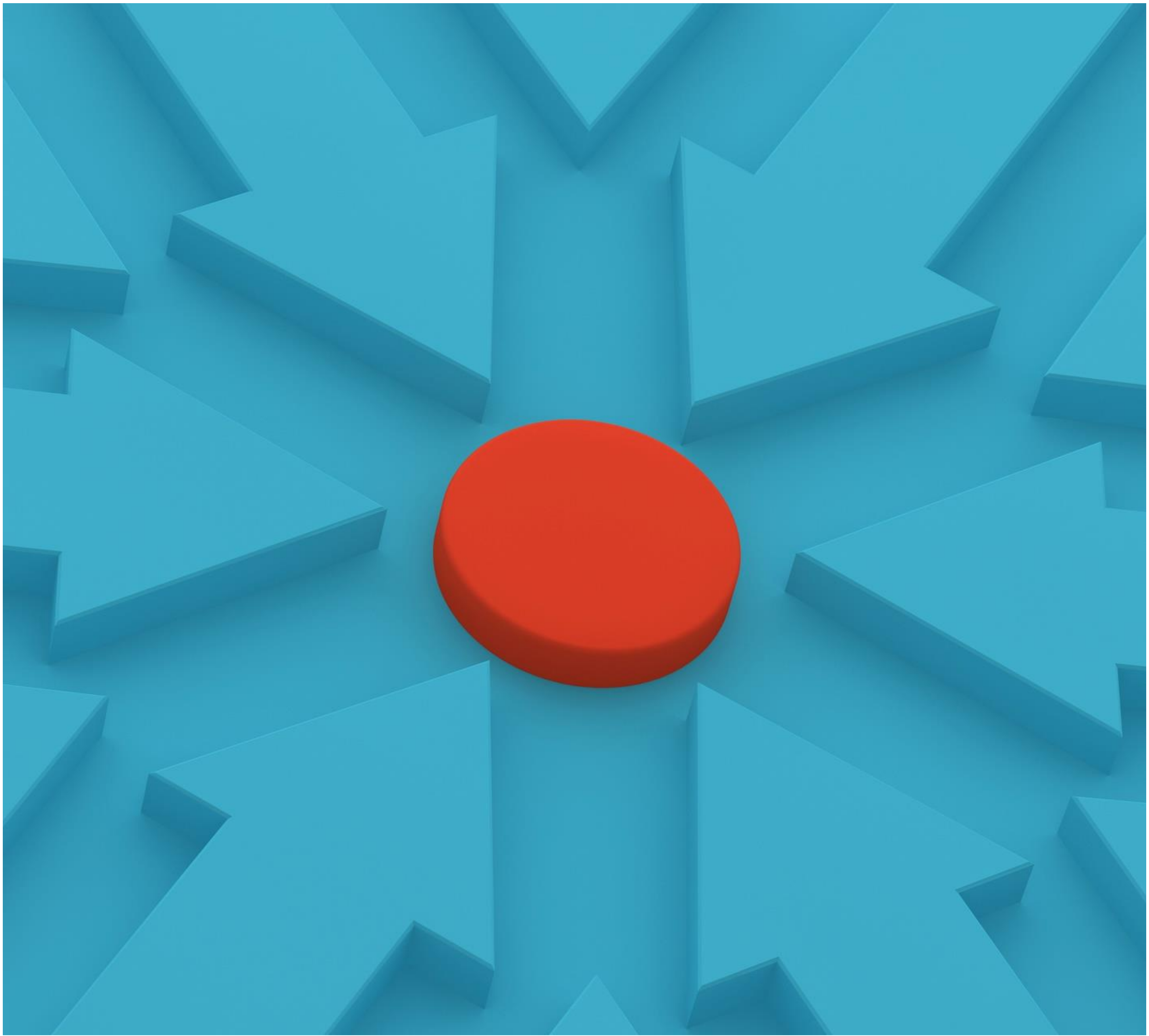


DATA MINING

PROJECT REPORT

SUNDAR RAM S
PGPDSBA
ONLINE DEC_C 2021
08-MAY-2022

Table of Contents

TABLE OF TABLES

TABLE OF FIGURES

TABLE OF EQUATIONS

EXECUTIVE SUMMARY

This report is based on two data sets that contain credit card usage of customers and tour insurance claims. Credit card usage of customers is analyzed by clustering them and suitable business recommendations for promotional purposes are made. The techniques of hierarchical clustering using dendrograms and non-hierarchical method of KMeans clustering are employed.

The insurance claims data of past few years are used to make a model that predicts the claim status. Depending on the claim status suitable recommendations are made to the management. The techniques of CART (Classification And Regression Tree), Random Forest & ANN (Artificial Neural Networks) are employed in the model building.

INTRODUCTION

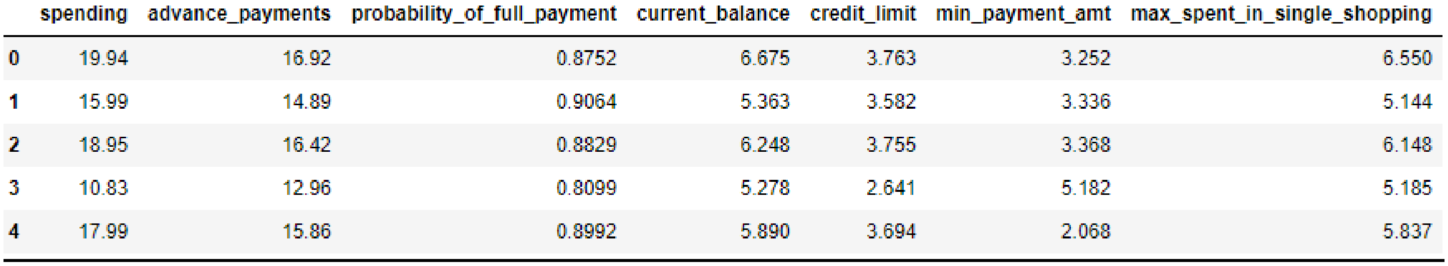
This report provides a detailed explanation on approach used, record inferences, insights and provide suitable business solutions/ recommendations. The techniques of Data Mining, both Supervised and unsupervised learning techniques, that employs Clustering (Hierarchical – Agglomerative & Non-hierarchical – K-means), Predictive Modelling techniques like CART, Random forest & Artificial Neural Networks are leveraged in this exercise.

PROBLEM 1 – CLUSTERING – CUSTOMER SEGMENTATION

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.

Sample of the Dataset



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

Table 1 Sample of the Dataset

The data consists of 7 columns having the details of customer activities during the past few months of 210 customers.

Dataset Description

1 spending : Amount spent by the customer per month (in 1000s)

2 advance_payments : Amount paid by the customer in advance by cash (in 100s)

3 probability_of_full_payment : Probability of payment done in full by the customer to the bank

4 current_balance : Balance amount left in the account to make purchases (in 1000s)

5 credit_limit : Limit of the amount in credit card (10000s)

6 min_payment_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)

7 max_spent_in_single_shopping : Maximum amount spent in one purchase (in 1000s)

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

EXPLORATORY DATA ANALYSIS

Variable type

spending : float64

advance_payments : float64

probability_of_full_payment : float64

current_balance : float64

credit_limit	: float64
min_payment_amt	: float64
max_spent_in_single_shopping	: float64

Variable type is the type of data each column is holding. Here we are able to notice that all the columns are of the type float64 and hence no bad data.

Check for missing values in the dataset

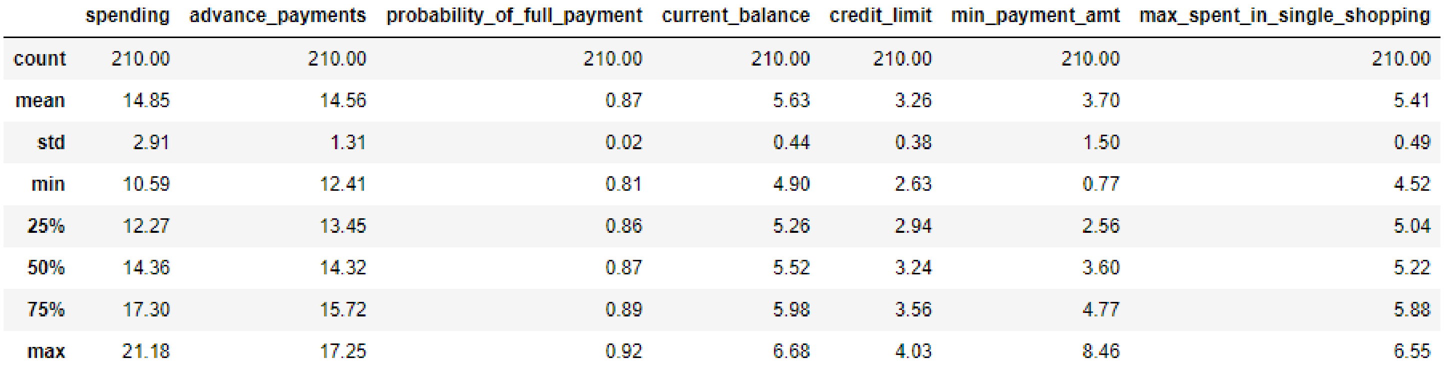
spending	: 210 non-null
advance_payments	: 210 non-null
probability_of_full_payment	: 210 non-null
current_balance	: 210 non-null
credit_limit	: 210 non-null
min_payment_amt	: 210 non-null
max_spent_in_single_shopping	: 210 non-null

The dataset consists of 210 customers data and it is evident that there are no missing values in the entire dataset.

Summary of the dataset

Summarizing briefly, the dataset has a total of 210 records. Mean spending is 14.85 k/ month, mean advance payment in cash is 1.456 k, mean probability that customer pays the full amount is 0.87 i.e. 87 of the customers are expected to pay the full amount, mean current balance is 5.63 k, mean credit limit is 32.6 k, mean minimum amount paid by the customer while making payments for purchases made monthly is 0.37 k and mean maximum amount spent in one purchase is 5.41 k.

From the below table, we can see that there are no anomalies/ bad data and we can also interpret that the variation between mean & median is low. Hence there might by very few/ no outliers in the data set. However, the presence of outliers can be determined by the box plot curve.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.00	210.00	210.00	210.00	210.00	210.00	210.00
mean	14.85	14.56	0.87	5.63	3.26	3.70	5.41
std	2.91	1.31	0.02	0.44	0.38	1.50	0.49
min	10.59	12.41	0.81	4.90	2.63	0.77	4.52
25%	12.27	13.45	0.86	5.26	2.94	2.56	5.04
50%	14.36	14.32	0.87	5.52	3.24	3.60	5.22
75%	17.30	15.72	0.89	5.98	3.56	4.77	5.88
max	21.18	17.25	0.92	6.68	4.03	8.46	6.55

Table 2 Statistical summary of the dataset

DATA VISUALIZATION

Histogram Plot, Skewness & Kurtosis

From the below histogram plots, we are able to understand the distribution of data in the dataset. There are double peaks for spending, advance_payments, & max_spent_in_single_shopping. Probability of full payment is left skewed while current balance is right skewed. Credit_limit has a flat top with more or less normal distribution, while min payment amount is more or less normally distributed with a slight right skew.

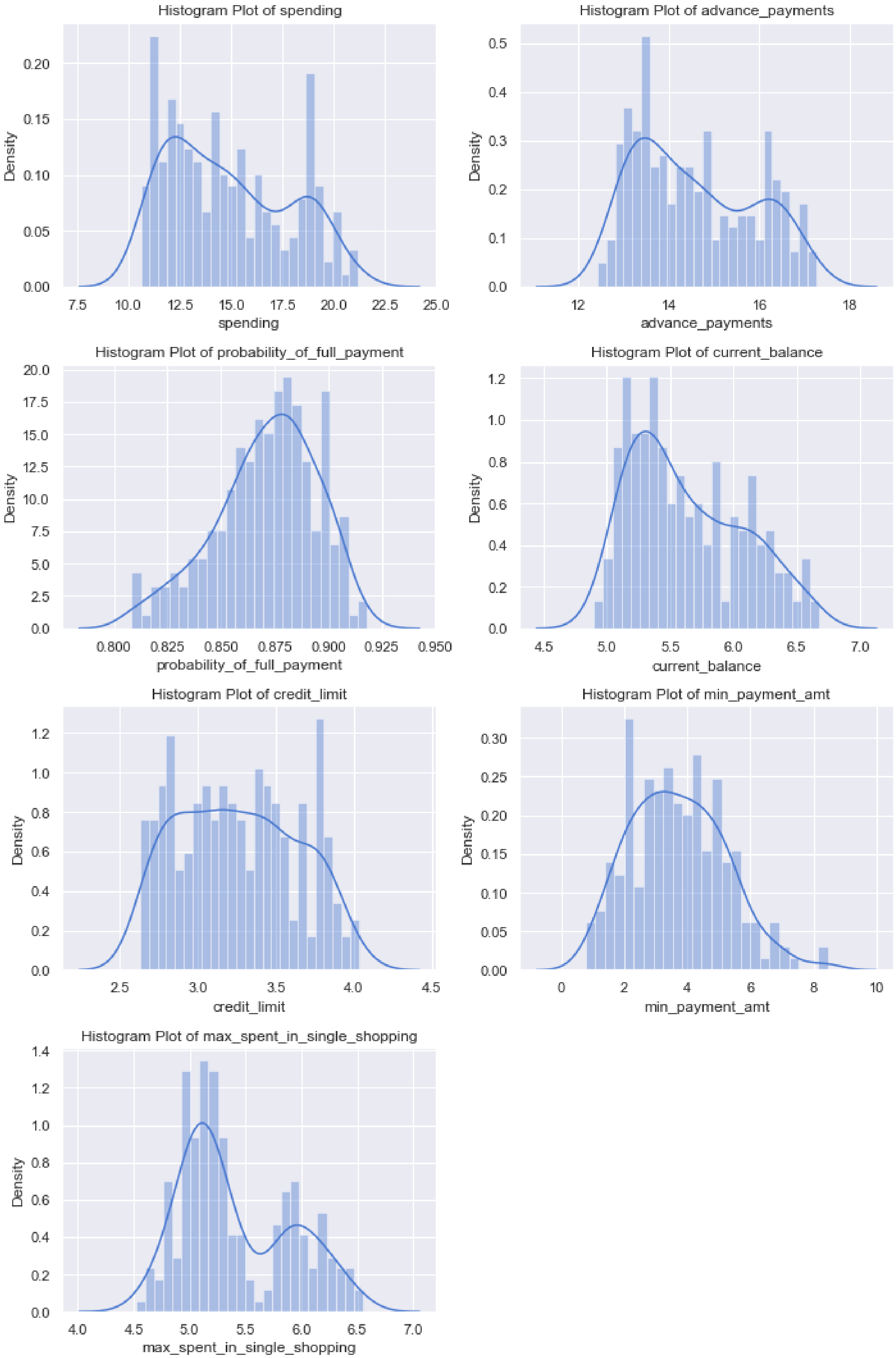
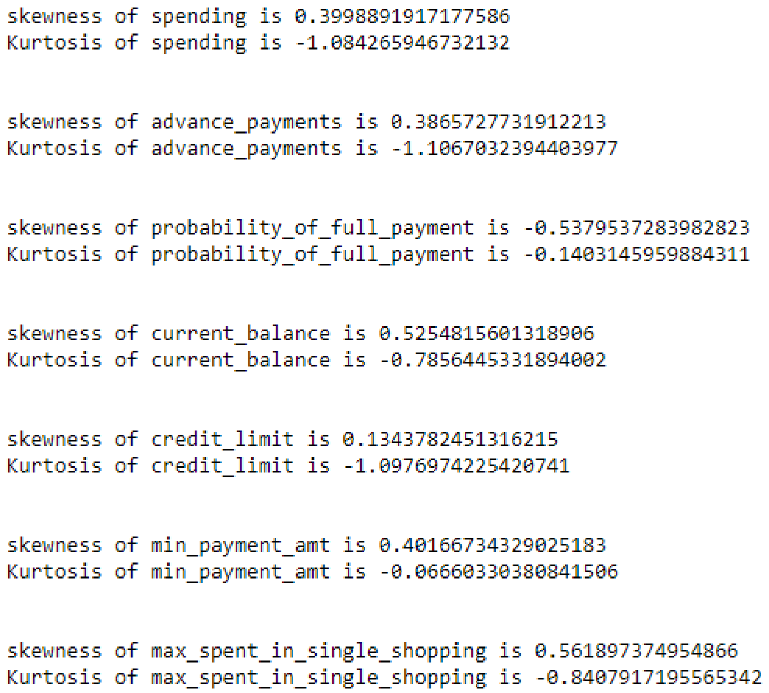


Figure 1 Histogram Plot

In addition to visualizing through histogram plots, we can also understand the distribution by calculating the skewness and kurtosis. Calculated values are as in the below figure.



```
skewness of spending is 0.3998891917177586
Kurtosis of spending is -1.084265946732132


skewness of advance_payments is 0.3865727731912213
Kurtosis of advance_payments is -1.1067032394403977


skewness of probability_of_full_payment is -0.5379537283982823
Kurtosis of probability_of_full_payment is -0.1403145959884311


skewness of current_balance is 0.5254815601318906
Kurtosis of current_balance is -0.7856445331894002


skewness of credit_limit is 0.1343782451316215
Kurtosis of credit_limit is -1.0976974225420741


skewness of min_payment_amt is 0.40166734329025183
Kurtosis of min_payment_amt is -0.06660330380841506


skewness of max_spent_in_single_shopping is 0.561897374954866
Kurtosis of max_spent_in_single_shopping is -0.8407917195565342
```

Figure 2 Calculated values of skewness & kurtosis

Skewness assesses the extent to which a variable's distribution is symmetrical. Kurtosis is a measure of whether the distribution is too peaked. For an ideal normal distribution (theoretical) Skewness and Kurtosis have to be between -1 to +1.

Thus, we can correlate the inference from histogram plot & the calculated values of skewness & kurtosis. For spending, we are able to infer from the calculated values of skewness (0.39), the data is slightly skewed towards the right. However, the value of kurtosis (-1.08) implies the distribution is less peaked/ flat at the top. A similar observation can be seen for advance_payments with a skewness value of 0.38 & kurtosis of -1.10. From Histogram & calculated values of skewness (-0.53) & kurtosis(-0.14) for probability_of_full_payment, we can understand that the distribution is left skewed (-ve value of skewness) and the lesser negative value of kurtosis implies the curve top is more or less peaked. The plot and calculated values (skew= 0.52, kurtosis= -0.78), current_balance implies that the distribution is right skewed and the curve has a top that is not too peaked. However, from the histogram plot, we can understand that, there are two peaks in the plot that impacts the kurtosis value. From Histogram & calculated values of skewness (0.13) & kurtosis (-1.09) for credit_limit, we can understand that the distribution is more or less normally distributed with slight inclination towards right skew and the negative value of kurtosis implies the curve top is flat. From Histogram & calculated values of skewness (0.40) & kurtosis (-0.06) for min_payment_amt, we can understand that the distribution is more or less normally distributed with slight inclination towards right skew and there is only a very mild negative value of kurtosis that implies, the peak is proper in the curve. The plot and calculated values (skew= 0.56, kurtosis= -0.84) of max_payment_in_single_shopping implies that the distribution is right skewed and the curve has a top that is not too peaked. However, from the histogram plot, we can understand that, there are two peaks in the plot that impacts the kurtosis value.

Box Plot – Outliers Detection

Outliers are values that are abnormally away from the other values in the dataset. They tend to affect the arithmetic mean of the dataset, abnormally skewing the value to one side (upper or lower), depending on the presence of the outlier. The values below minimum **[Q1 - 1.5(Inter Quartile Range)]** are the outliers on the lower side of the dataset, while values above maximum **[Q3 + 1.5(Inter Quartile Range)]** are the outliers on the upper side of the dataset, where Q1 & Q3 are the 25th and 75th percentile respectively.

Boxplot is an excellent plot that gives us the 5 number summary (minimum, Q1, median, Q3, maximum). Q1, median & Q3 are represented by the box and the whiskers denote the values of maximum and minimum on either side of the box.

The outliers are denoted by the points that fall either after the maximum whisker or below the minimum whisker. The figure 3 shows the box plot for all the columns. From the box plot it is evident that there are outliers in two of the column's probability_of_full_payment (lower side) & min_payment_amt (higher side). The remaining columns don't have any outliers.

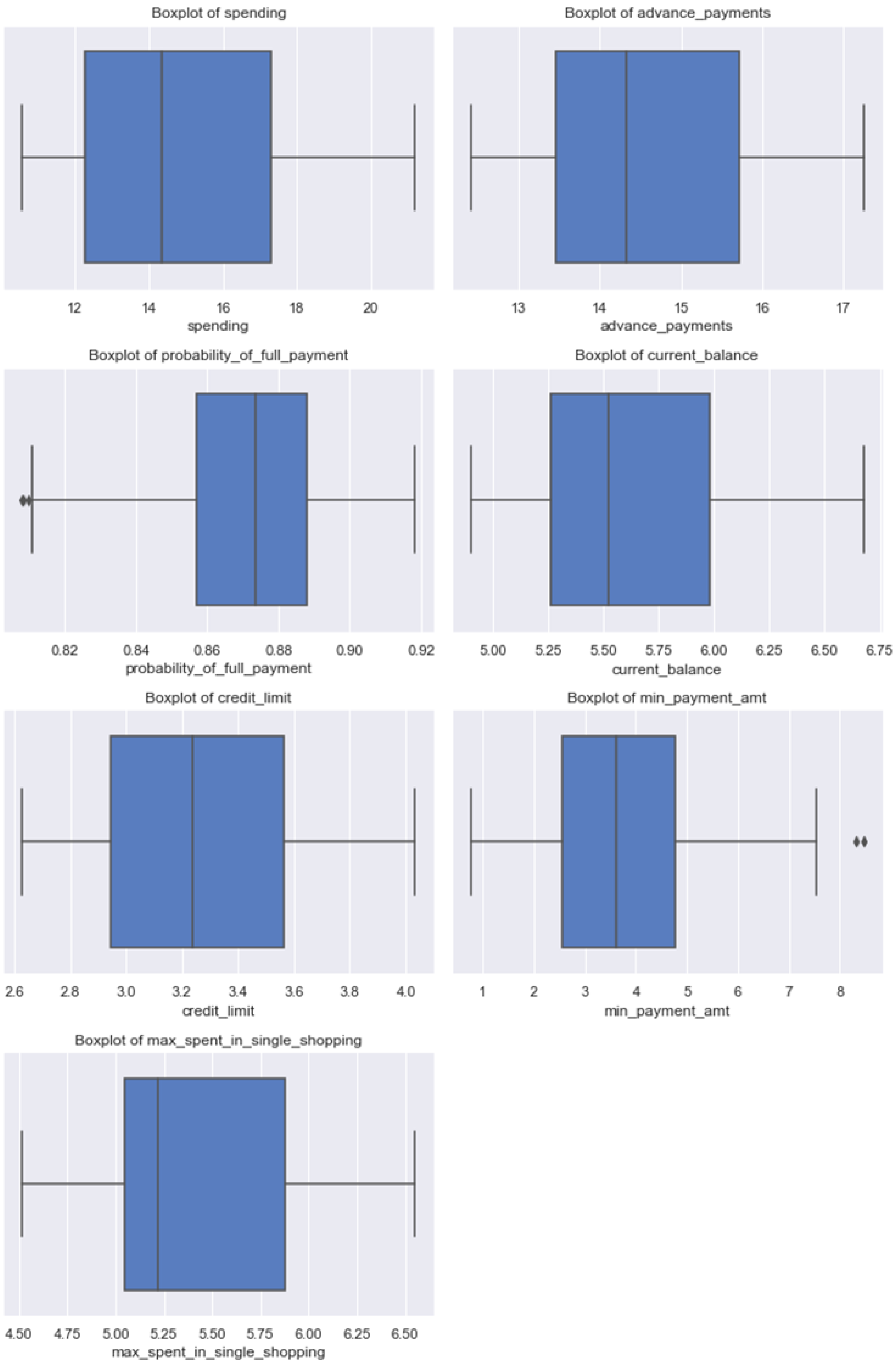


Figure 3 Box Plot

Outlier Treatment

Clustering is affected by the presence of outliers. Hence, it becomes essential to either impute the outliers with the highest/ lowest whisker value, whichever is apt or remove the records from the dataset. Considering imputation of whisker values, the box plot of the updated columns as in the below figure.

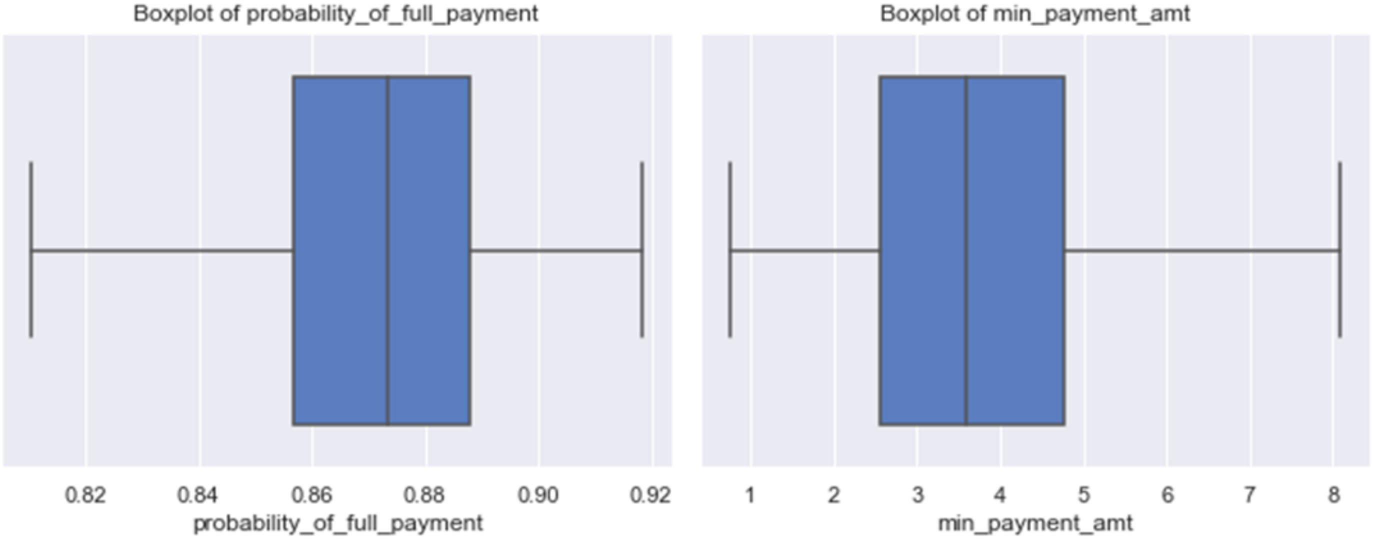


Figure 4 Box Plot - Post Outlier treatment

Correlation Plot – Relation amongst columns

The below figure 4 shows the correlation plot (heat map) of the columns present in the dataset. The figure indicates that all the columns except min_payment_amt has a positive correlation amongst them. Spending has the highest positive correlation (0.99) with advance_payments and the negative correlation of -0.23 with min_payment_amt. Similarly advance_payment has the highest positive correlation (0.99) with spending while has a negative correlation of -0.22 with min_payment_amt. probability_of_full_payment has a moderate positive correlation with all the variables except min_payment_amt. It has the highest positive correlation (0.76) with credit_limit and has a negative correlation of -0.34 with min_payment_amt. current_balance has the highest positive correlation (0.97) with advance_payments and has a negative correlation of -0.17 with min_payment_amt. credit_limit has the highest positive correlation (0.97) with spending and a negative correlation of -0.26 with min_payment_amt. The min_payment_amt has a negative correlation with all the other columns. The highest negative correlation is with probability_of_full_payment (-

0.34) and the least negative correlation with max_spent_in_single_shopping (-0.01). max_spent_in_single_shopping has the highest positive correlation with current_balance and a lower positive correlation with probability_of_full_payment (0.23). It has a negative correlation value of -0.01 with min_payment_amt.

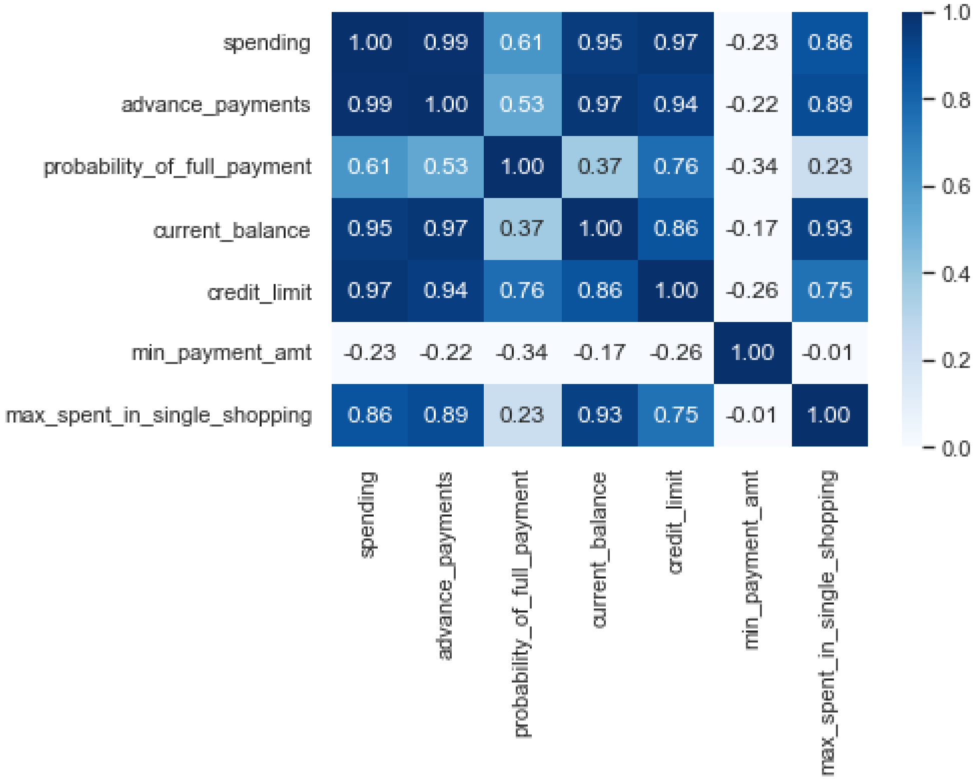


Figure 5 Correlation plot (heat map)

Pair Plot

From the above pair plot it can be seen that spending, advance_payment, current_balance and credit_limit is directly proportional to each other.

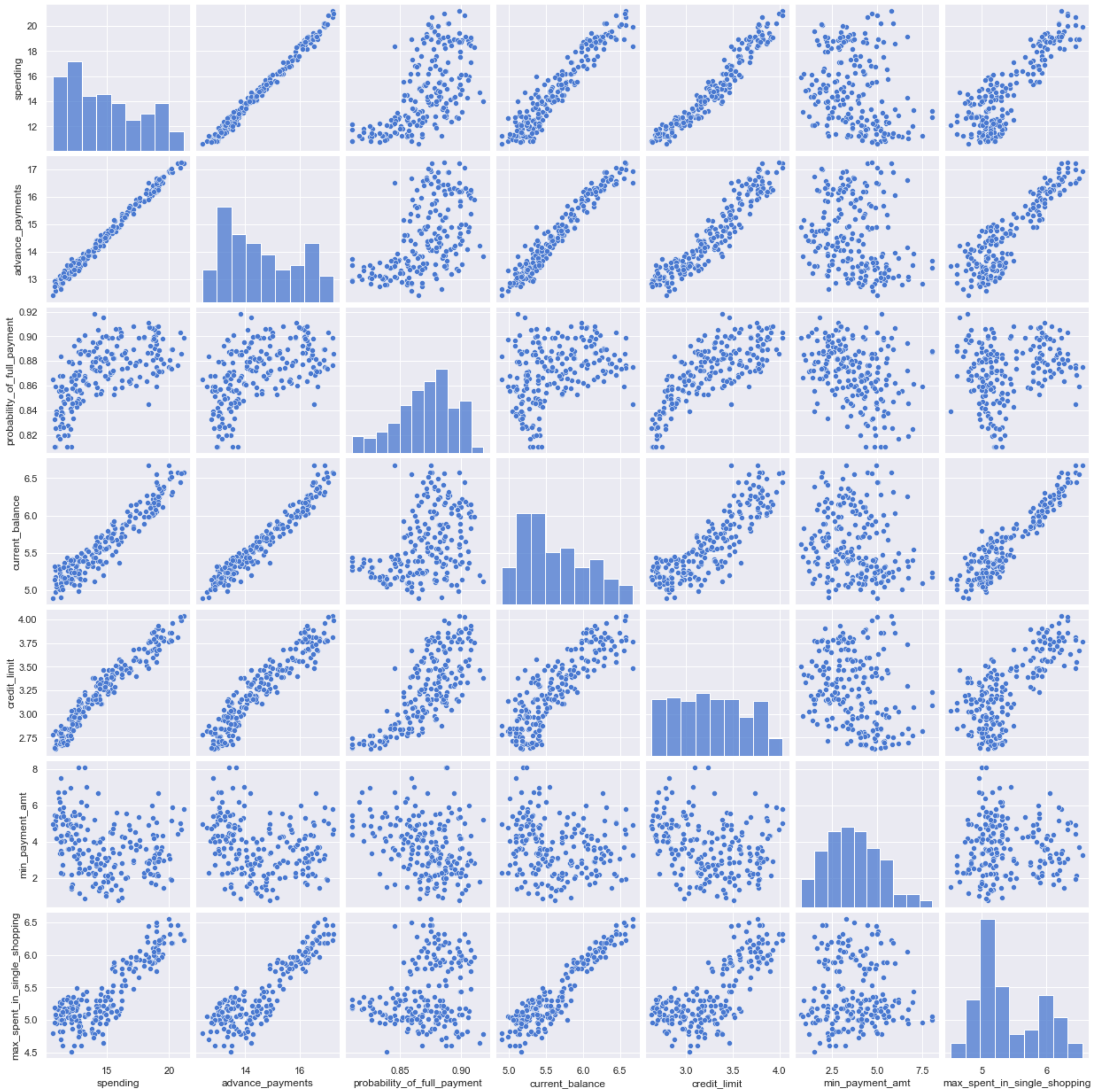
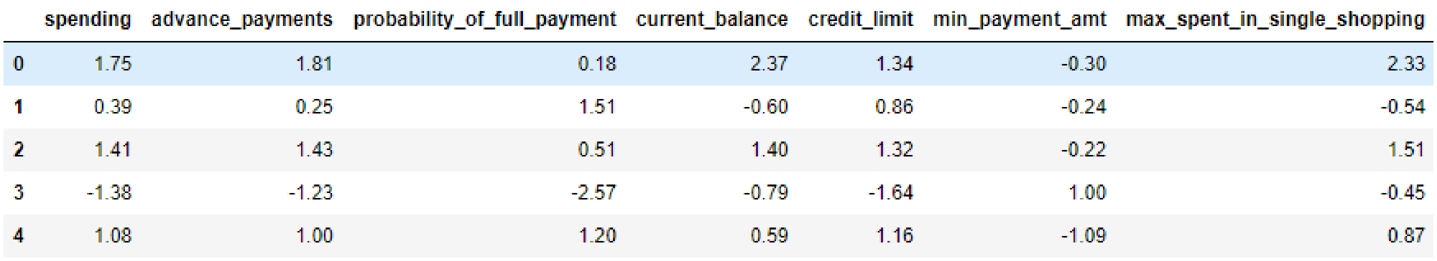


Figure 6 Pair Plot

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

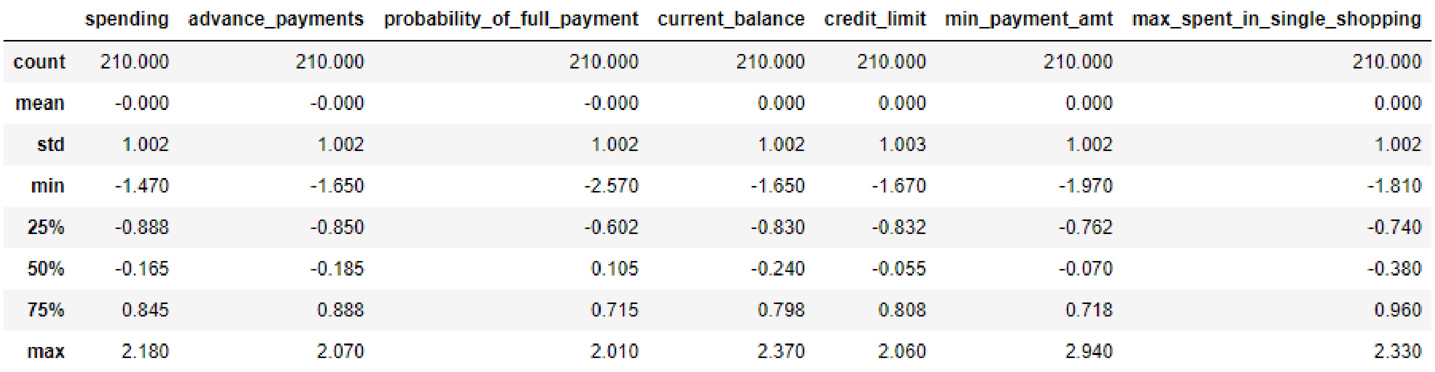
Yes, Scaling is necessary for clustering in this case. Clustering is a technique that works based on distance calculation and is sensitive to magnitude of the values in the dataset. Hence, it becomes absolutely essential to scale the dataset to a common magnitude, or else, the clustering algorithm would give biased importance to one column that has the higher magnitude. In other words, if variables in the data set have large differences in their variances, then all variables need to be scaled. Otherwise, the variables with large variance will have disproportionately more influence on. In this case, spending & advance_payments have values in double digit (higher values greater than 10), while current balance, credit limit, min payment amt, max spent in single shopping have values less than 7. The probability of full payment, as the name suggests is less than 1. Similarly, ethe variance/ standard deviation is higher in spending, advance payments and min payment amt (greater than 1.3), while the other columns have the values less than 1. Hence, the columns of the dataset are on different scale, and need to be scaled to a common magnitude before subjecting them to clustering. The data on values, variance/ std can be viewed in tables 1 & 2.

The different ways in which we can scale the data set are Z score, Standard Scaler and min max scaling. Z score and Standard scaler scales the values such that mean of the data tends to 0 and standard deviation tends to 1. The min max scaling will scale data such that the values will lie between 0 and 1. In this case, we will perform scaling through the standard scaler technique. Scaled dataset and summary as in the below tables. We can see that the variance/ standard deviation is more or less the same across columns.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.75	1.81	0.18	2.37	1.34	-0.30	2.33
1	0.39	0.25	1.51	-0.60	0.86	-0.24	-0.54
2	1.41	1.43	0.51	1.40	1.32	-0.22	1.51
3	-1.38	-1.23	-2.57	-0.79	-1.64	1.00	-0.45
4	1.08	1.00	1.20	0.59	1.16	-1.09	0.87

Table 3 Sample of Scaled Dataset



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000	210.000	210.000	210.000	210.000	210.000	210.000
mean	-0.000	-0.000	-0.000	0.000	0.000	0.000	0.000
std	1.002	1.002	1.002	1.002	1.003	1.002	1.002
min	-1.470	-1.650	-2.570	-1.650	-1.670	-1.970	-1.810
25%	-0.888	-0.850	-0.602	-0.830	-0.832	-0.762	-0.740
50%	-0.165	-0.185	0.105	-0.240	-0.055	-0.070	-0.380
75%	0.845	0.888	0.715	0.798	0.808	0.718	0.960
max	2.180	2.070	2.010	2.370	2.060	2.940	2.330

Table 4 Summary of the Scaled Dataset

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

Hierarchical clustering method is based on hierarchy representation of clusters where parent cluster node is connected to further to child cluster node. It is further divided into two types:

- Agglomerative Clustering
- Divisive Clustering

Agglomerative Clustering

The methods start by considering each data point as a single cluster. In the next step the singleton clusters are merged into a big cluster based on the similarity between them. The procedure is repeated until all the datapoints are merged into one big cluster.

Divisive Clustering

The divisive clustering works completely opposite to agglomerative clustering. The method starts from one big cluster considering all data points within it. In the next the big cluster is divided into the most heterogeneous two clusters. The procedure is repeated until each data point is in its own cluster

Out of these 2 types, agglomerative clustering is the most popular and commonly used clustering methods.

Linkage

The similarity between two data points/ records in agglomerative clustering is mathematically calculated as the distance between the two data points. The distance is calculated in the following ways:

- Eucledian
- Manhattan
- Chebyshev
- Minkowski

If (x1, y1) and (x2, y2) are two data points their distance in each of the above method is calculated as following

Euclidean Distance = $\sqrt{}((x2-x1)^2+(y2-y1)^2$………………………………………………………………………… *Equation 1 Euclidean Dist.*

Manhattan Distance = $|x2-x1|+|y2-y1|$……………………………………………………………………… Equation 2 Manhattan Dist.

Chebyshev Distance = $\max(|x2-x1|,|y2-y1|)$………………………………………………………………… Equation 3 Chebyshev Dist.

Minkowski Distance = $\sum|((x1-y1),(x2-y2))^{p}|$………………………………………………………………….. Equation 4 Minkowski Dist.

The most commonly used distance measuring method is the Euclidean Distance. For measuring distance between clusters, the linkage comes in to play. Linkage process merges two clusters into one cluster based on the distance or similarity between them. The similarity between two clusters is very important parameter for merging and dividing of cluster. Following methods are popularly used to calculate similarity between two clusters.

- Minimum or single linkage
- Maximum or complete linkage
- Mean or average linkage
- Centroid linkage
- Ward's method or minimum variance method

Minimum or single linkage method

Minimum linkage between two clusters is defined as the minimum distance between all pairs of points within two clusters.

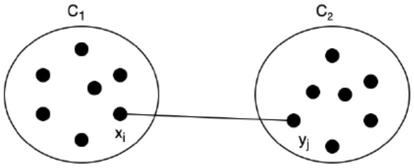


Figure 7 Single Linkage

Maximum or complete linkage method

Maximum linkage between two clusters is defined as the maximum distance between all pairs of points within two clusters.

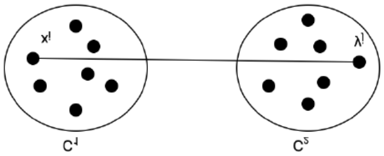


Figure 8 Complete Linkage

Mean or Average linkage method

The distance between two clusters is calculated by taking mean of similarity among all pair of points.

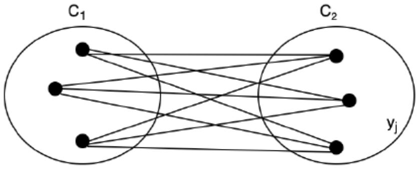


Figure 9 Average Linkage

Centroid linkage method

The distance between two clusters is measured as the distance between the centroids of the two clusters.

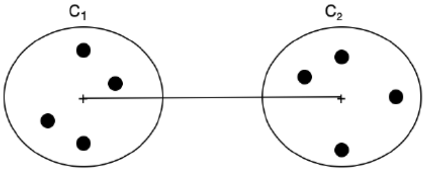


Figure 10 Centroid Linkage

Ward's method

The distance between two clusters is based on the similarity calculated as the sum of square of the of the distances x_i and y_j. This is very much similar to average method except it works on the sum of squares.

Steps in Agglomerative clustering

Consider data points a, b, c, d, e

- In the first step, we consider each point into a individual cluster.
- Let suppose point a and b closer to each other as compared to other points and formed one cluster by merging them together. Now we left with points {ab, c, d, e}
- Again, we calculate proximity with left out points. Let suppose again d and e are much closer than others so to merge into a cluster and left with the points { ab, c,de}.
- Once again calculate proximity matrix with remaining clusters. Let suppose cluster c and de are closer to each other and formed a new cluster cde and left with { ab, cde} clusters.
- Now we have left with only two cluster { ab, cde } so to merge them into one cluster as { abcde }.

Dendrogram

A dendrogram is a pictorial way to visualize hierarchical clustering. It is mainly used to show the outcome of hierarchical clustering a tree like diagram that records the sequences of merges and splits. The x axis represents the data points while y axis represents the distance/ heigh between the merged data points/ clusters. The below dendrogram plot is obtained by employing ward linkage method on the scaled dataset.

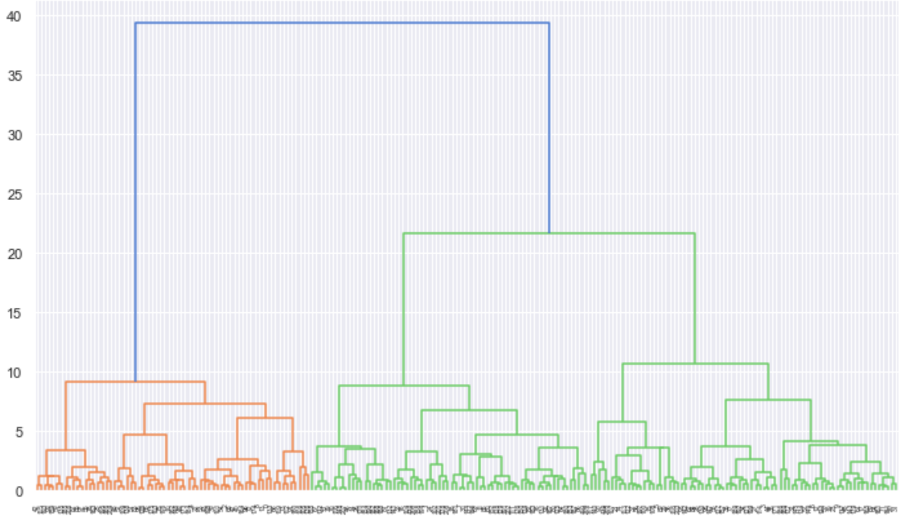


Figure 11 Dendrogram

The below truncated dendrogram is obtained by truncating it to get only the last 10 merges of the cluster. The truncating is performed to get a comprehensible view of the Dendrogram.

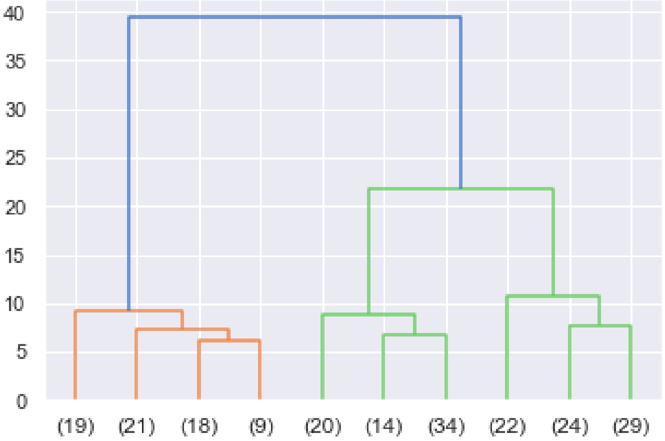


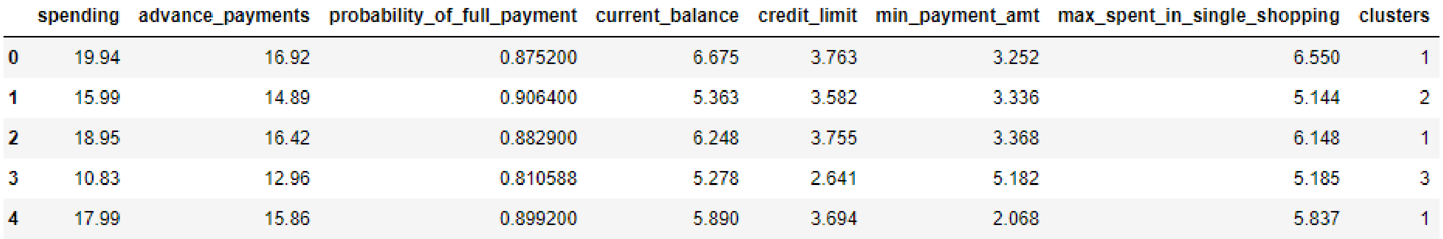
Figure 12 Truncated Dendrogram

Optimum Number of clusters

As per the above dendrogram, it is suggested to have 2 clusters represented by colors orange and green. However, considering the business scenario of Credit card usage, we would take 3 clusters considering the agglomerative height change approximately after 15 in the y axis. The cluster labels for each of the data point can be obtained by one of the two criteria using f cluster.

- Max clust
- Distance in the dendrogram

In max clust we need to mention the number of cluster that need to be formed, while in the distance in the dendrogram, we need to mention the cut off distance at the dendrogram plot, below which the number of first vertical lines denote the number of clusters. Thus, considering max clust of 3 or distance in the dendrogram as 15, we will be getting the cluster details of each data point in the dataset. The cluster details appended in the original dataset is as below.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
0	19.94	16.92	0.875200	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.906400	5.363	3.582	3.336	5.144	2
2	18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	3
4	17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1

Table 5 Sample of Clustered dataset

Note: *The fcluster functions from scipy.cluster.hierarchy library is an alternate method to get the clusters instead of agglomerative clustering.*

Customer Segmentation – Hierarchical Clustering

From the obtained clusters, customers can be segmented in to 3 segments High Spending, Medium Spending and low spending groups which is evident from the table 6. The mean spending of cluster 1 is 18.45 k while for cluster 2 is 14.38 k and 12.05 k for cluster 3. It can be seen that the advance payment made by them also follows a pattern similar to spending pattern where the cluster 1 pays more in advance and cluster 2 and 3 following up. The probability of full payment doesn't have a significant say in clustering as most of the customers have good past records. Hence, the probability column can be ignored during segmentation. The current balance column also follows a similar trend, wherein the cluster 1 records maintain a mean current balance more than that of the cluster 2 and cluster 3 respectively following up. It is also evident that the customers in cluster 3

(LSG) have a tendency to minimize credit card usage wherever possible, that can be witnessed by the highest average min amount paid on monthly purchases. However, the variable is inversely proportional to all the other columns. Similar to the other variables except min payment amount, max spending in single shopping also follows the similar pattern, wherein the cluster 1 (HSG) spend the highest average amount in single shopping, followed by cluster 2(MSG) and cluster 3 (LSG).

Clusters	Mean						
	Spending	advance payment	Probability of full payment	current balance	Credit limit	min payment amt	max spending in single shopping
1	18.45	16.19	0.88	6.17	3.69	3.69	6.04
2	14.38	14.31	0.88	5.51	3.25	2.45	5.12
3	12.05	13.33	0.85	5.25	2.88	4.83	5.10

Table 6 Mean patterns of clusters – Hierarchical clustering

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.

K-means clustering is an unsupervised learning algorithm with non-hierarchical approach whose goal is to find groups or assign the data points to clusters on the basis of their similarity. Which means the points in same cluster are similar to each other and in different clusters are dissimilar with each other. Here K means the number of clusters. The difference between K means and Hierarchical clustering is that, in K means we need to specify the number of clusters even before starting the clustering process.

Working of K means clustering

- Derive # no of clusters i.e., the optimal k value
- Partition dataset to k initial clusters (assume centroids)
- Assign each record to cluster with the nearest centroid (Eucledian distance)
- Re calculate centroid for losing & receiving clusters
- Check for reassignments, if yes, redo assignment of record to cluster to nearest centroid. If no, finalize the clusters.

Optimal K value

There are some techniques from which can be used to find the approximate or optimal value of **k**. These techniques include:

- Elbow Method
- Silhouette method

Elbow Method

It is the most popular and well-known method to find the optimal no. of clusters or the value of k in the process of clustering. This method is based of plotting the value of cost function against different values of k. As the number of clusters (k) increase lesser number of points fall within clusters or around the centroids. Hence the average distortion decreases with the increase of number of clusters. The point where the distortion declines most is said to be the elbow point and define the optimal number of clusters for dataset.

Silhouette Method

Silhouette is a different method to determine optimal number of clusters for given dataset. It defines as a coefficient of measure of how similar an observation is to its own cluster compared to that of other clusters. The range of silhouette coefficient varies between -1 to 1. 1 indicates that an observation is far from its neighbouring cluster and close to its own whereas -1 denotes that an observation is close to neighbouring cluster than its own cluster. The 0 value indicate the presence of observation on boundary of two clusters.

Optimal value of K in this case study

In this case study, we have found the optimal number of clusters through the scree plot obtained from the elbow method. The within cluster sum of squares are calculated for different values of k and are plotted as a scree plot as in the plot below (Figure13). The values for 10 k values are as follows 1469.94, 659.18, 430.14, 370.81, 327.10, 289.59, 262.29, 239.84, 220.57, 210.89. The y axis denotes the within sum of square values and the x axis denotes the k values. It can be seen from the plot that, the variation in within sum of squares is flattened after 3 values of k i.e., the variation with in clusters is minimized after 3 values of k. Hence, the optimal value of k in this case would be 3.

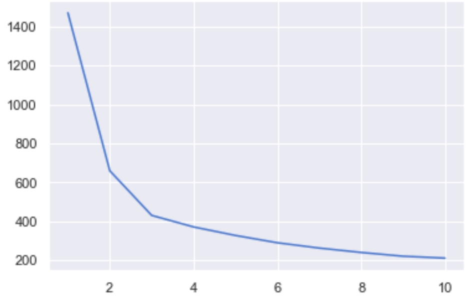
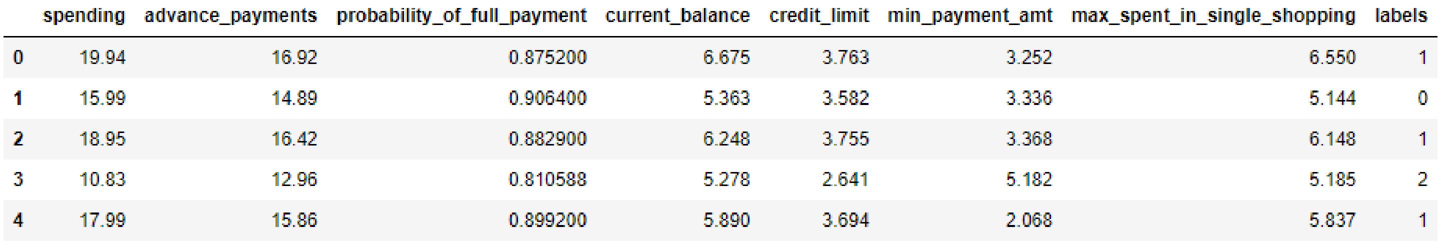


Figure 13 Scree Plot to find optimal number of clusters

Clustered Dataset

The cluster values are obtained and are assigned to the respective records in the data set. Sample of the clustered data set as below.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	labels
0	19.94	16.92	0.875200	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.906400	5.363	3.582	3.336	5.144	0
2	18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1

Table 7 Sample of KMeans clustered dataset

Model Evaluation – Silhouette width/ Silhouette score

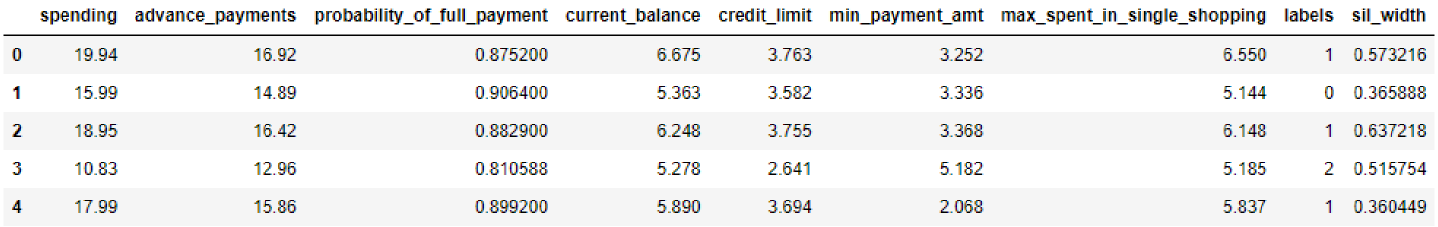
The Clusters thus obtained need to be validated for correctness of assignment of the records to respective clusters. There is a simple way to compute a standardized metric to arrive at this conclusion. The following formula that is:

Silhouette width = b / max(a,b)………………………………………………………………………………..*Equation 5 Silhouette width*

We can calculate silhouette width for each record, and when we take the average of silhouette widths that is called as silhouette score for a dataset.

If the silhouette score is close to +1 then we can say the clusters are well separated from each other on an average. If the silhouette score is close to 0, then we can say the clusters are not separated from each other. If the silhouette score is close to -1 then we can say the model has done a blunder in terms of clustering the data.

In this case study the silhouette widths are calculated for each of the record and appended in the dataset as in the below table. The minimum value of Silhouette width calculated is 0.003. Hence, we can conclude that there are no negative values in the silhouette widths. Value of Silhouette score is found to be 0.401, which indicates the clusters are approximately well separated from each other.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	labels	sil_width
0	19.94	16.92	0.875200	6.675	3.763	3.252	6.550	1	0.573216
1	15.99	14.89	0.906400	5.363	3.582	3.336	5.144	0	0.365888
2	18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1	0.637218
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	2	0.515754
4	17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1	0.360449

Table 8 Sample of KMeans clustered dataset with Silhouette width

Customer Segmentation – KMeans Clustering

From the obtained clusters, customers can be segmented in to 3 segments High Spending, Medium Spending and low spending groups which is evident from the table 6. The mean spending of cluster 1 is 18.50 k while for cluster 2 is 14.44 k and 11.86 k for cluster 0. It can be seen that the advance payment made by them also follows a pattern similar to spending pattern where the cluster 1 pays more in advance and cluster 2 and 0 following up. The probability of full payment doesn't have a significant say in clustering as most of the customers have good past records. Hence, the probability column can be ignored during segmentation. The current balance column also follows a similar trend, wherein the cluster 1 records maintain a mean current balance more than that of the cluster 2 and cluster 0 respectively following up. It is also evident that the customers in cluster 0 (LSG) have a tendency to minimize credit card usage wherever possible, that can be witnessed by the highest average min amount paid on monthly purchases. However, the variable is inversely proportional to all the other columns. Similar to the other variables except min payment amount, max spending in single shopping also follows the similar pattern, wherein the cluster 1 (HSG) spend the highest average amount in single shopping, followed by cluster 2(MSG) and cluster 0 (LSG). From the average values of Silhouette width column, we are able to infer that the clusters are well separated from each other.

Clusters	Mean							
	Spending	advance payment	Probability of full payment	current balance	Credit limit	min payment amt	max spending in single shopping	Silhouette width
1	18.50	16.20	0.88	6.18	3.70	3.63	6.04	0.47
2	14.44	14.34	0.88	5.51	3.26	2.71	5.12	0.34
0	11.86	13.25	0.85	5.23	2.85	4.73	5.10	0.40

Table 9 Mean patterns of clusters – KMeans clustering

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

Considering the small number of records in the dataset, we shall consider the clusters obtained through hierarchical clustering for profiling and recommendations.

Customer Profiling

We have already seen from the obtained clusters, customers can be segmented in to 3 segments High Spending, Medium Spending and low spending groups which is evident from the *table 6* of the *Customer Segmentation – Hierarchical Clustering.* We have also discussed on the means of the different variables and how well they are clustered. It can be seen that customers in the high spending group pay the highest advance payment by cash and also maintain a higher current balance and credit limit.

Recommendations

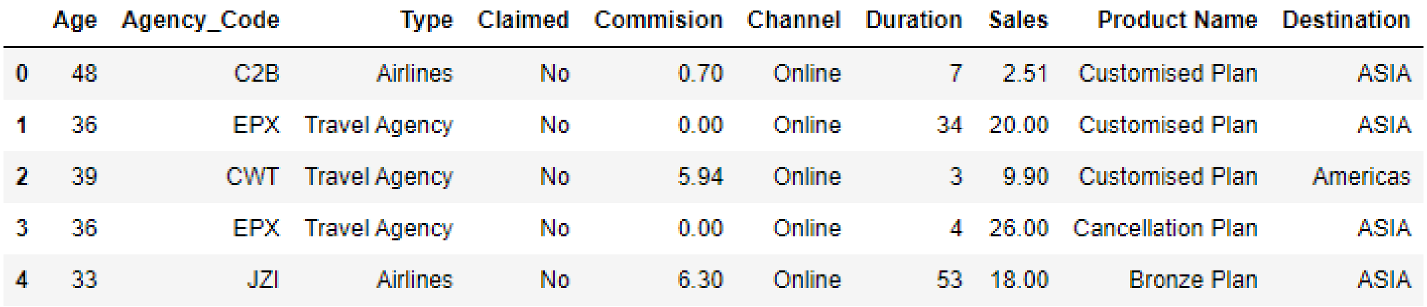
- Customers in the 3 clusters can be rewarded when their spending crosses a specific value (fixed based on the cluster they are present – higher amount for high spending customers, medium amount for medium spending customers and lower value for low spending customers). This helps in retaining the customers and encourages higher credit card usage.
- The customers in cluster 2 and 3 can also be motivated to increase their spending patterns by offering exclusive discounts/ cash back options in purchases for cluster 1 customers.
- Similar to the rewards offered when spending crosses a specific value, the customers can also be rewarded for maintaining a minimum current balance (the value of which is fixed based on their cluster)
- The credit limits of High spending customers can be relatively increased periodically, based on their continuous usage and similarly for the other clusters.

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.

Sample of the Dataset



	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.70	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0.00	Online	34	20.00	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.90	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0.00	Online	4	26.00	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.30	Online	53	18.00	Bronze Plan	ASIA

Table 10 Sample of dataset - Insurance problem

There are 3000 records with 10 columns containing the various characteristics of the travel plan for the past few years along with whether the insurance was claimed by the particular record.

Dataset Description

Age	Age of insured
Agency_Code	Code of tour firm (EPX/ C2B/ CWT/ JZI)
Type	Type of tour insurance firms (Travel Agency/ Airlines)
Claimed	Claim Status (Yes/ No)
Commission	The commission received for tour insurance firm as % of sales
Channel	Distribution channel of tour insurance agencies (Online/ Offline)
Duration	Duration of the tour
Sales	Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)

Product Name	Name of the tour insurance products (Customized Plan/ Cancellation Plan/ Bronze Plan/ Silver Plan/ Gold Plan)
Destination	Destination of the tour (Asia/ America/ Europe)

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

EXPLORATORY DATA ANALYSIS

Variable type

Age	int64
Agency_Code	object
Type	object
Claimed	object
Commission	float64
Channel	object
Duration	int64
Sales	float64
Product Name	object
Destination	object

Here, we are able to notice that there are two columns with int64 data type, two columns with float64 data type & six columns with object data type. For Decision Trees, Random Forest & ANN, it is mandatory for the variables to be of numerical continuous type. Hence, we need to convert the datatype of six columns from object to integer. This is achieved by converting them in to categorical types and extracting the codes for each category. Thus, post conversion, the data type of the variables is as below.

Age	int64
Agency_Code	int8
Type	int8
Claimed	int8
Commission	float64

Channel	int8
Duration	int64
Sales	float64
Product Name	int8
Destination	int8

The corresponding assignment of labels to the categorical conversion fields are as in the below table

Labels	Agency Code	Type	Claimed	Channel	Product Name	Destination
0	C2B	Airlines	No	Offline	Bronze Plan	Asia
1	CWT	Travel Agency	Yes	Online	Cancellation Plan	America
2	EPX				Customised Plan	Europe
3	JZI				Gold Plan	
4					Silver Plan	

Table 11 Categorical Assignment Mapping

Check for missing values in the dataset

Age	3000 non-null
Agency_Code	3000 non-null
Type	3000 non-null
Claimed	3000 non-null
Commission	3000 non-null
Channel	3000 non-null
Duration	3000 non-null
Sales	3000 non-null
Product Name	3000 non-null
Destination	3000 non-null

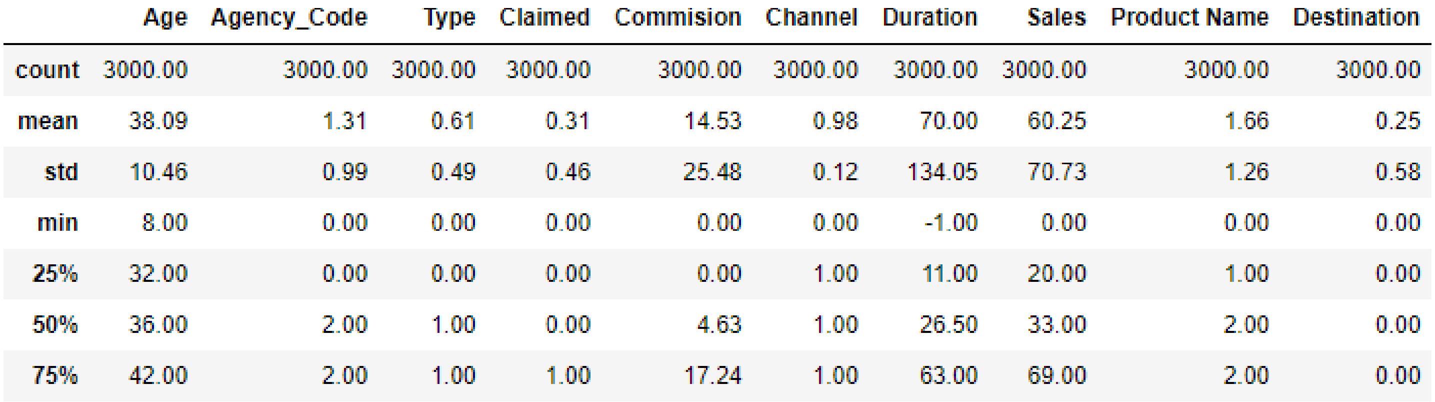
There are no missing values in the entire dataset.

Summary of the dataset

Summarizing briefly, the dataset has a total of 3000 records. Mean age of the insured is 38.09. We are aware that the Agency code, type, claimed, channel, product name and destination are categorical variables.

They don’t provide any meaningful information and hence can be ignored while summarizing the statistical information. The mean commission received for tour insurance firm is 14.53 % of the sales. The mean duration is 70 days and mean sales value is 6025. There is variation in the mean and median values of the discussed columns (except age) which implies that there might be outliers that are skewing the curve towards one side.

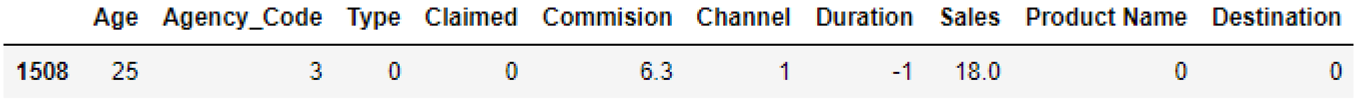
It can also be seen that there is negative value of -1 in the duration, which is bad data, as there can be no negative duration. Hence, we need to treat these data by imputing or dropping the records from our dataset.



	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000.00	3000.00	3000.00	3000.00	3000.00	3000.00	3000.00	3000.00	3000.00	3000.00
mean	38.09	1.31	0.61	0.31	14.53	0.98	70.00	60.25	1.66	0.25
std	10.46	0.99	0.49	0.46	25.48	0.12	134.05	70.73	1.26	0.58
min	8.00	0.00	0.00	0.00	0.00	0.00	-1.00	0.00	0.00	0.00
25%	32.00	0.00	0.00	0.00	0.00	1.00	11.00	20.00	1.00	0.00
50%	36.00	2.00	1.00	0.00	4.63	1.00	26.50	33.00	2.00	0.00
75%	42.00	2.00	1.00	1.00	17.24	1.00	63.00	69.00	2.00	0.00

Table 12 Statistical summary of the dataset

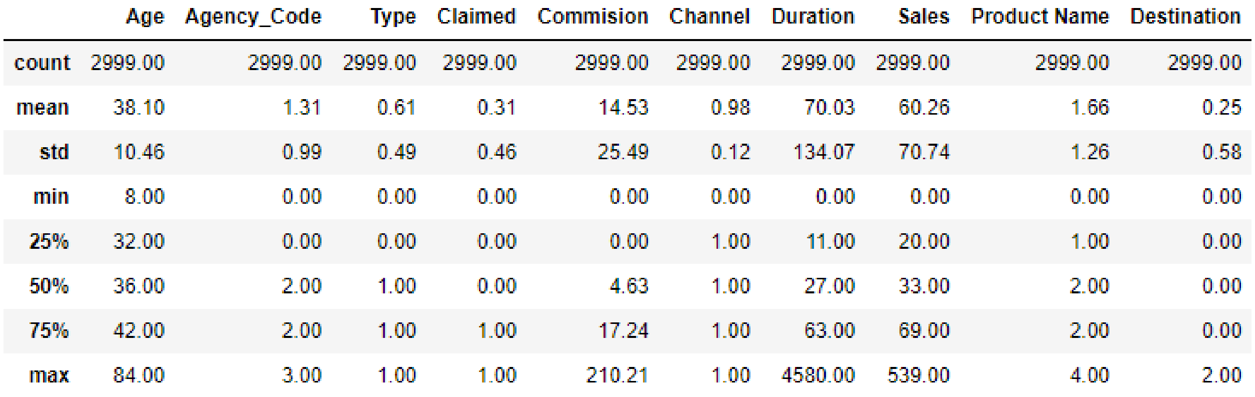
On checking the dataset, there was only one record with negative duration which might be a typographical error. Considering only 1 value, we can go ahead removing that row from the dataset.



	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
1508	25	3	0	0	6.3	1	-1	18.0	0	0

Table 13 Negative duration record

The revised summary as in the below table.



	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	2999.00	2999.00	2999.00	2999.00	2999.00	2999.00	2999.00	2999.00	2999.00	2999.00
mean	38.10	1.31	0.61	0.31	14.53	0.98	70.03	60.26	1.66	0.25
std	10.46	0.99	0.49	0.46	25.49	0.12	134.07	70.74	1.26	0.58
min	8.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	32.00	0.00	0.00	0.00	0.00	1.00	11.00	20.00	1.00	0.00
50%	36.00	2.00	1.00	0.00	4.63	1.00	27.00	33.00	2.00	0.00
75%	42.00	2.00	1.00	1.00	17.24	1.00	63.00	69.00	2.00	0.00
max	84.00	3.00	1.00	1.00	210.21	1.00	4580.00	539.00	4.00	2.00

Table 14 Summary of the cleaned dataset

It can be seen that the cleaned dataset consists of only 2999 records.

DATA VISUALIZATION

Histogram Plot

For the continuous variables in the dataset, the histograms are plotted and the corresponding skewness and kurtosis values are calculated. The high values of skewness and kurtosis suggests that the data are not symmetrical and not having proper peaks. This implies that the data is not normally distributed.

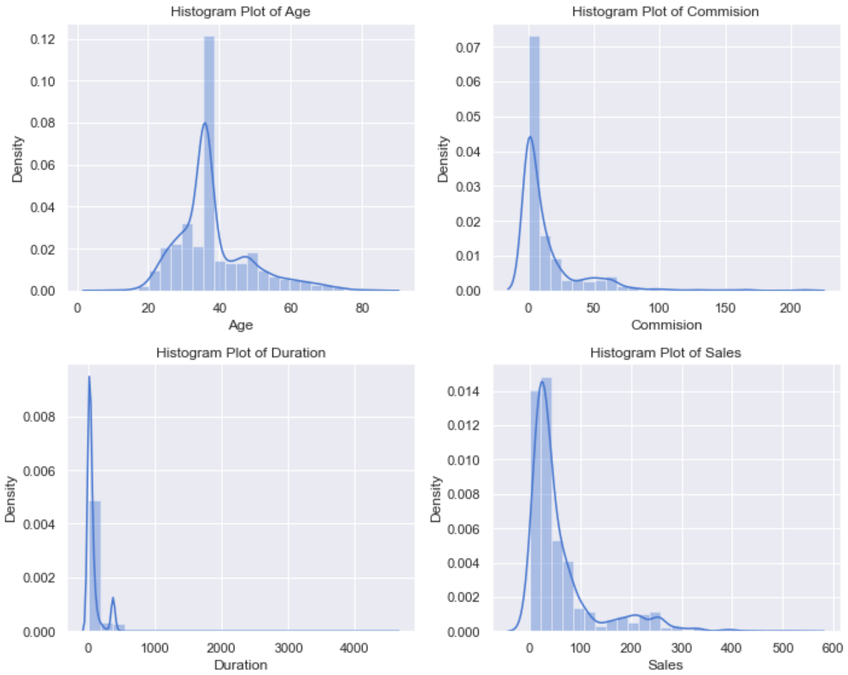
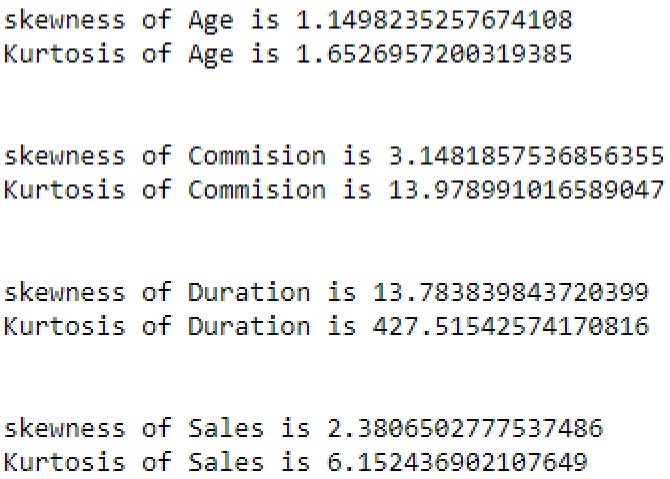


Figure 14 Histogram Plot - Continuous Variables



```
skewness of Age is 1.1498235257674108
Kurtosis of Age is 1.6526957200319385


skewness of Commision is 3.1481857536856355
Kurtosis of Commision is 13.978991016589047


skewness of Duration is 13.783839843720399
Kurtosis of Duration is 427.51542574170816


skewness of Sales is 2.3806502777537486
Kurtosis of Sales is 6.152436902107649
```

Figure 15 Calculated values of skewness & kurtosis

Count Plot

For the categorical variables in the dataset, the frequency of occurrence is plotted and the corresponding plots are as below. With respect to claims made by Agencies, it can be seen that the highest number of claims are made by C2B followed by EPX, CWT & JZI. However, EPX has recorded the least proportion of claims. With regards to the type of firm, Airlines has made the greatest number of claims, while the travel agencies have the greatest number of no claims. We can also see that the entire data has an imbalance in the claims & no claims that can be seen from the count plot of claimed in the below figure. The data is imbalanced in the ratio 70:30 for no claim: claim. It can also be seen that the claims are mostly made online. With respect to the plans, silver plan insured have claimed the most, while cancellation plan insured have claimed the least. It can also be seen that there are a greater number of customized plans. With respect to destination, we can note that there are lot of people traveling to Asia, and they have made most of the claims and the least claims are made by people travelling to Europe.

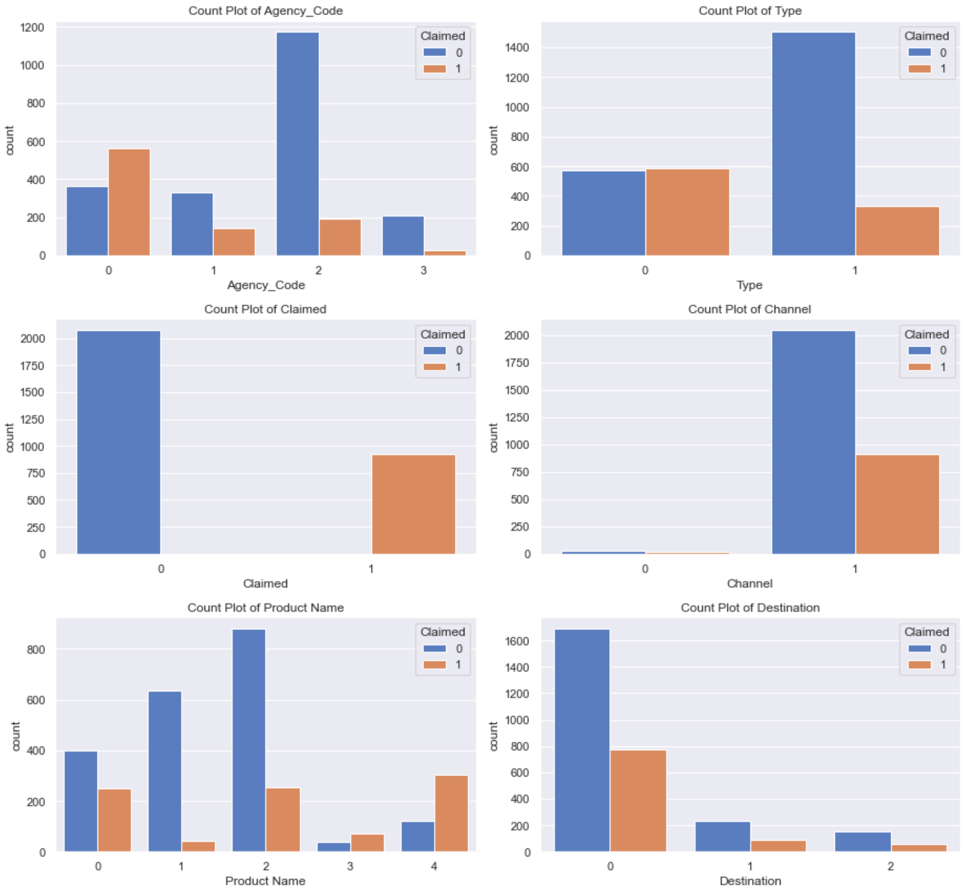


Figure 16 Count Plot of Categorical Variables

Box Plot

From the below box plot, it can be seen that there are a lot of outliers in all the variables. However, these outliers are not going to be treated, as these are original data and would have influence in the final out come of the prediction. Hence it is not advisable to treat the outliers.

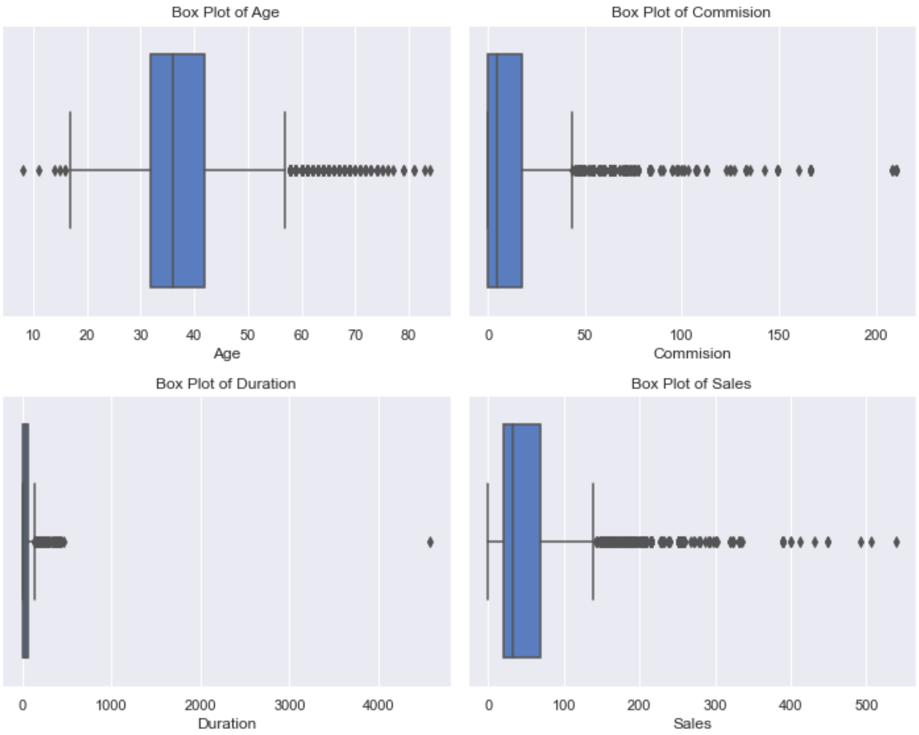


Figure 17 Box Plot of Continuous Variables

Correlation Plot – Relation amongst columns

The below figure shows the correlation plot (heat map) of the variables in the dataset. The figure indicates that there is weak correlation among the variables. However, there is measurable correlation between commission and duration, commission and sales, commission and product name, duration and sales, sales and Product name & Type and Agency code. This correlation is not so significant to affect the prediction model.

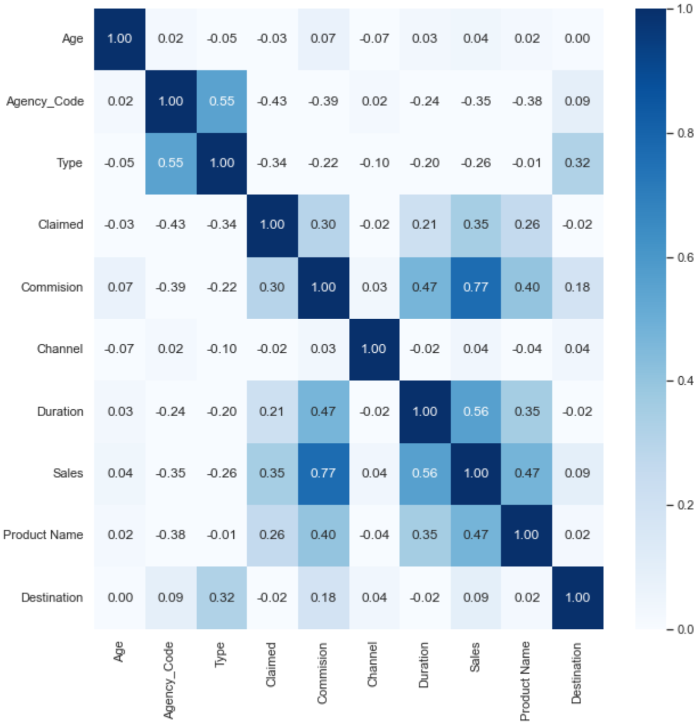


Figure 18 Correlation Plot (Heat Map) - Insurance Dataset

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

DATA SPLIT

To split the data in to training & testing data, we need to first create separate data frames for the independent & dependent variables. This is achieved by performing drop and pop functions on the Claimed variable. Once the data set is separated, the split of training & testing data can be achieved through train_test_split function in sklearn.model_selection library. This function requires the independent & dependent variables, percentage data to be considered for testing and the random state. The independent & dependent variables can be passed as we have created them. Usually, to obtain good results, we need to train the model with a larger number of data points, so that it would be able to capture most of the characteristics of the dataset. Hence, we usually select the train: test samples in the ratio 70:30. Hence, the input for the testing percentage to be considered will be 0.3. The random state considered here is 1. Random state is to ensure that the set of codes when run in a different jupyter notebook will give the same values for the defined random state.

CART MODEL BUILDING:

CART – Classification And Regression Tree is a binary decision tree that can give both categorical & continuous output variable. The most commonly used splitting criteria is the Gini index criteria. Gini index is the measure of impurity of a node. For a highly pure node, Gini index will be equal to 0. The purity here is nothing but the measure of independence of a variable.

In order to build a CART model, we first need to import the DecisionTreeClassifier from sklearn.tree. Post importing, we need to build the model using the gini criterion and fit and transform the training data (both dependent & independent variables). Once the model is built, we need to visualize it as a decision tree, for which we create a dot file and using the tree.export_graphviz function, we can impute the model to create the code required to visualize the tree. The tree will be too large, depending up on the number of records in the dataset. However, from the Decision tree built, we can see that the tree is over grown/ over fit. Hence, we need to select the hyper parameters to get the optimum tree. The optimum parameters can be obtained by the Grid search Cross validation function of model selection library in the SKlearn module. The parameters for building the CART model include maximum depth, min sample leaf & min sample split. Maximum depth refers to the number of branches that the tree can be split along vertically, min sample leaf refers to the minimum number of records that a node must contain after splitting and the min sample split refers to the minimum number of records that a node must have so that it can be split. These parameters are initialized and the Grid Search cross validation is run for the initialized hyper parameters. The term cross validation refers to the number of times the Grid Search CV runs to revalidate the hyper parameters. We select a value of 3 for Cross validation. Greater the CV number, higher the time to execute the python codes. Once the Grid search CV is run, we need to fit our training dataset in to the Grid Search CV model and the best parameters are obtained. Using these best hyper parameters, we shall predict the values for both training & testing dataset. However, depending on the classification report, we can re-run the code multiple times with different values for initializing the hyper parameters to get the best results. The best parameters obtained in this insurance case study is maximum depth of 4, minimum sample leaf of 15 and min sample split of 120.

RANDOM FOREST MODEL BUILDING:

Random Forest is a technique, where in multiple CART models are built to get the accuracy in the predicted value. The model building is same except that we need to build a random forest classifier model instead of decision tree classifier model. We will not be able to visualize the trees in this case, as there will be multiple decision trees built. The optimum hyper parameters here are also obtained by grid search cross validation. In addition to the hyper parameters of Decision tree classifier, we have max_features and n_estimators. Max_features refer to the number of variables to be considered and the n_estimators define the number of decision trees to be built. Using these best hyper parameters, we shall predict the values for both training & testing dataset. However, depending on the classification report, we can re-run the code multiple times with different values for initializing the hyper parameters to get the best results. The best parameters obtained in this insurance case study is maximum depth of 7, maximum features: 7, minimum sample leaf of 20, min sample split of 90 and n_estimators of 501.

ARTIFICIAL NEURAL NETWORK MODEL BUILDING:

A Machine Learning algorithm that is roughly modelled around what is currently known about how the human brain functions. They have the ability to learn, generalize, & adapt. This method is also called the black box method. Neural networks are made with many layers of inter connected nodes. There are 3 main layers – Input layer, Hidden Layers & Output layer. Hidden layers can be one or more. Input nodes process the incoming data exactly as received. Adds one or more hidden layers that process the signals from the input nodes prior to reaching the output node. The output node gives a predicted value. The difference between predicted value and actual value is the error. Error is propagated backward by apportioning them to each node's weights. The artificial neural networks employ a weight-based algorithm. Hence the data need to be scaled before building the model. The scaling is performed using the Standard Scaler function. The train data are fit and transformed in the standard scaler, while test data is only transformed. The hyper parameters here are the hidden layer sizes, Activation, solver, tolerance and maximum iterations. The hidden layer sizes defines the number of nodes in hidden layers, activation denotes the type of activation function to be used, tolerance defines the learning rate and maximum iteration defines the number of times he model will iterate during training. The best parameters are obtained by running the hyper parameters in the grid search cross validation function. However, depending on the classification report, we can re-run the code multiple times with different values for initializing the hyper parameters to get the best results. The best parameters thus obtained are hidden layers – 100, activation – relu, maximum iteration – 10000, solver – adam, tol – 0.001. Using these hyper parameters the dependent variable is predicted in the train and test data.

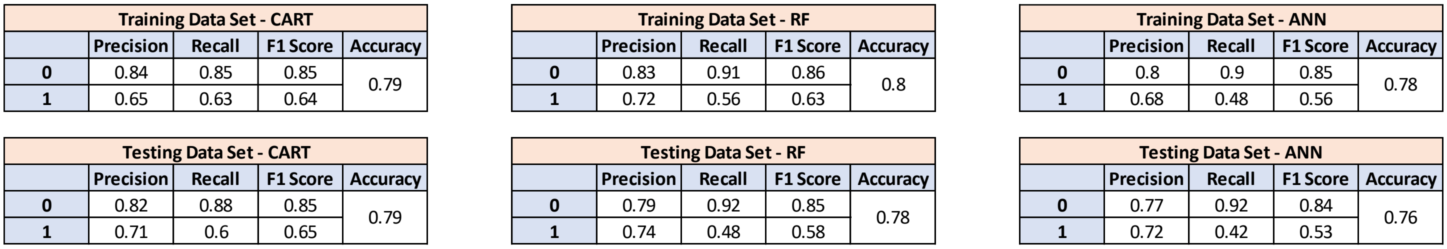
2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score, classification reports for each model.

&

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

The Performance metrics are obtained for each of the models as below.

Classification Reports:



Training Data Set - CART				
	Precision	Recall	F1 Score	Accuracy
0	0.84	0.85	0.85	0.79
1	0.65	0.63	0.64	

Training Data Set - RF				
	Precision	Recall	F1 Score	Accuracy
0	0.83	0.91	0.86	0.8
1	0.72	0.56	0.63	

Training Data Set - ANN				
	Precision	Recall	F1 Score	Accuracy
0	0.8	0.9	0.85	0.78
1	0.68	0.48	0.56	

Testing Data Set - CART				
	Precision	Recall	F1 Score	Accuracy
0	0.82	0.88	0.85	0.79
1	0.71	0.6	0.65	

Testing Data Set - RF				
	Precision	Recall	F1 Score	Accuracy
0	0.79	0.92	0.85	0.78
1	0.74	0.48	0.58	

Testing Data Set - ANN				
	Precision	Recall	F1 Score	Accuracy
0	0.77	0.92	0.84	0.76
1	0.72	0.42	0.53	

Table 15 Performance Metrics - All Models

From the above table, we can see that accuracy of prediction is almost same ranging from 0.78 to 0.80 in training dataset and 0.76 to 0.79 in the testing dataset. However, the CART model shows the highest accuracy. But, the same is not reliable when compared to all the 3 models, as neural network is much more advanced and takes multiple parameters in to consideration. The main parameter to be considered here is the recall value of the model as the prime objective in this case study is to reduce the False negative values i.e., predicting Not claimed when the insured has actually claimed.

Confusion Matrix:

Training Data - CART		Predicted	
		0	1
Actual	0	1251	214
	1	235	399

Training Data - RF		Predicted	
		0	1
Actual	0	1327	138
	1	277	357

Training Data - ANN		Predicted	
		0	1
Actual	0	1325	140
	1	332	302

Testing Data - CART		Predicted	
		0	1
Actual	0	538	72
	1	117	173

Testing Data - RF		Predicted	
		0	1
Actual	0	562	48
	1	151	139

Testing Data - ANN		Predicted	
		0	1
Actual	0	562	48
	1	167	123

Table 16 Confusion Matrix - All Models

The main objective of this case study is to reduce the False Negative value during prediction i.e., index (1,0). We can see here that the FN value is high. However, this occurs due to imbalance in the dataset of training and also the original complete dataset. Hence this is the best possible that can be achieved.

AUC:

Area under the curve values are as below. It is evident that all the models have higher area for training sets while the testing sets have lower area. However, the difference in area is the least for ANN and hence considered the best model for prediction.

	CART	RF	ANN
AUC - Train	0.825	0.849	0.729
AUC - Test	0.8	0.827	0.708

Table 17 Area under the curve - All models

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

Based on the analysis of the 3 models, we can conclude that the Artificial Neural Networks model is the best, as it is evident from the above metrics.

Recommendation:

- It is recommended that the management focus on the agency that has the highest claim rate. In this case it is the C2B & the reasons for the claims. This would help the management alter their policies to reduce the claims. It can also make a comparative analysis with EPX, so that it can get an hold on what went wrong.

- The management must also look in to the insurance policies of the Airlines, as most of the claims are made by the insured travelling via air. There must be stricter norms.
- The management can also consider altering the existing policies of Product names to leverage the most.
- The management can also consider framing policies based on the destination they travel. It is obvious that the insured travelling to Asia has made the greatest number of claims. This can be due to the developing air space in the countries of Asia.

<u>END</u>