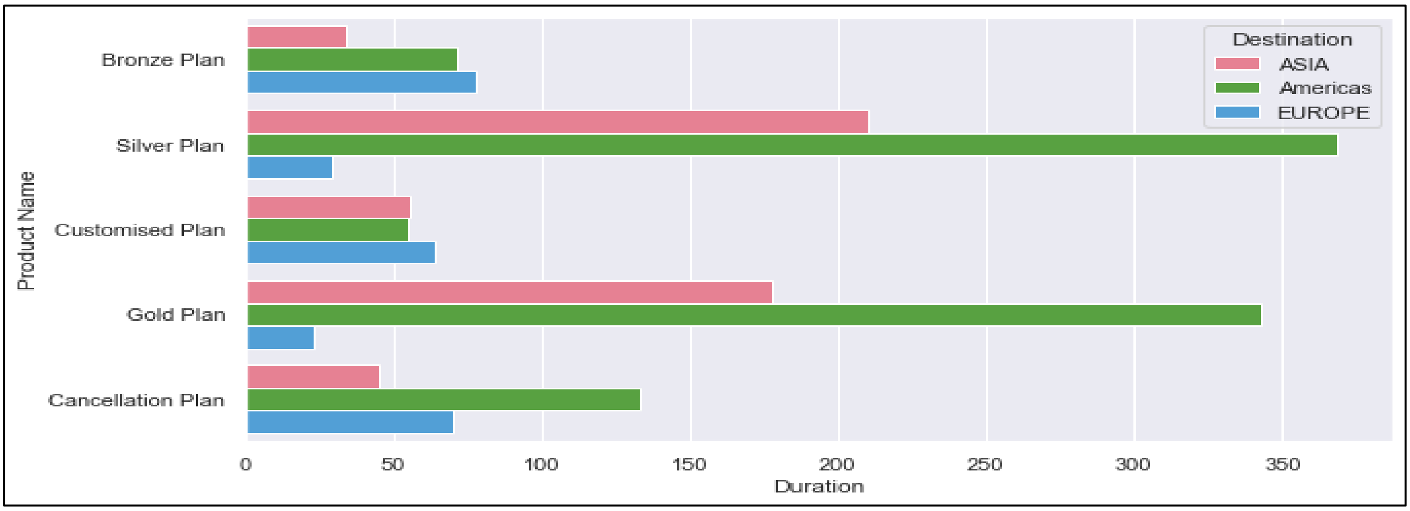


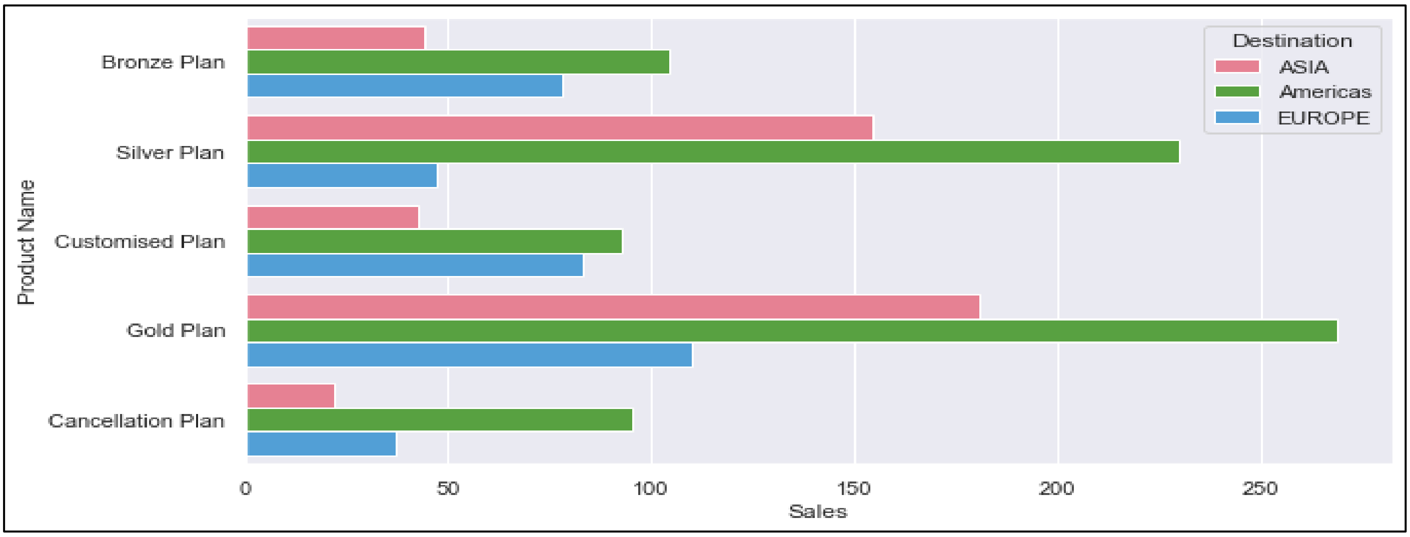
- Silver Plan: It has the lowest gini index hence, chosen as a decision node. The Silver plan under the Product name has highest number of claimed cases. The given business problem is about the higher claim frequency. The model has correctly captured the feature which caused high claims.

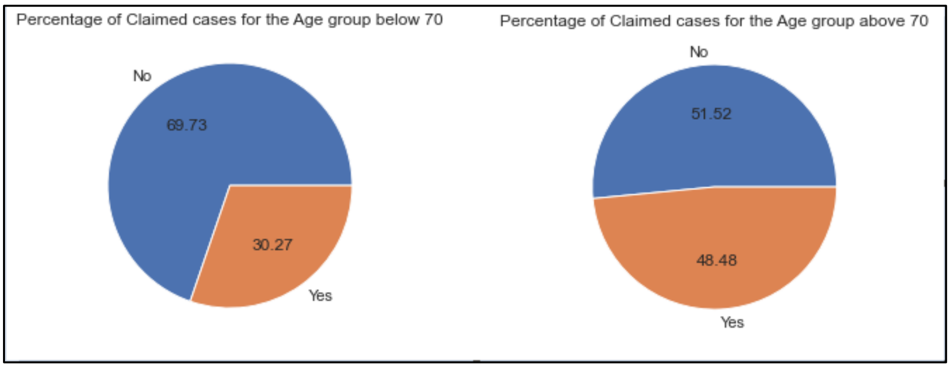


Product Names	Not Claimed	Claimed	Claimed %	Total
Bronze Plan	399	251	38.62	650
Cancellation Plan	635	43	6.34	678
Customised Plan	882	254	22.36	1136
Gold Plan	39	70	64.22	109
Silver Plan	121	306	71.66	427

- Commission: 1677 observations with commission less than 13.88 have not claimed the policy.
- Type: Among the Travel agency and Airlines, the travel tickets booked through airlines have equal number of claimed and non-claimed cases.
- The Customized and Cancellation plans are highly selling plans and the number of claims is less when compared with other plans.
- The Bronze, Customized and Cancellation plans are preferred for less duration travels and their corresponding sales are lesser than the other plans.
- The Gold and Silver are preferred for longer duration travel i.e., for more than 250 days. These plans are highly used by the passengers whose destination is America. As the duration of travel is more, the claims are high as well. The passenger faces more risks and uncertainties when his/her duration of travel is longer.
- Below is the graphical representation of the analysis on only the claimed data.







Conclusion:

- The above graphical representation shows the Sales, Duration and Product plan preference among the people who have claimed. Out of 3000 observations, the claimed counts 924, is used for the analysis.
- People who travel using the Bronze, Cancellation plan and Customised plan have less number claims. The mentioned plans majorly are used for travel with shorter durations.

- People who travel using Gold and silver plan have significant number of claims. Longer the distance higher is the risks and uncertainties. Hence, more claims from these observations are recorded.
- Though the claims are higher for Silver plan and Gold plan, these plans are highly selling plan policy and yielding an average commission of 38 and 47 respectively.
- Almost 2950 people preferred online channel. 30% of travellers have claimed their policies. Out of the 30% claimed, 40% travellers have booked through Silver and Gold plans.
- The pie plot representation shows the Claimed and non-claimed percentage between below 70 and above 70 age group. Generally, if the travellers age is more than 60/70 the Medical risk is higher. Travel insurance companies should have appropriate medical benefits and medical expenses claims. In the given data, almost 50% of people who are aged above 70 have claimed the insurance. Whereas, only 30% of people who are aged below 70 have claimed the insurance. One of the reasons for high claims.
- Travellers tend to prefer the insurance providers with Lower claim rejection ratio. Hence, the travel insurance providers should create robust packages which covers maximum benefits.

Libraries Imported for the project:

Common Libraries:

```
import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

sns.set()
```

Problem Statement 1:

```
from scipy.cluster.hierarchy import dendrogram, linkage

from scipy.cluster.hierarchy import fcluster

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

from scipy.stats import zscore

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette_samples, silhouette_score
```

Problem Statement 2:

```
from sklearn import tree

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.neural_network import MLPClassifier

from sklearn.model_selection import train_test_split

from sklearn.metrics import roc_auc_score,roc_curve,classification_report,confusion_matrix,accuracy_score

from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import GridSearchCV

from sklearn.tree import export_graphviz
```