DATA MINING PROJECT REPORT

By

M.P.KARTHIKEYAN

Contents Problem

Problem 1: Clustering

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.

Problem 2: CART-RF-ANN

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.

List of Figures

List of Tables

Problem 1: Clustering

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.

Executive summary

A leading bank collected the user data. The sample data summarizes the activities of users .With this dataset we need to segment customers based on credit card usage so that promotional offers can be designed individually for different segments

Introduction

The given dataset contains details about 210 customer's credit card usage.Exploratory data analysis is done.The clustering methods such as hierarchical clustering and K-Means clustering are adopted to segment the customers into different clusters.

Sample Dataset

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

Table-1 Sample 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
5	12.70	13.41	0.8874	5.183	3.091	8.456	5.000
6	12.02	13.33	0.8503	5.350	2.810	4.271	5.308
7	13.74	14.05	0.8744	5.482	3.114	2.932	4.825
8	18.17	16.26	0.8637	6.271	3.512	2.853	6.273
9	11.23	12.88	0.8511	5.140	2.795	4.325	5.003

Exploratory Data Analysis

Let us check the type of variables

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

The dataset contains 210 rows and 7 columns.All 7 columns are of float data type.

Check for missing values in dataset

spending	210 non-null	float64
advance_payments	210 non-null	float64
probability_of_full_payment	210 non-null	float64
current_balance	210 non-null	float64
credit_limit	210 non-null	float64
min_payment_amt	210 non-null	float64
max_spent_in_single_shopping	210 non-null	float64

From the above values it is clear that there are no missing values in dataset.

Descriptive Statistics

Descriptive statistics are used to describe about the variables in dataset by giving short summaries about the sample and the measures of data.

The most recognized types of descriptive statistics are measures of centre**: the mean, median, and mode**, which are used at almost all levels of math and stati stics.

Table1.1-Summary of data

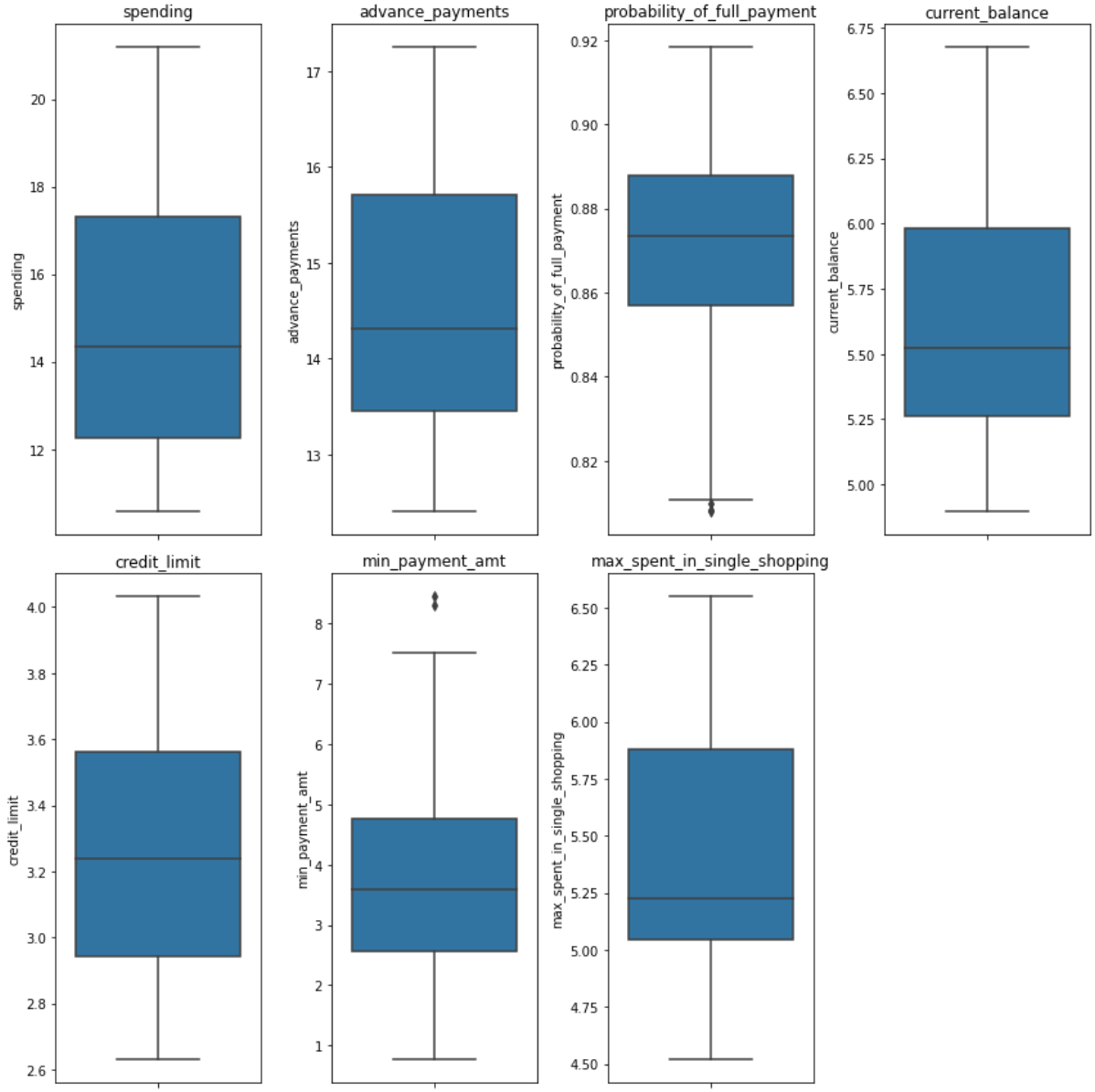
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

From the above table it is evident that the average amount spent by customers is 14.35 (in thousands).A customer pays 14.32 (in hundreds) as advance on an average.The average probability of customer paying full payment is 0.87.The Current balance maintained by customers range between 4.89 to 6.67 (in thousands).The average credit card limit of customer is 3.23(in thousands).The maximum amount spent in single shopping is 6.55(in 1000s)

Uni-Variate Analysis

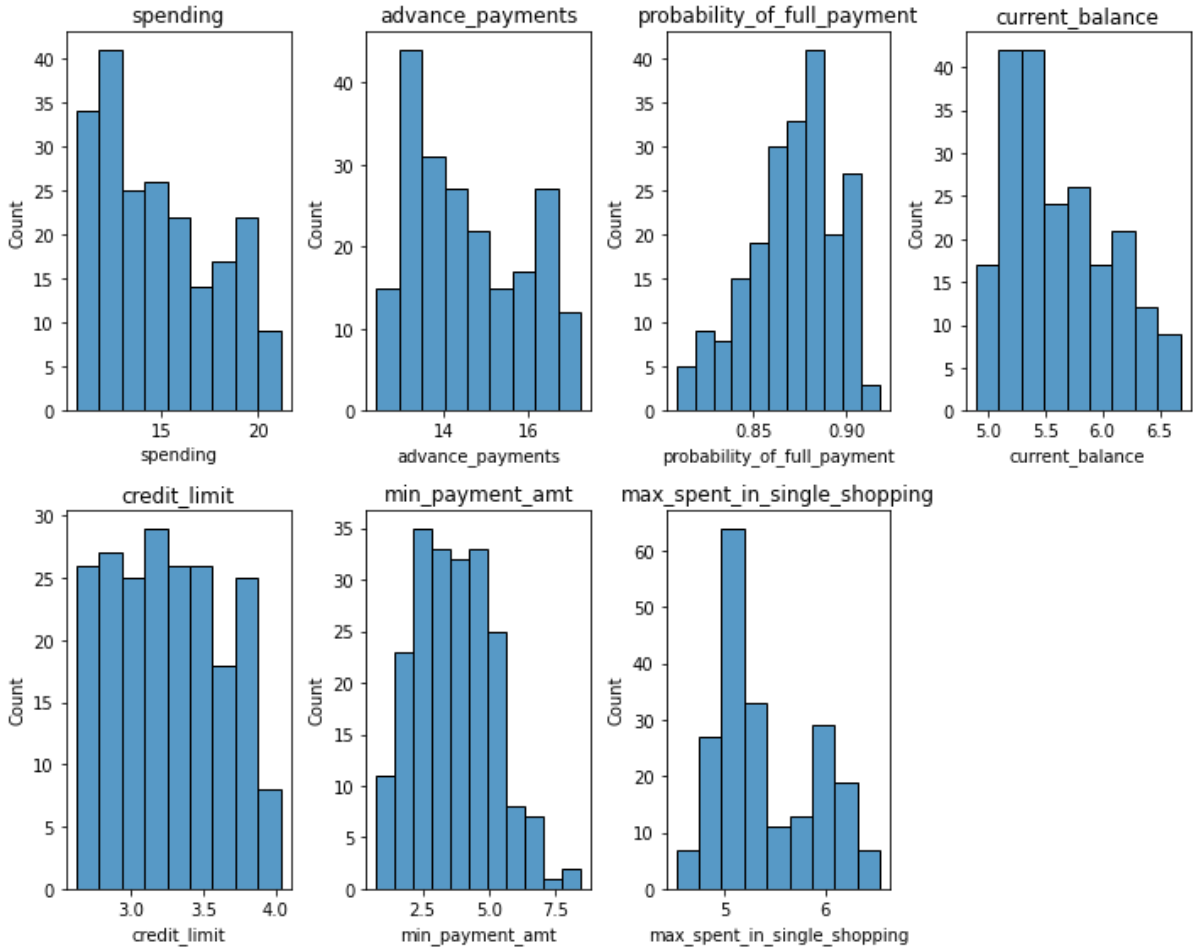
Check for outliers

Fig 1.1 Boxplot of Variables



There are many outliers in probability of full payment & min_payment.

Fig 1.2 Histogram of Variables



From the histogram the distribution of variables seems to be slightly skewed.

Bi-Variate Analysis

Fig -1.3 Scatter plot

Max spent in single shopping vs Min Payment Amount

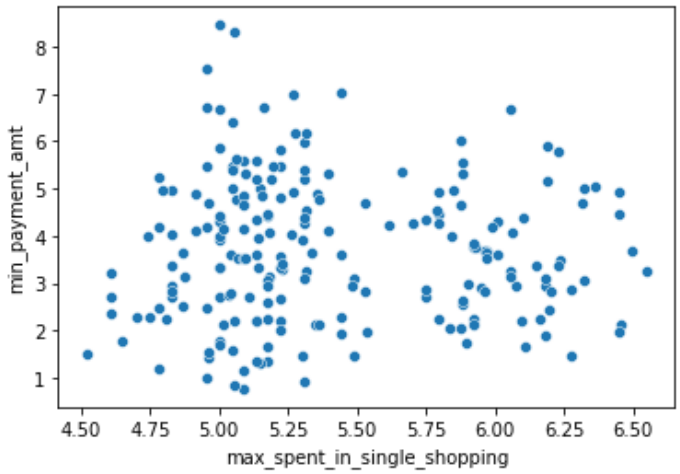


Fig -1.4 Scatter plot

Spending vs Advance Payment

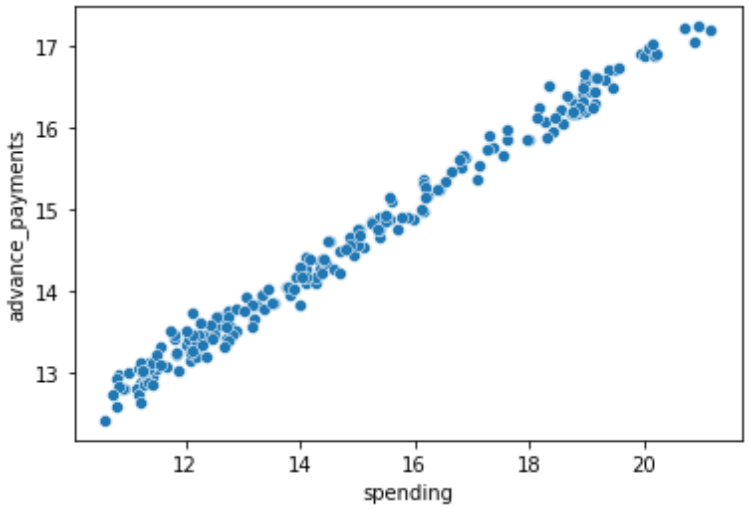


Fig -1.5 Scatter plot

Spending vs Probability of full Payment

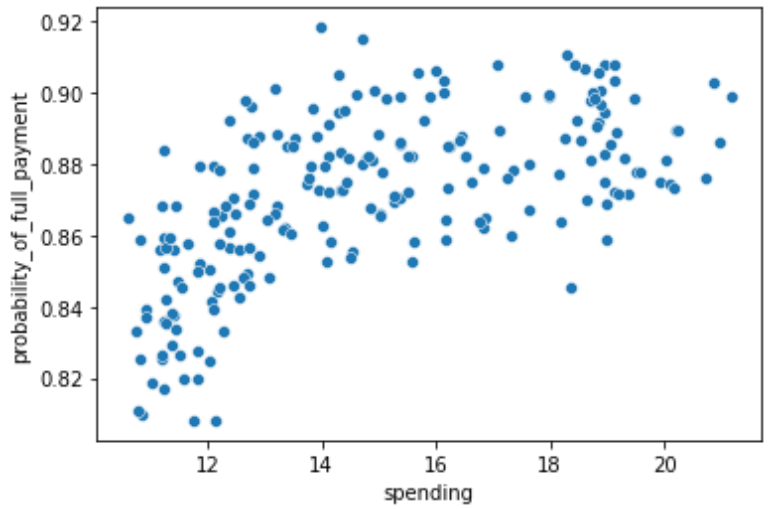


Fig -1.6 Scatter plot

Spending vs Current balance

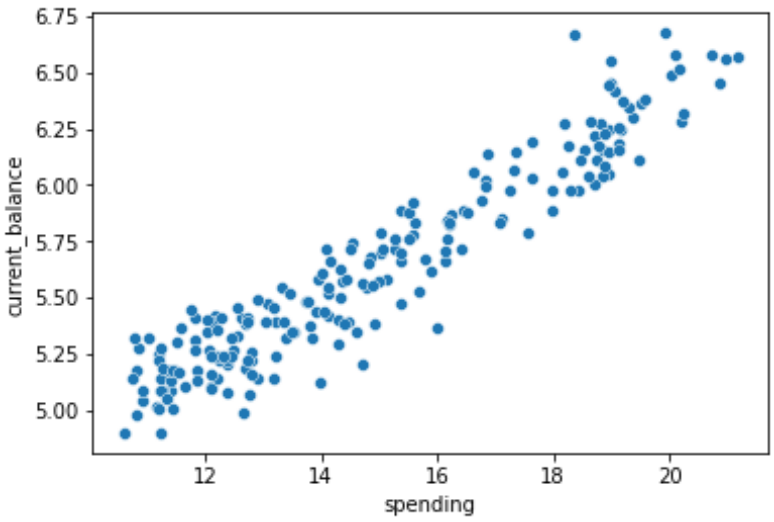
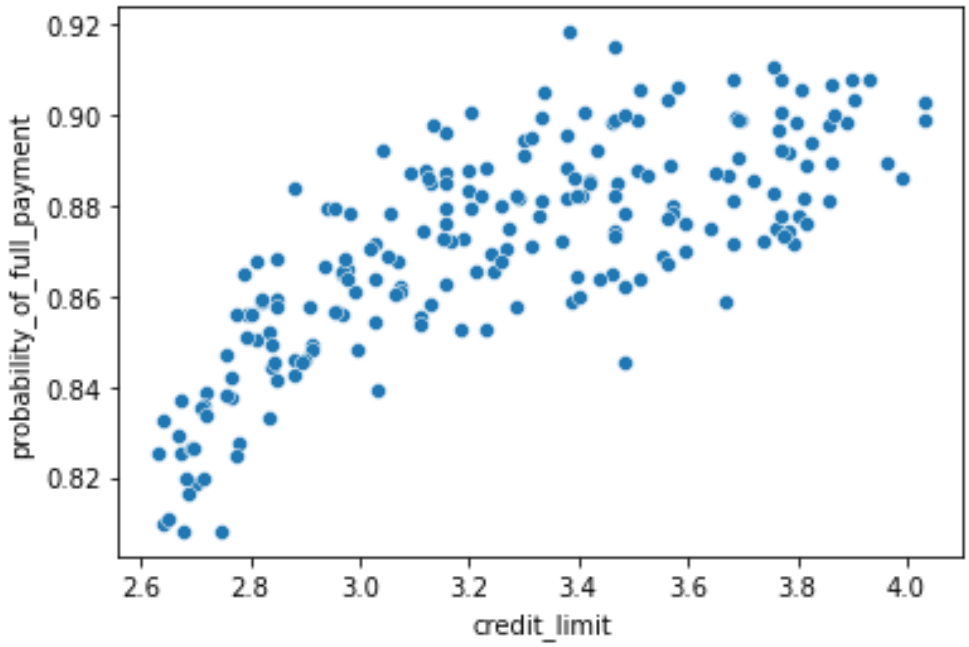


Fig -1.7 Scatter plot

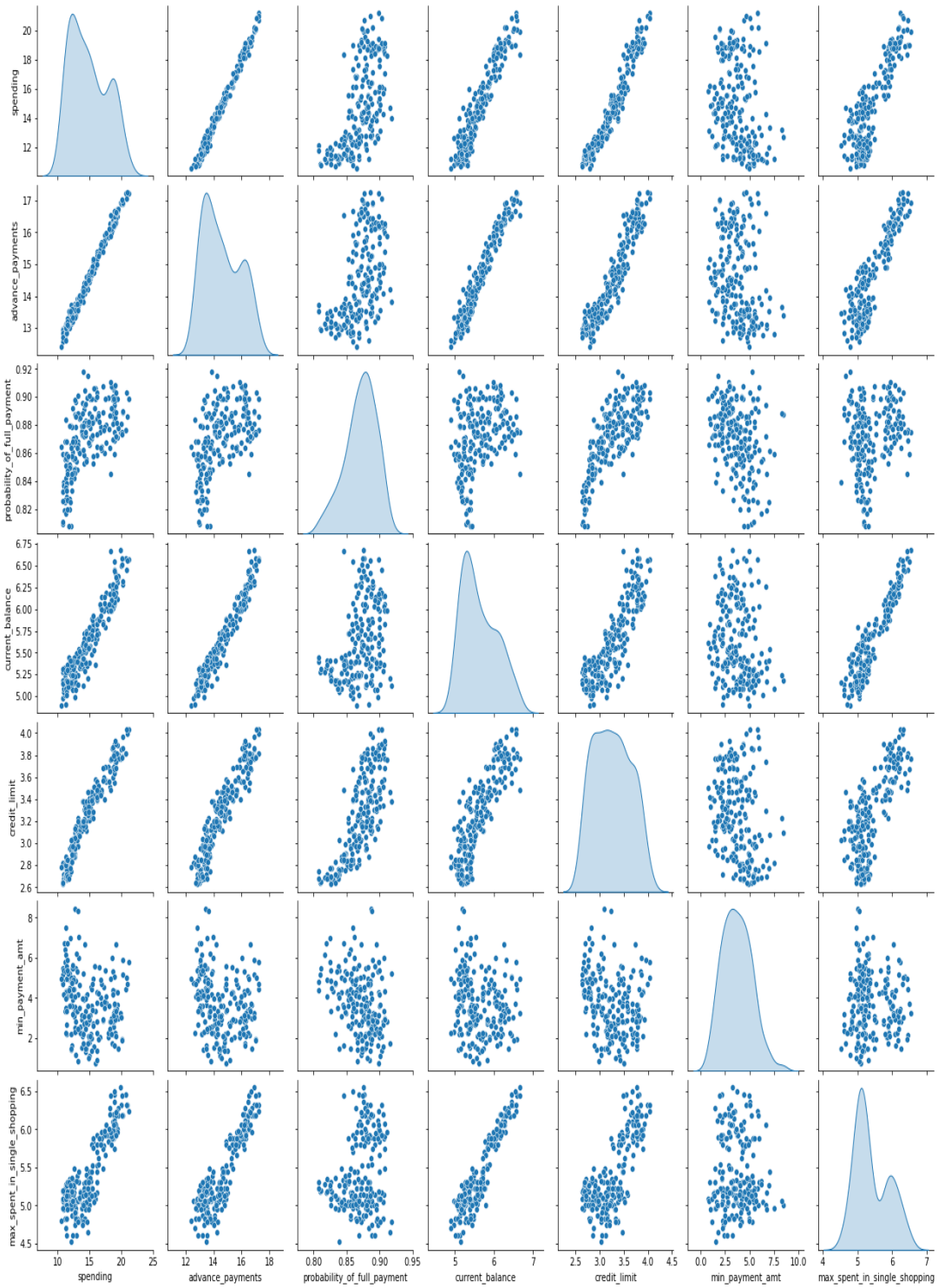
Probability of full payment vs Credit Limit



The scatterplots are plotted between two variables to identify the relationship between the variables.From the plots we can see the spending is in linear relationship with advance payments and current balance.

Multi-Variate Analysis

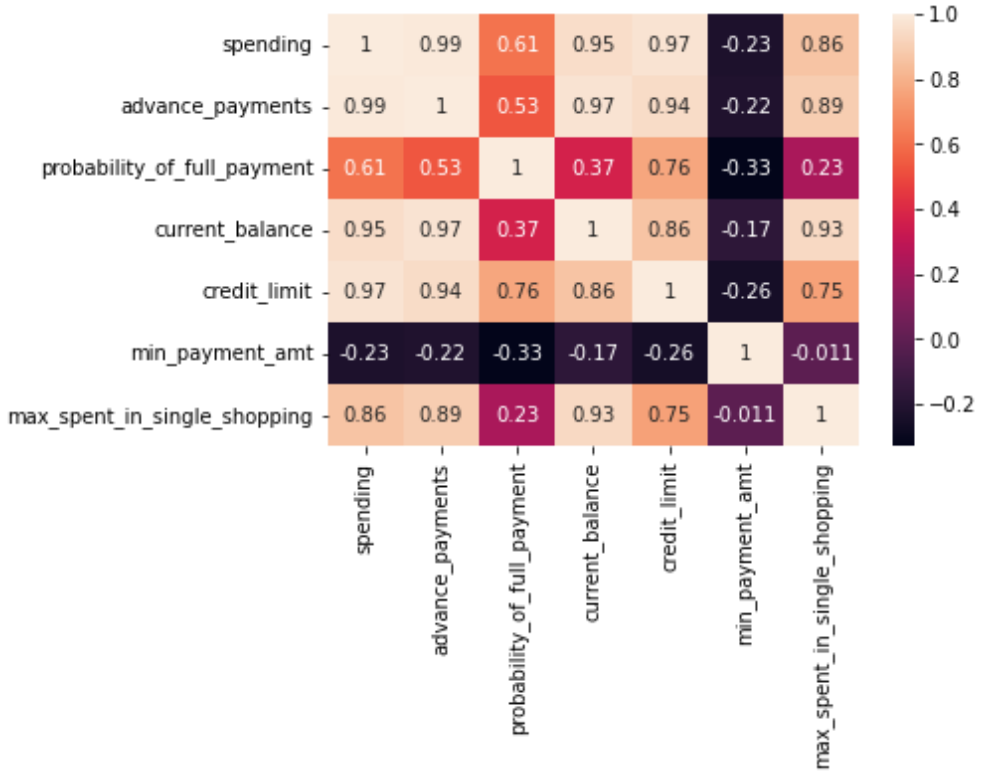
Fig -1.8 Pair plot



Checking the Pairwise distribution of continuous variables.

The Pair plot is plotted between the variables to check the distribution pattern of one variable with the influence of other variable

Fig -1.9 Heat Map



The correlation between the variables are established using heatmap.From the plot we can see many variables are highly correlated to each other such as spending and advance payments,spending and credit limit,advance payment and credit limit .

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

Yes ,I prefer to scale the data before performing any kind of clustering techniques as all those techniques use distance as metric to group the data into various clusters.The scaling can be done by using various methods such as Z-Scale/Standard scaler/Min-max scaler.

Here the scaling is done using Z-score.

Table1.2-Summary of Scaled data

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02
mean	9.148766e-16	1.097006e-16	1.243978e-15	-1.089076e-16	-2.994298e-16	5.302637e-16	-1.935489e-15
std	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00
min	-1.466714e+00	-1.649686e+00	-2.668236e+00	-1.650501e+00	-1.668209e+00	-1.956769e+00	-1.813288e+00
25%	-8.879552e-01	-8.514330e-01	-5.980791e-01	-8.286816e-01	-8.349072e-01	-7.591477e-01	-7.404953e-01
50%	-1.696741e-01	-1.836639e-01	1.039927e-01	-2.376280e-01	-5.733534e-02	-6.746852e-02	-3.774588e-01
75%	8.465989e-01	8.870693e-01	7.116771e-01	7.945947e-01	8.044956e-01	7.123789e-01	9.563941e-01
max	2.181534e+00	2.065260e+00	2.006586e+00	2.367533e+00	2.055112e+00	3.170590e+00	2.328998e+00

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

Hierarchical Clustering (Agglomerative)

A dendrogram is a visual representation of cluster-making. On the x-axis are the item names or item numbers. On the y-axis is the distance or height. The vertical straight lines denote the height where two items or two clusters combine. The higher the level of combining, the distant the individual items or clusters are. By definition of hierarchical clustering, all items must combine to make one cluster.

Ward's Linkage: Here we adopted Ward's Linkage method also known as minimum variance clustering method, an iterative method where after every merge, the distances are updated successively. Ward's method often creates compact and even-sized clusters.

Fig – 1.10 Dendrogram of Clusters

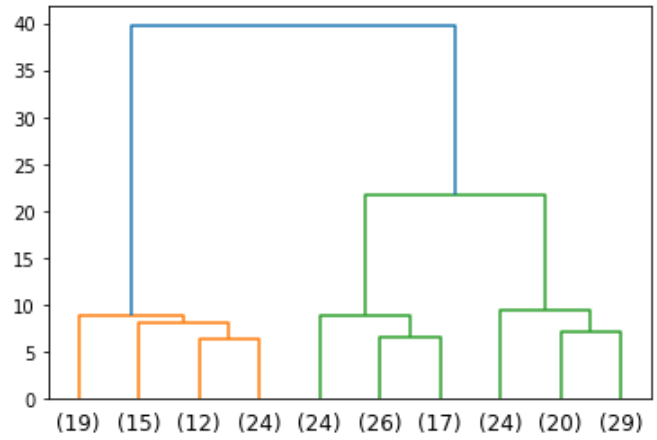


Figure 1.10 does not show the full grown tree but just the number of Users in each cluster. Cluster size is not equal for all 10 clusters, as indicated in the above diagram. It is important to note that based on the method selected for linkage and affinity, cluster membership and cluster size could vary.

Table-1.3 Sample Dataset with Clusters

	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	0.8752	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1

Here the optimum number of clusters is identified as 3 and the dataset has been clustered into 3 clusters.

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

k-means clustering is the most used non-hierarchical clustering technique. It aims to partition n observations into k clusters in which each observation belongs to the cluster whose mean (centroid) is nearest to it, serving as a prototype of the cluster. It minimizes within-cluster variances (squared Euclidean distances).

Elbow Curve

For a given number of clusters, the total within-cluster sum of squares (WCSS) is computed. That value of k is chosen to be optimum, where addition of one more cluster does not lower the value of total WCSS appreciably. The Elbow method looks at the total WCSS as a function of the number of clusters.

Fig-1.11 Elbow Curve

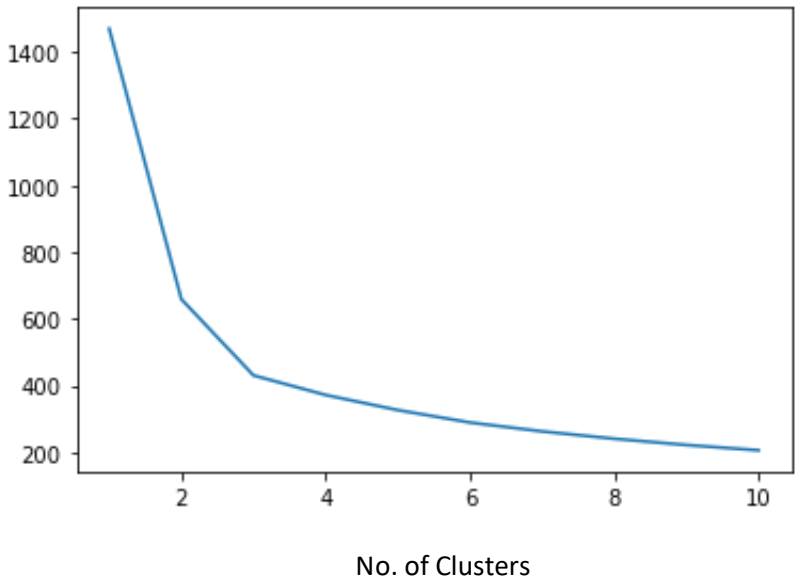


Fig 1.11 does not indicate any clear break in the elbow after k=3. Hence one option for optimum number of clusters is 3. This will be clearer in subsequent analysis of model performance.

Silhouette Method

This method measures how tightly the observations are clustered and the average distance between clusters. For each observation a silhouette score is

constructed which is a function of the average distance between the point and all other points in the cluster to which it belongs, and the distance between the point and all other points in all other clusters, that it does not belong to. The maximum value of the statistic indicates the optimum value of k.

Silhouette score

0.40072705527512986

Minimum Sil width

0.002713089347678376

As the minimum sil width is positive we can conclude the clusters are seperable and the model works fine.

Table 1.4 Sample cluster data

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

By using k-Means technique the dataset is grouped into 3 clusters.

Cluster 0 contains 71 user details

Cluster 1 contains 67 user details

Cluster 2 contains 72 user details

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

Table 1.5 Summary of cluster 0

	count	mean	std	min	25%	50%	75%	max
spending	71.0	14.437887	1.056513	12.080000	13.820000	14.430000	15.260000	16.440000
advance_payments	71.0	14.337746	0.525706	13.150000	14.030000	14.390000	14.760000	15.270000
probability_of_full_payment	71.0	0.881597	0.015502	0.852700	0.871300	0.881900	0.893350	0.918300
current_balance	71.0	5.514577	0.225266	4.984000	5.380000	5.541000	5.689500	5.920000
credit_limit	71.0	3.259225	0.154766	2.936000	3.155000	3.258000	3.378000	3.582000
min_payment_amt	71.0	2.707341	1.176440	0.765100	1.951000	2.640000	3.332000	6.685000
max_spent_in_single_shopping	71.0	5.120803	0.269558	4.605000	4.958500	5.132000	5.263500	5.879000
Clus_kmeans	71.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sil_width	71.0	0.339816	0.165898	0.005457	0.234095	0.371077	0.479615	0.554103

The Average amount spent by users belonging to cluster 0 is 14.43(in 1000s).The cluster 0 users makes an advance payment of 14.33(in 100s).The probability of making full payment is 0.88 on an average.They have a current balance of 5.51(in 1000s) on average.The average credit limit is 3.25(in 10000s).The average minimum payment amount made is 2.70(in 100s).On average they spend 5.12(in 1000s) maximum in single shopping.

Table 1.6 Summary of cluster 1

	count	mean	std	min	25%	50%	75%	max
spending	67.0	18.495373	1.277122	15.560000	17.590000	18.750000	19.14500	21.180000
advance_payments	67.0	16.203433	0.546439	14.890000	15.855000	16.230000	16.58000	17.250000
probability_of_full_payment	67.0	0.884210	0.014917	0.845200	0.874650	0.882900	0.89805	0.910800
current_balance	67.0	6.175687	0.237807	5.718000	6.011500	6.153000	6.32800	6.675000
credit_limit	67.0	3.697537	0.166014	3.387000	3.564500	3.719000	3.80800	4.033000
min_payment_amt	67.0	3.632373	1.211052	1.472000	2.848000	3.619000	4.42100	6.682000
max_spent_in_single_shopping	67.0	6.041701	0.229566	5.484000	5.879000	6.009000	6.19250	6.550000
Clus_kmeans	67.0	1.000000	0.000000	1.000000	1.000000	1.000000	1.00000	1.000000
Sil_width	67.0	0.468772	0.153712	0.029792	0.419827	0.523482	0.57434	0.639285

The Average amount spent by users belonging to cluster 1 is 18.49(in 1000s).The cluster 1 users makes an advance payment of 16.20(in 100s).The probability of making full payment is 0.88 on an average.They have a current balance of 6.17(in 1000s) on average.The average credit limit is 3.69(in 10000s).The average minimum payment amount made is 3.63(in 100s).On average they spend 6.04(in 1000s) maximum in single shopping.

Table 1.7 Summary of cluster 2

	count	mean	std	min	25%	50%	75%	max
spending	72.0	11.856944	0.714801	10.590000	11.255000	11.825000	12.395000	13.340000
advance_payments	72.0	13.247778	0.355208	12.410000	12.992500	13.250000	13.482500	13.950000
probability_of_full_payment	72.0	0.848253	0.019953	0.808100	0.835000	0.848600	0.861475	0.888300
current_balance	72.0	5.231750	0.141795	4.899000	5.139250	5.225000	5.337250	5.541000
credit_limit	72.0	2.849542	0.138689	2.630000	2.738500	2.836500	2.967000	3.232000
min_payment_amt	72.0	4.742389	1.354711	1.502000	4.032250	4.799000	5.463750	8.456000
max_spent_in_single_shopping	72.0	5.101722	0.184012	4.519000	5.001000	5.089000	5.223500	5.491000
Clus_kmeans	72.0	2.000000	0.000000	2.000000	2.000000	2.000000	2.000000	2.000000
Sil_width	72.0	0.397473	0.159526	0.002713	0.314599	0.453462	0.515146	0.587277

The Average amount spent by users belonging to cluster 2 is 11.85(in 1000s).The cluster 2 users makes an advance payment of 13.24(in 100s).The probability of making full payment is 0.84 on an average.They have a current balance of 5.23(in 1000s) on average.The average credit limit is 2.84(in 10000s).The average minimum payment amount made is 4.74(in 100s).On average they spend 5.10(in 1000s) maximum in single shopping.

Cluster 0 :Medium Spending Group

These customers have shown a good track of maintaining their credit score.They make purchases and their probability of making full payment to the bank is also good.

They can be encouraged to make purchases by offering incentives like cashback offers and other rewards by clubbing with other consumer brands.

We can also increase their credit limit.

Cluster 1 :High Spending Group

They are most potential customers.They Spend higher compared to other clusters.The maximum max spent in single shopping is also higher for this group.

They can be incentivized by offering best deals from luxury consumer brands.Special rewards can be offered in form of loyalty points.

Their credit limit can be further increased to encourage them to make frequent purchases.

Cluster 2 : Low Spending Group

The customers grouped in this segment have shown less probability of making full payments comparatively.The payment due reminders should be made frequently to improve their repayment rate and their credit score.

Rewards can be announced to the customers making repayment on time.

These customers spend less than customers in other clusters.Special deals and discounts can be given in categories like grocery and utilities to make them transact frequent.

Problem 2: CART-RF-ANN

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.

Executive sumary

The management of an Insurance firm providing tour insurance decides to collect data from the past few years as they face higher claim frequency. EDA is performed on the dataset and predictive models are built to provide recommendations to management using CART, RF & ANN .

Introduction

The given dataset contains details about 3000 customer's insurance details.Exploratory data analysis is done. Predictive Models such as CART,RF & ANN are built to predict the claim status.Performance of various models in train & test sets are evaluated.

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

Sample Dataset

Table 2.1-Sample 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

Exploratory Data Analysis

Let us check the type of variables

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

The dataset contains 3000 rows and 10 columns.Out of 10 columns 6 columns are Object type,2 columns are integer type and 2 columns are float type.

Check for missing values in dataset

Age	3000 non-null	int64
Agency_Code	3000 non-null	object
Type	3000 non-null	object
Claimed	3000 non-null	object
Commision	3000 non-null	float64
Channel	3000 non-null	object
Duration	3000 non-null	int64
Sales	3000 non-null	float64
Product Name	3000 non-null	object
Destination	3000 non-null	object

From the above values it is clear that there are no missing values in dataset.

Descriptive statistics

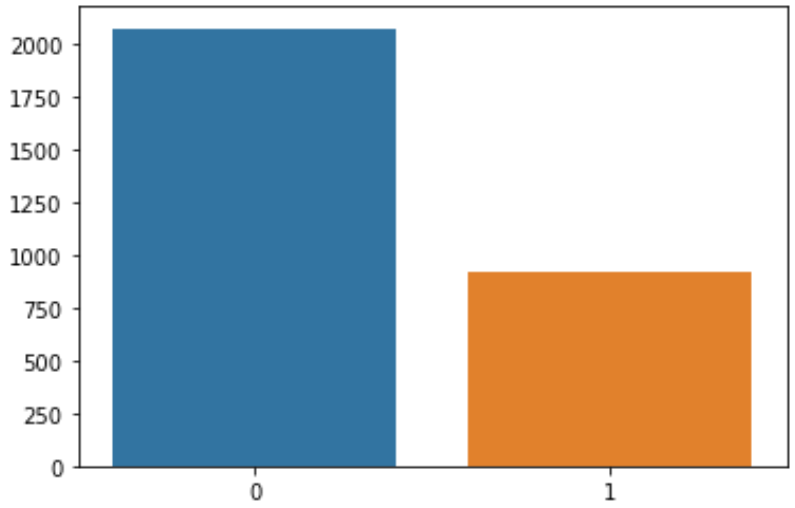
Descriptive statistics are used to describe about the variables in dataset by giving short summaries about the sample and the measures of data. The most recognized types of descriptive statistics are measures of centre**: the mean, median, and mode**, which are used at almost all levels of math and statistics.

Table2.2 -Summary of data

	Age	Type	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
mean	38.091000	0.612333	14.529203	0.984667	70.001333	60.249913	1.661667	0.250000
std	10.463518	0.487299	25.481455	0.122895	134.053313	70.733954	1.258726	0.575277
min	8.000000	0.000000	0.000000	0.000000	-1.000000	0.000000	0.000000	0.000000
25%	32.000000	0.000000	0.000000	1.000000	11.000000	20.000000	1.000000	0.000000
50%	36.000000	1.000000	4.630000	1.000000	26.500000	33.000000	2.000000	0.000000
75%	42.000000	1.000000	17.235000	1.000000	63.000000	69.000000	2.000000	0.000000
max	84.000000	1.000000	210.210000	1.000000	4580.000000	539.000000	4.000000	2.000000

From the above table it is evident that the average age of clients is 38.09.The av erage commission received by firms is 4.63% of sales. The average amount worth of sales per customer in procuring tour insurance policies in rupees (in 100's) is 33.00.

Fig-2.1 Barplot of Claimed

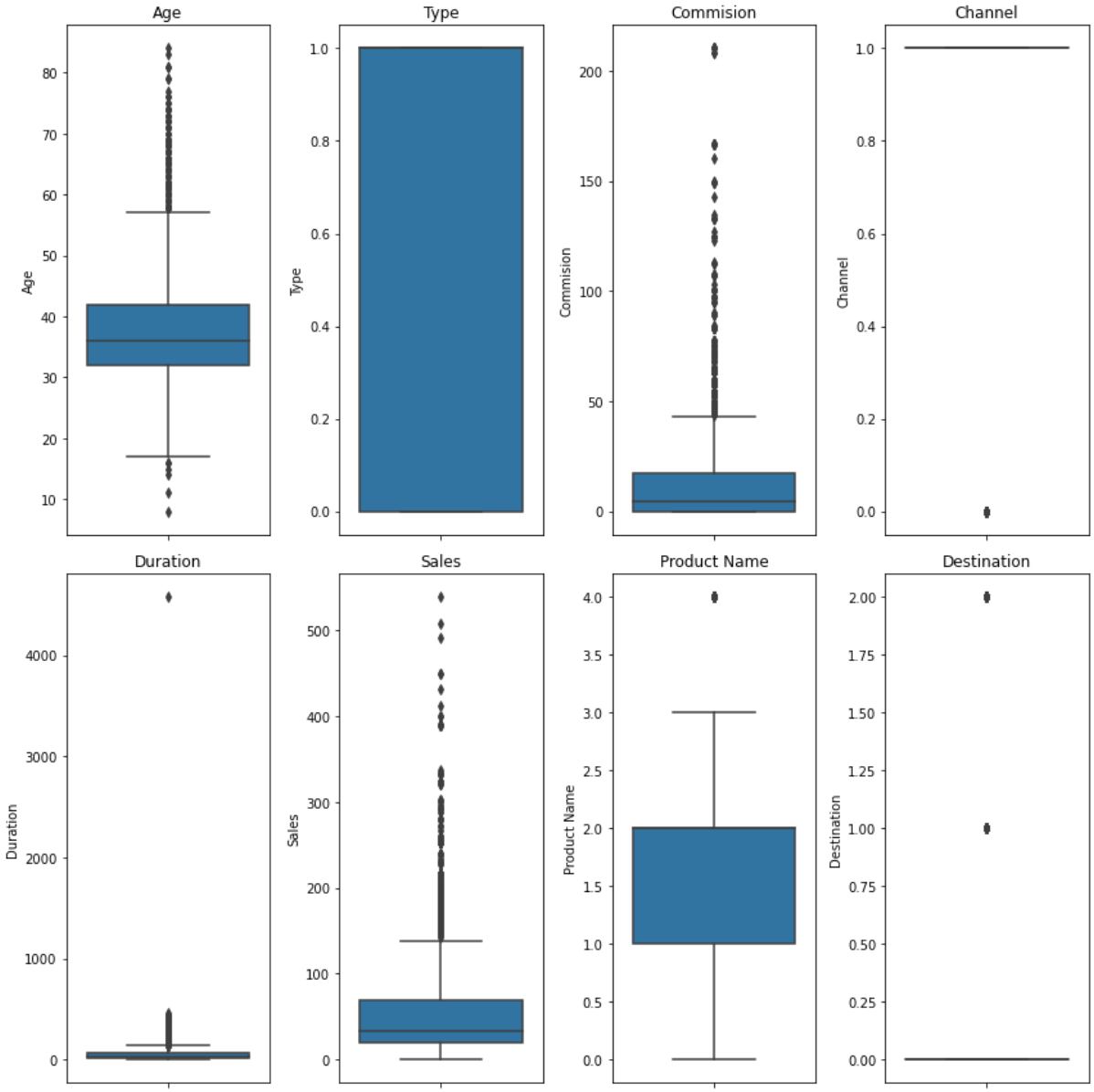


The dataset has no equal representation for the claim status 'Yes' & 'No'. 69.2% of dataset represents 'NO' & 30.8% of dataset represents 'YES'.

Univariate Analysis

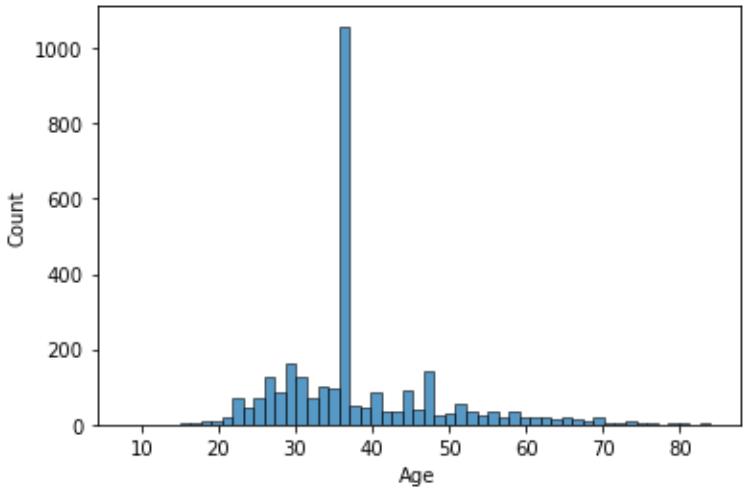
Check for Outliers

Fig -2.2 Boxplot of variables



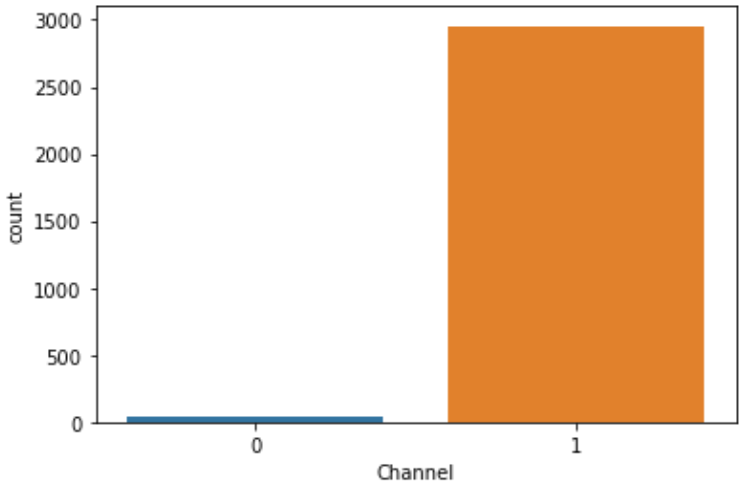
From boxplot of variables we find there are many outliers in the dataset.

Fig-2.3Histogram of Age



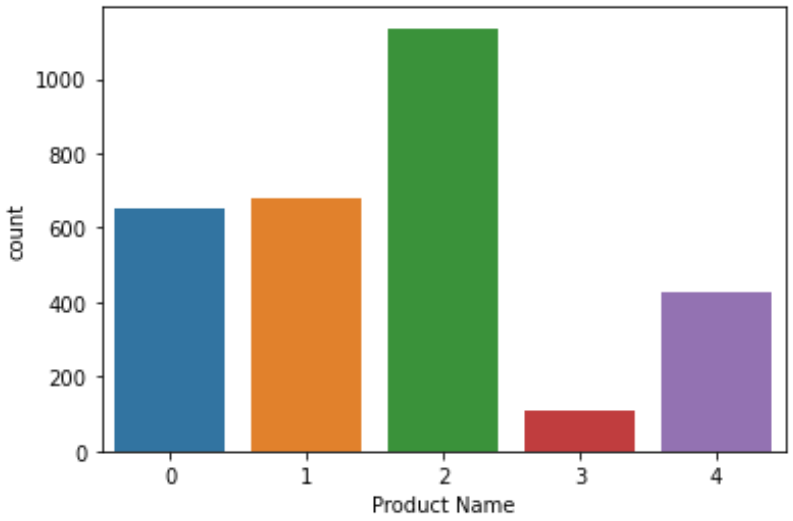
The distribution of age seems to positive skewed.The minimum age is 8 and maximum age is 84.

Fig-2.4 Count plot of Channel



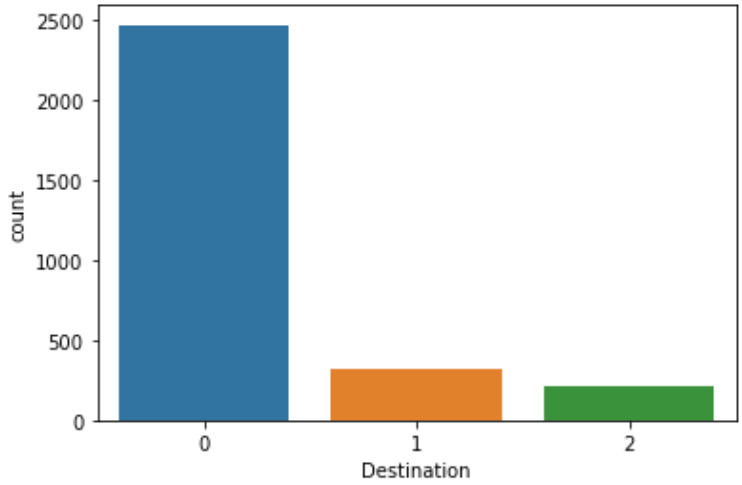
From the plot we find that the most preferred channel is online-(1) when compared to offline(0).

Fig-2.5 Count plot of Product Name



The most preferred product is customised plan(2) and the least preferred product is Gold Plan(3).

Fig-2.6 Countplot of Destination

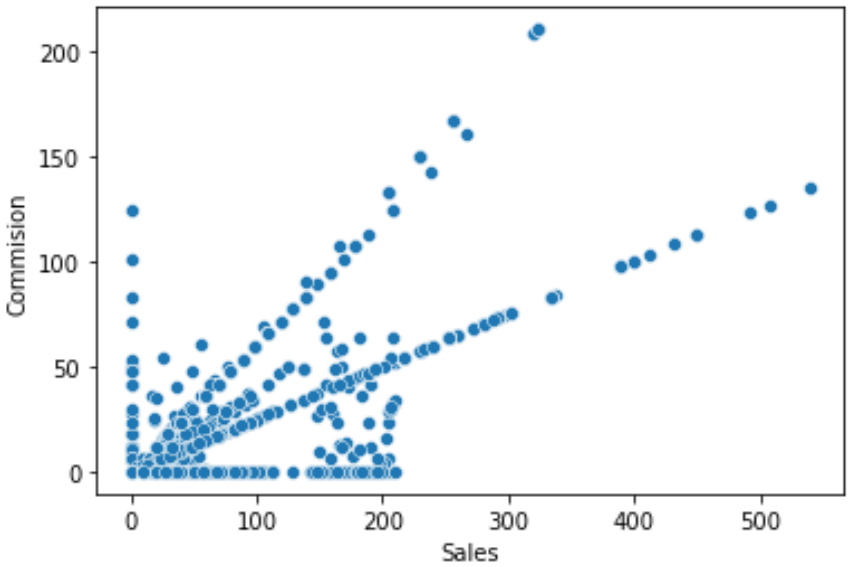


The most preferred destination is Asia(0) and the least preferred destination is Europe(2).

Bi-variate Analysis

Fig -2.7 Scatter plot

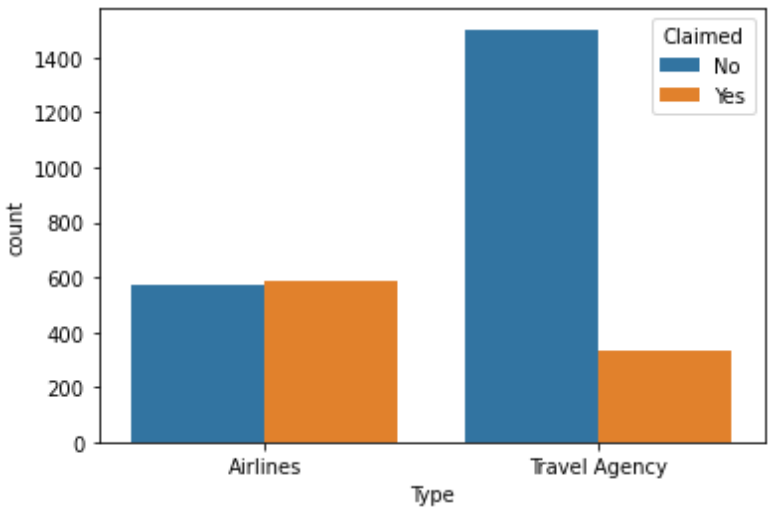
Comission vs Sales



The scatter plot shows that there exists a linear relationship between the commission and sales.

Fig -2.8 Count plot

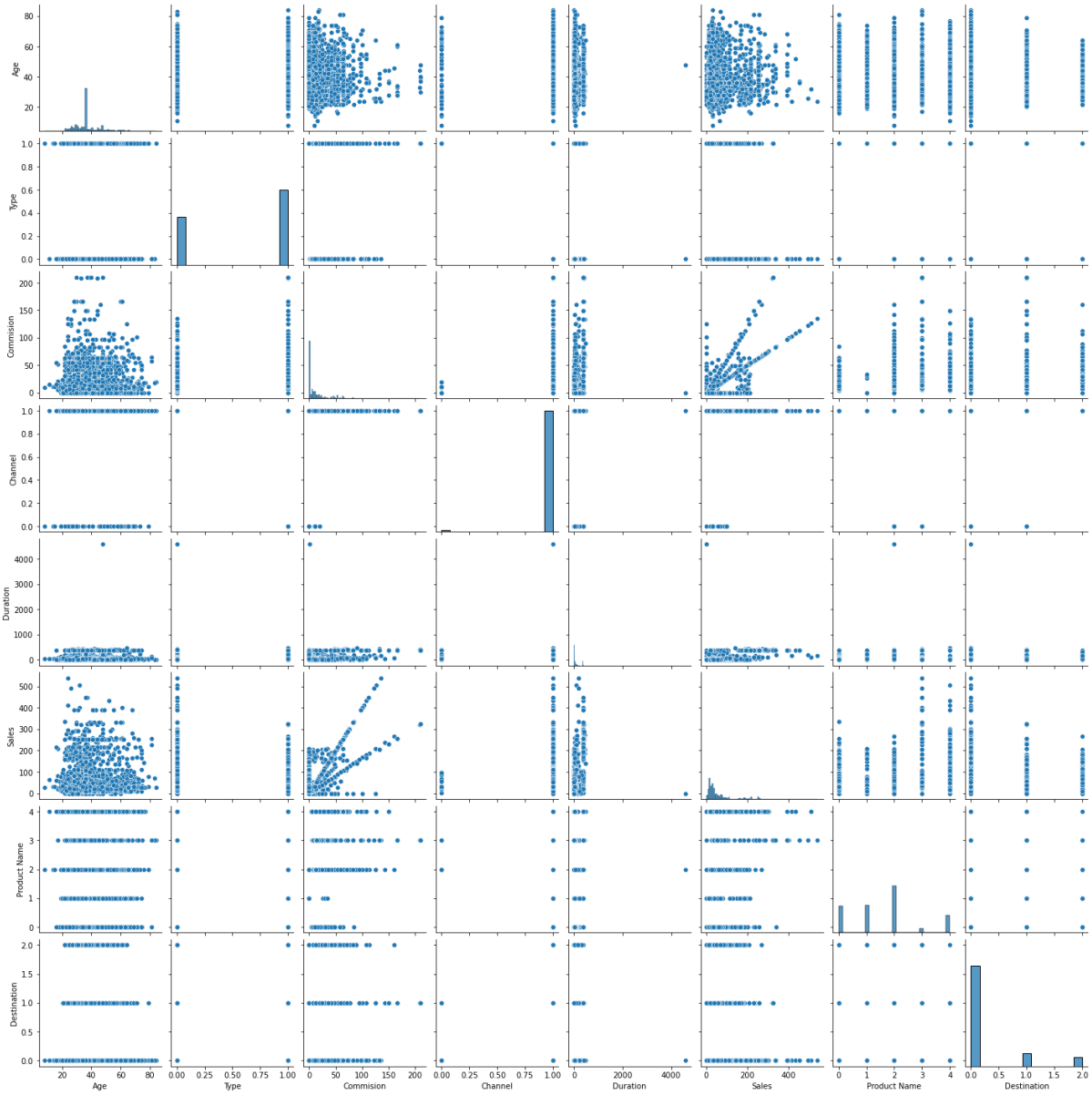
Type vs claim status.



The count plot shows most preferred insurance firms are Travel Agencies.

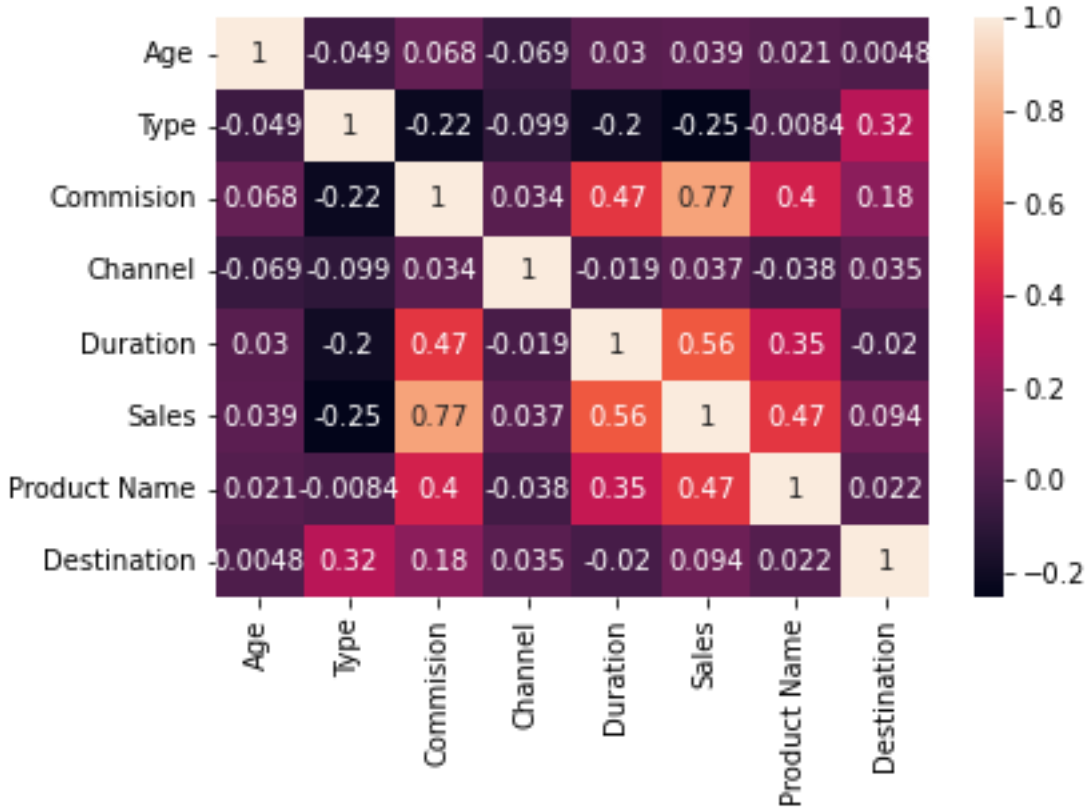
Multivariate Analysis

Fig -2.9 Pair plot of Variables



Checking the pairwise distribution of the continuous variables.

Fig -2.10 Heatmap of Variables



The heatmap is plotted to find the correlation between the variables.The variables such as sales and commission,sales and duration shows significant correlation.

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

Train -Test Split

The dataset is divided into training set and test set.70% of dataset is taken as training dataset and 30% is taken as test set.The variable 'Claimed' is dependent variable and all other variables are independent variables.

Table-2.3 Sample of Independent Train data set

	Age	Type	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0.70	1	7	2.51	2	0
1	36	1	0.00	1	34	20.00	2	0
2	39	1	5.94	1	3	9.90	2	1
3	36	1	0.00	1	4	26.00	1	0
4	33	0	6.30	1	53	18.00	0	0

Classification Models

The following classification models are built to predict the dependent variable.

1.CART Model

2.Random Forest

3.Artificial Neural Network.

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.

Model Performance Measure

a. Accuracy – How accurately / cleanly does the model classify the data points. Lesser the false predictions, more the accuracy

b. Sensitivity / Recall – How many of the actual True data points are identified as True data points by the model . Remember, False Negatives are those data points which should have been identified as True.

c. Specificity – How many of the actual Negative data points are identified as negative by the model

d. Precision – Among the points identified as Positive by the model, how many are really Positive

Confusion Matrix – A 2X2 tabular structure reflecting the performance of the model in four blocks

1.CART Model

Accuracy

Training set – 0.78

Test set - 0.76

Confusion Matrix of Training set

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	TN =1265	FP=206
Actual Positive	FN=260	TP=369

Confusion Matrix of Test set

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	TN =538	FP=67
Actual Positive	FN=148	TP=147

ROC -Curve

Fig-2.11 ROC- Curve (Training set)

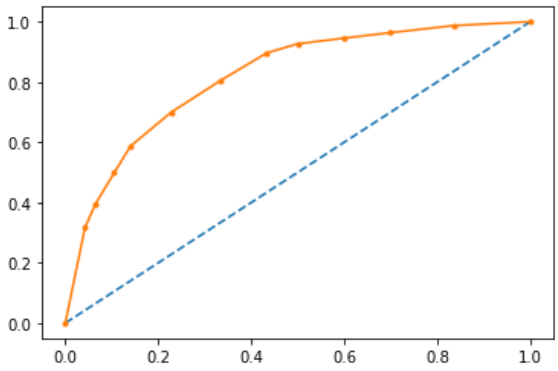
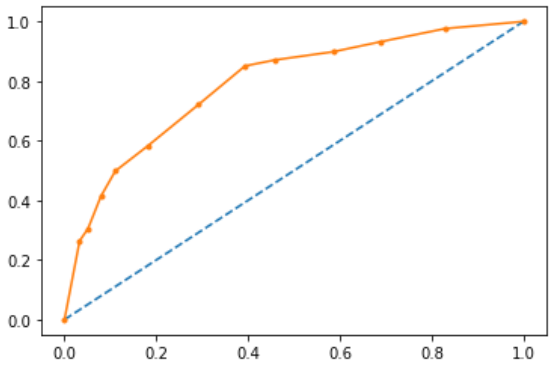


Fig-2.12 ROC-Curve of Test set



ROC_AUC score

Training set - 0.808

Test set - 0.790

CART -Model

Classification report for Training set

Table -2.4 Classification report for Training set

	precision	recall	f1-score	support
0	0.83	0.86	0.84	1471
1	0.64	0.59	0.61	629
accuracy			0.78	2100
macro avg	0.74	0.72	0.73	2100
weighted avg	0.77	0.78	0.78	2100

Classification report for Testing set

Table -2.5 Classification report for Test set

	precision	recall	f1-score	support
0	0.78	0.89	0.83	605
1	0.69	0.50	0.58	295
accuracy			0.76	900
macro avg	0.74	0.69	0.71	900
weighted avg	0.75	0.76	0.75	900

2.RANDOM FOREST

Accuracy

Training set – 0.78

Test set - 0.75

Confusion Matrix of Training set

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	TN =1352	FP=119
Actual Positive	FN=337	TP=292

Confusion Matrix of Test set

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	TN =565	FP=40
Actual Positive	FN=184	TP=111

ROC -Curve

Fig-2.13 ROC- Curve (Training set)

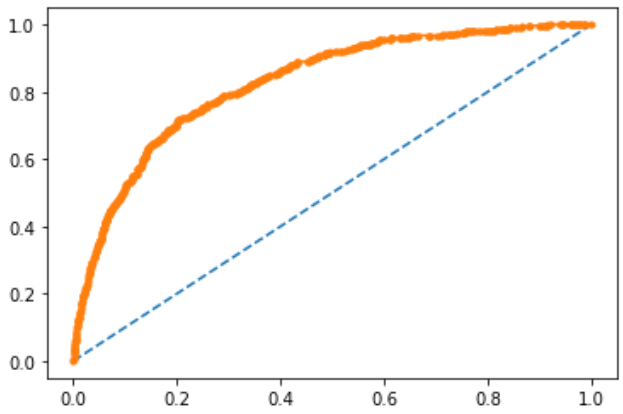
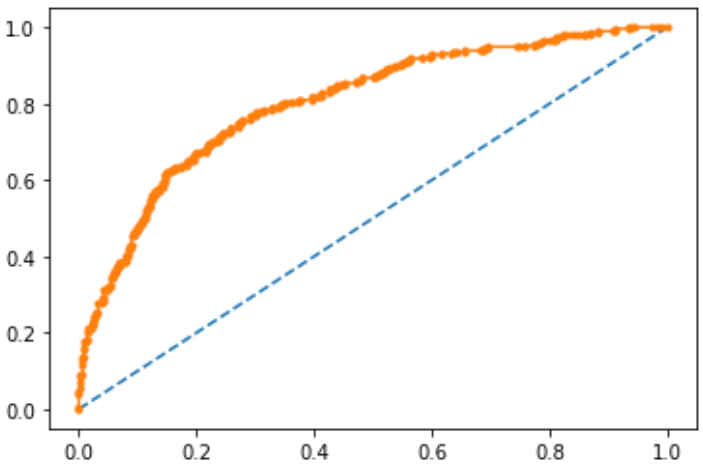


Fig-2.14 ROC-Curve of Test set



ROC_AUC score

Training set – 0.828

Test set - 0.790

RANDOM FOREST

Table -2.6 Classification report for Training set

	precision	recall	f1-score	support
0	0.80	0.92	0.86	1471
1	0.71	0.46	0.56	629
accuracy			0.78	2100
macro avg	0.76	0.69	0.71	2100
weighted avg	0.77	0.78	0.77	2100

Table -2.7 Classification report for Test set

	precision	recall	f1-score	support
0	0.75	0.93	0.83	605
1	0.74	0.38	0.50	295
accuracy			0.75	900
macro avg	0.74	0.66	0.67	900
weighted avg	0.75	0.75	0.72	900

3.ARTIFICIAL NEURAL NETWORK

Accuracy

Training set –0.78

Test set - 0.74

Confusion Matrix of Training set

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	TN =1350	FP=121
Actual Positive	FN=332	TP=297

Confusion Matrix of Test set

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	TN =559	FP=46
Actual Positive	FN=184	TP=111

ROC -Curve

Fig-2.15 ROC- Curve (Training set)

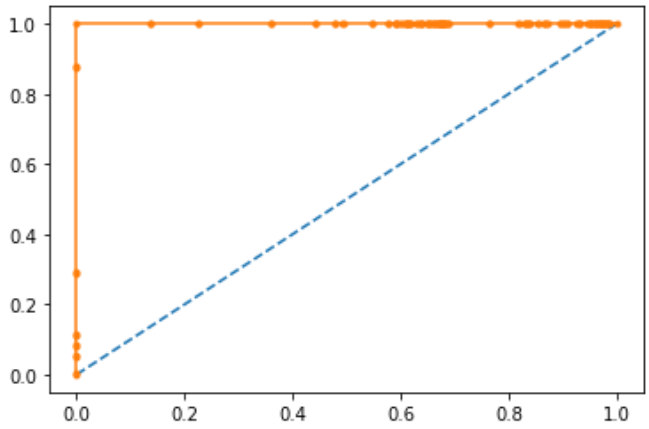
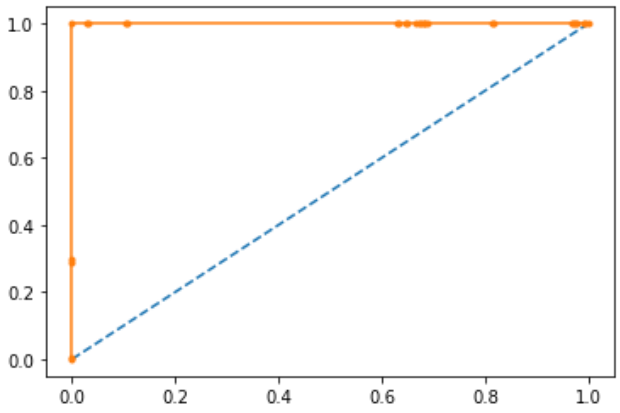


Fig-2.16 ROC-Curve of Test set



ROC_AUC score

Training set - 1.00

Test set - 1.00

ARTIFICIAL NEURAL NETWORK

Table -2.8 Classification report for Training set

	precision	recall	f1-score	support
0	0.80	0.92	0.86	1471
1	0.71	0.47	0.57	629
accuracy			0.78	2100
macro avg	0.76	0.69	0.71	2100
weighted avg	0.78	0.78	0.77	2100

Table -2.9 Classification report for Test set

	precision	recall	f1-score	support
0	0.75	0.92	0.83	605
1	0.71	0.38	0.49	295
accuracy			0.74	900
macro avg	0.73	0.65	0.66	900
weighted avg	0.74	0.74	0.72	900

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

PERFORMANCE METRICS	MODELS					
		CART		RF		ANN
AUC	TRAIN	0.808		0.828		1
	TEST	0.79		0.79		1
PRECISION	TRAIN	0.64		0.71		0.71
	TEST	0.69		0.74		0.71
RECALL	TRAIN	0.59		0.46		0.47
	TEST	0.5		0.38		0.38
ACCURACY	TRAIN	0.78		0.78		0.78
	TEST	0.76		0.75		0.74

With a **recall of 0.50** for test data and **accuracy of 0.76** the **CART Model** is considered as the best model .Here the metric recall is important considering the model has failed to predict **148(FN)** customers who claimed insurance so major focus can be upon improving the recall score which can provide some insights for the firm to take proactive steps in analysing those customers who might claim and take necessary steps

AUC on the training data is **80%** and on test data is **79%.**

The Overall model performance is moderate enough to start predicting if any new customer will claim insurance or not.

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

From the data we can conclude that 98% of the customers purchased their insurance plans via online medium, making it the most preferred sales channel.To increase the purchase rate further the user experience can be improved.

Most of the customers prefer to choose custom insurance plans thus by providing more flexibility in choosing the plans the sales can be further improved**.**

The offline sales can be improved by training the agents with necessary pitching techniques and sales strategies.Promotional campaigns can be run.

Compared to Airlines more sales happens at Agency on contrast most claims are processed at Airline.The reason for higher claims at airlines must be further investigated.

Efficiency of Processing claims can be improved.Fraud detection methods can be employed.