DATA MINING PROJECT

Submitted by Areti Vinay

FEBRUARY 27, 2022

GREAT LEARNING

Contents

Problem-1 Clustering

Introduction:

A leading bank wants to profile its customers to give promotional activities. They have collected a sample of Data that summarizes the credit card usage activities of its customers over last few months.

Objective:

Objective of this study is to label bank customers based on their credit card usage. Unsupervised learning model like Hierarchal clustering and K-means techniques will be applied to profile Customers.

Exploratory Data Analysis:

Data dictionary:

The following are Attributes in the data:

- Spending – Amount spent by customer per each month (1000s)
- Advance_payments- Amount paid by the customer in advance by cash (in 100s)

- probability_of_full_payment: Probability of payment done in full by the customer to the bank
- current_balance: Balance amount left in the account to make purchases (in 1000s)
- credit_limit: Limit of the amount in credit card (10000s)
- min_payment_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)
- max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s)

Data Dimension:

There are 7 attributes and 210 observations in the sample data frame.

Data Information:

The following output provides information about each attribute data type and total number of Non-null values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

Above results indicate that there are total of 7 numerical variables and total of 210 non-null values. As non-values count is equal to total observations, indicates that there are no Missing Values in the data. However the missing values will be checked using isna() function in the next step.

Data head:

The following output depicts the first five observation of the data

	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

Checking for Missing Values:

Following is the output generated while checking for missing values in the sample data.

```
spending                      0
advance_payments              0
probability_of_full_payment   0
current_balance               0
credit_limit                  0
min_payment_amt               0
max_spent_in_single_shopping  0
dtype: int64
```

0 corresponding to all the attributes confirm that there no missing values in the data.

Checking for duplicate values:

The duplicate values if any in the data frame are checked using duplicate function in python, produced an output FALSE which indicate that there are no duplicate observations in the data frame.

Univariate analysis:

Following tables and figures depict the respective descriptive statistics and distribution of each attribute. Describe function is used to produce descriptive stats of each attribute and 1) Box plot 2) Histogram are used to visualize distribution of each attribute.

1)Spending

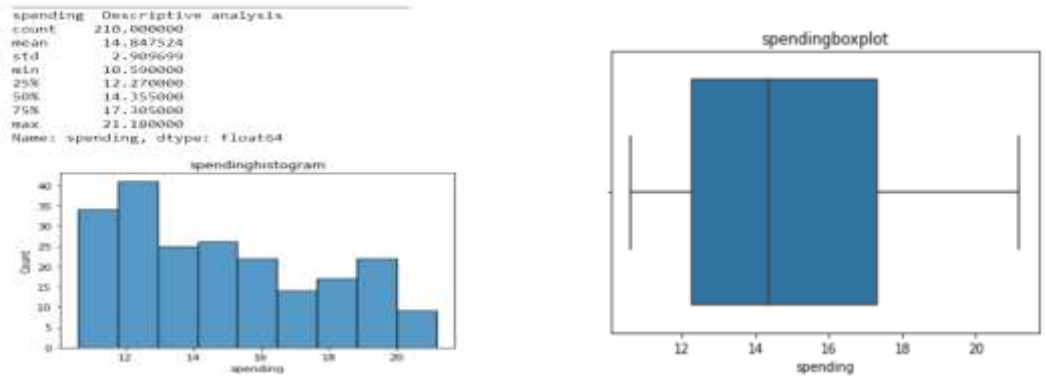


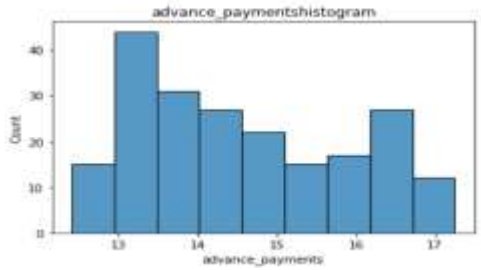
Figure 1- Spending analysis

Observations:

- The customers spend 14847 on an average per month and 50% percent of customers spend less than 14355, with minimum and maximum spending ranging from 10590 to 21180 respectively.
- Median is less than mean which means that data is slightly right skewed.
- Boxplot confirms that there are no outliers in the data.

2)advance payments

```
advance_payments  Descriptive analysis
count    210.000000
mean      14.559286
std        1.305959
min       12.410000
25%       13.450000
50%       14.320000
75%       15.715000
max       17.250000
Name: advance_payments, dtype: float64
```



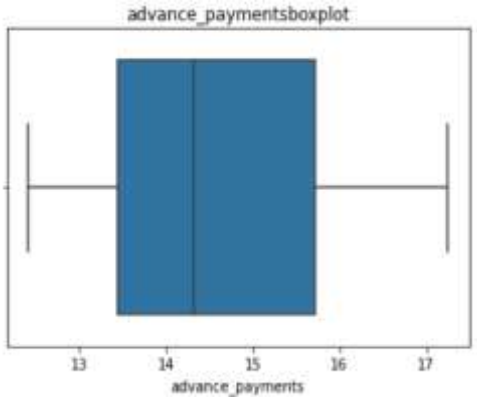
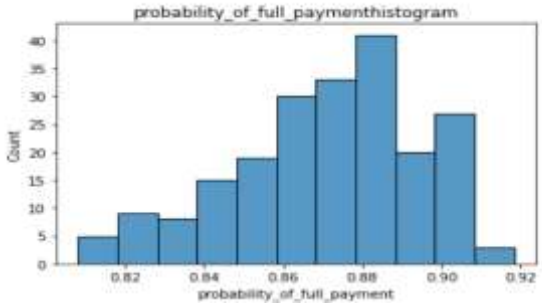


Figure 2-advance payments analysis

- Customers make around 1455 on average as advance payments with 50% of customers making less than 1432. Minimum and maximum ranging from 1241 to 1725 respectively
- Median is slightly less than Mean indicating slight right skewness.
- Box plot confirms there are outliers in the data.

3) probability of full payment

```
probability_of_full_payment  Descriptive analysis
count    210.000000
mean       0.870999
std        0.023629
min        0.808100
25%        0.856900
50%        0.873450
75%        0.887775
max        0.918300
Name: probability_of_full_payment, dtype: float64
```



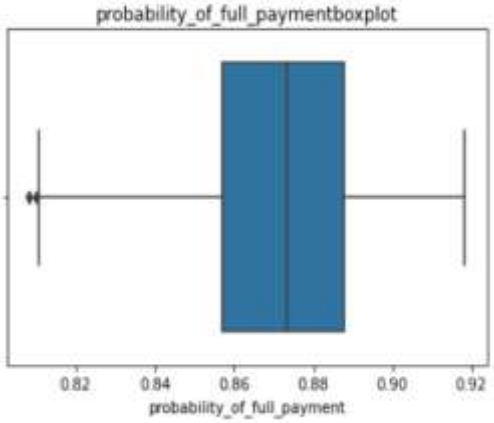


Figure 3-probability of full payment analysis

- average probability of full payment is 0.87 and 50% of customers make full payments less than 87.3% of times

- Median is greater than mean, indicating left skewness in the distribution.
- There are outliers present as per boxplot

4)current balance

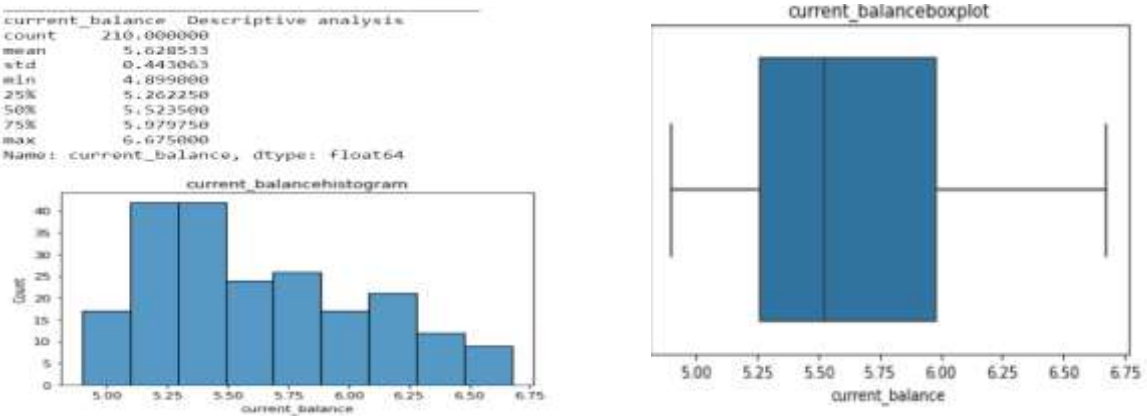


Figure 4-current balance analysis

- Average current balance available is 5628, with 50% of customers having current balance less than 5523
- Box plot confirms absence of outliers.
- Median is less than mean indicating presence of right skewness in the distribution.

5)credit limit

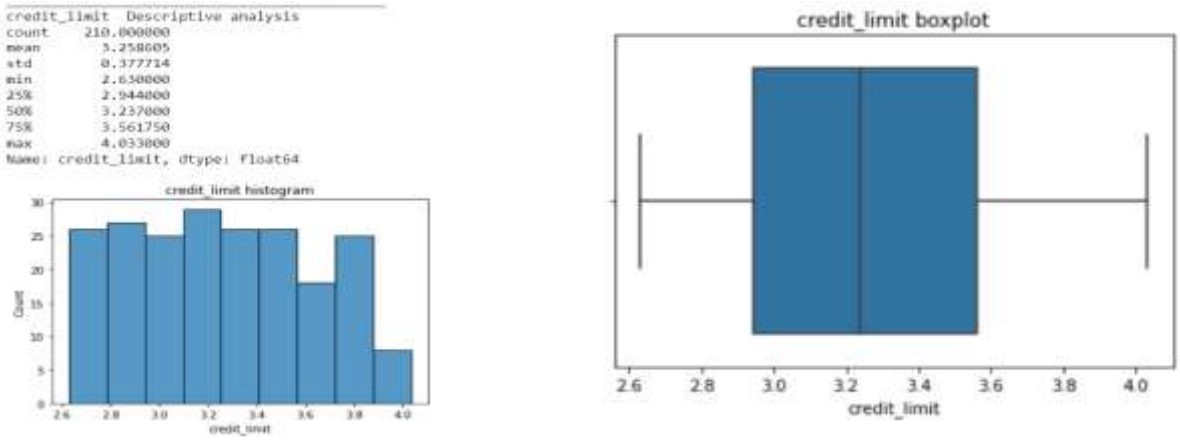


Figure 5- credit limit analysis

- On an average available credit limit is 32586, with 50% of customers having credit limit less than 32370, Mean is slightly higher than median indicating slight right skewness.
- Boxplot confirms absence of outliers.

6)min payment amount

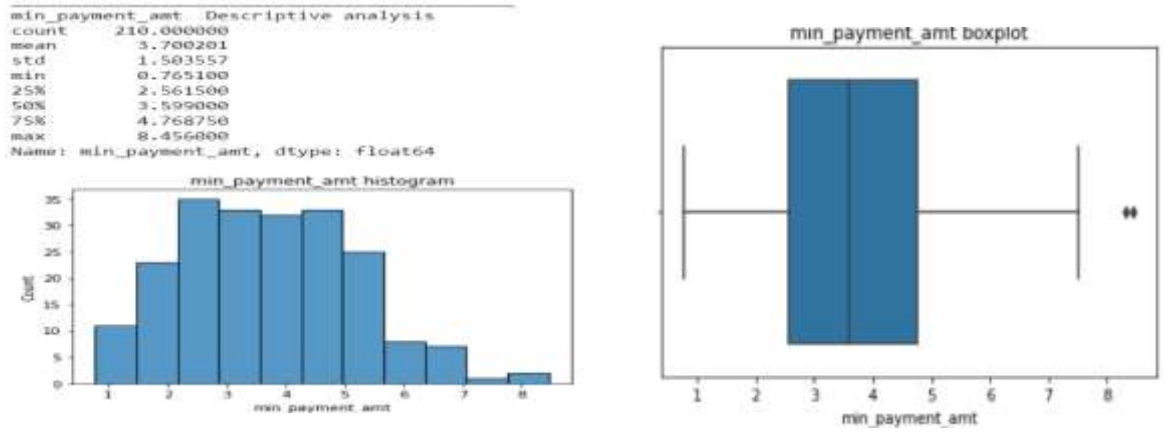


Figure 6-min payment analysis

- Minimum payment paid on average is 370 and with 50% of customers making minimum payment less than 359.9. Mean is higher than median indicating a right skewness. Minimum value as less as 76 and maximum value 845
- From boxplot is evident that there are outliers in the variable (approximately 2)

7)max spent in single shopping

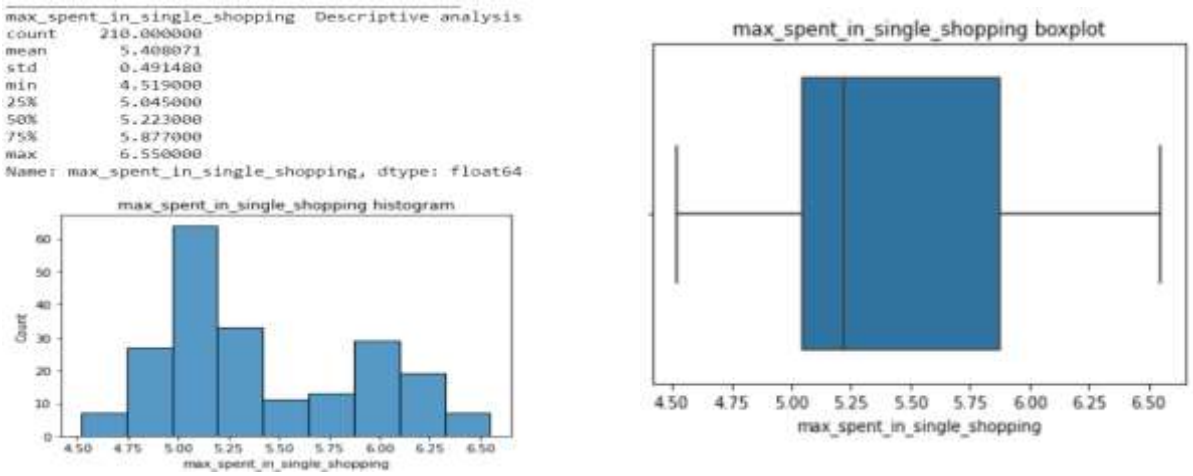


Figure 7 max spent in single shopping

- On an average max spent in single shopping is 5408 and 50% of customers spent higher than 5223. Mean is slightly higher than Median indicating right skewness.
- Box plot confirms the absence of outliers in the data.

Multivariate analysis:

As all the variables are on a continuous scale, Pair plot will help us understand the linear relationship between each variable.

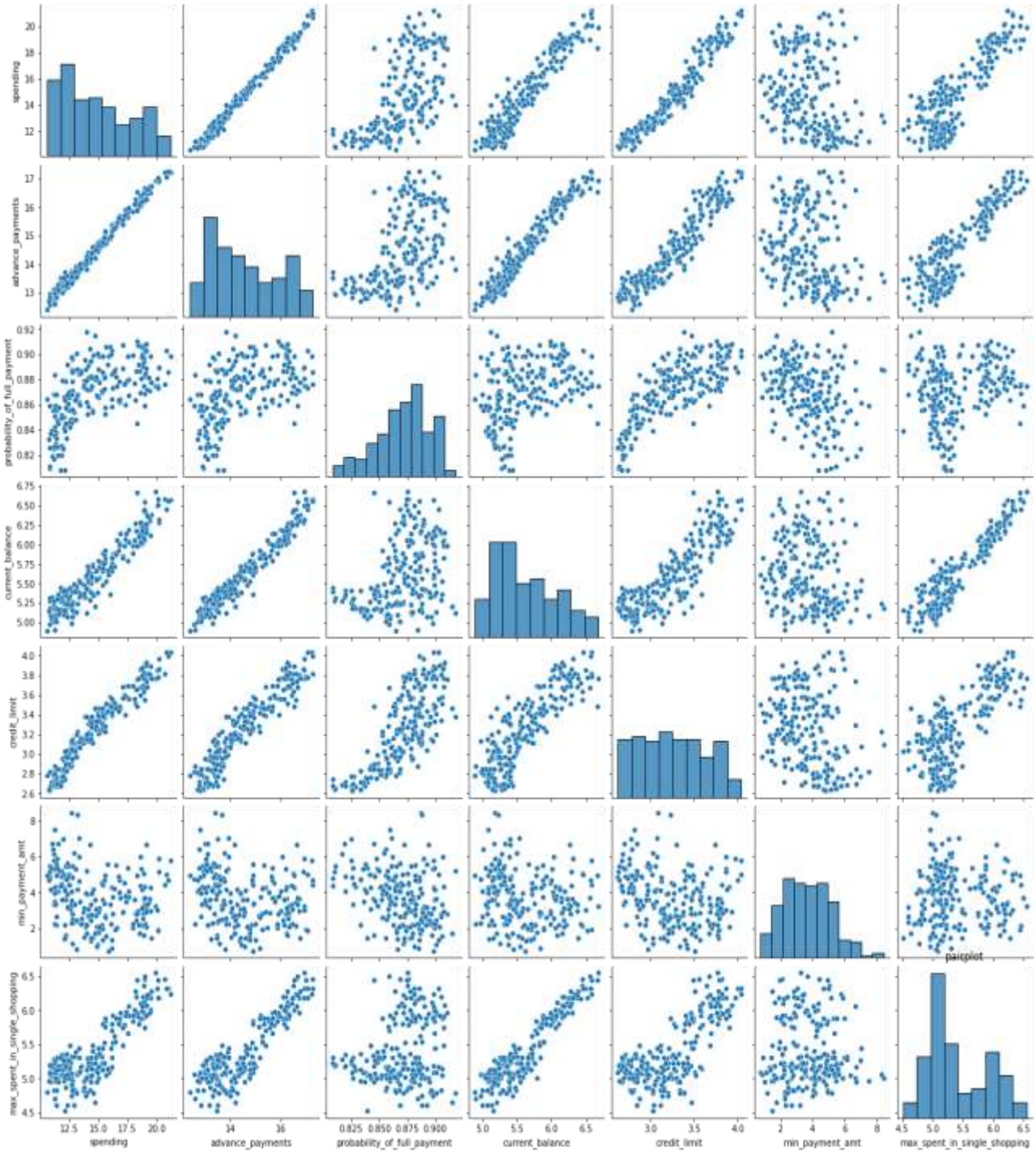


Figure 8-Pair plot

- Above pair plot depicts individual histogram for each variable and its linear relationship to other variable with the help of scatterplot.

Inference from pair plot:

- From the above graph its evident that attribute spending has a strong linear movement with attribute advance payments.
- Current balance has strong linear movement with attribute spending and advance payments.
- Attribute credit limit has strong linear movement with attribute spending and advance payments and current balance.
- Max spent in single shopping has strong linear movement with spending, advance payments, current balance, credit limit.

Note: Pair plot visualizes on the linear relationship between continuous variables, however does not provide the magnitude of the relationship between the continuous variables. Correlation plot will help us identify the magnitude of linear relationship between continuous variables.

Correlation plot:

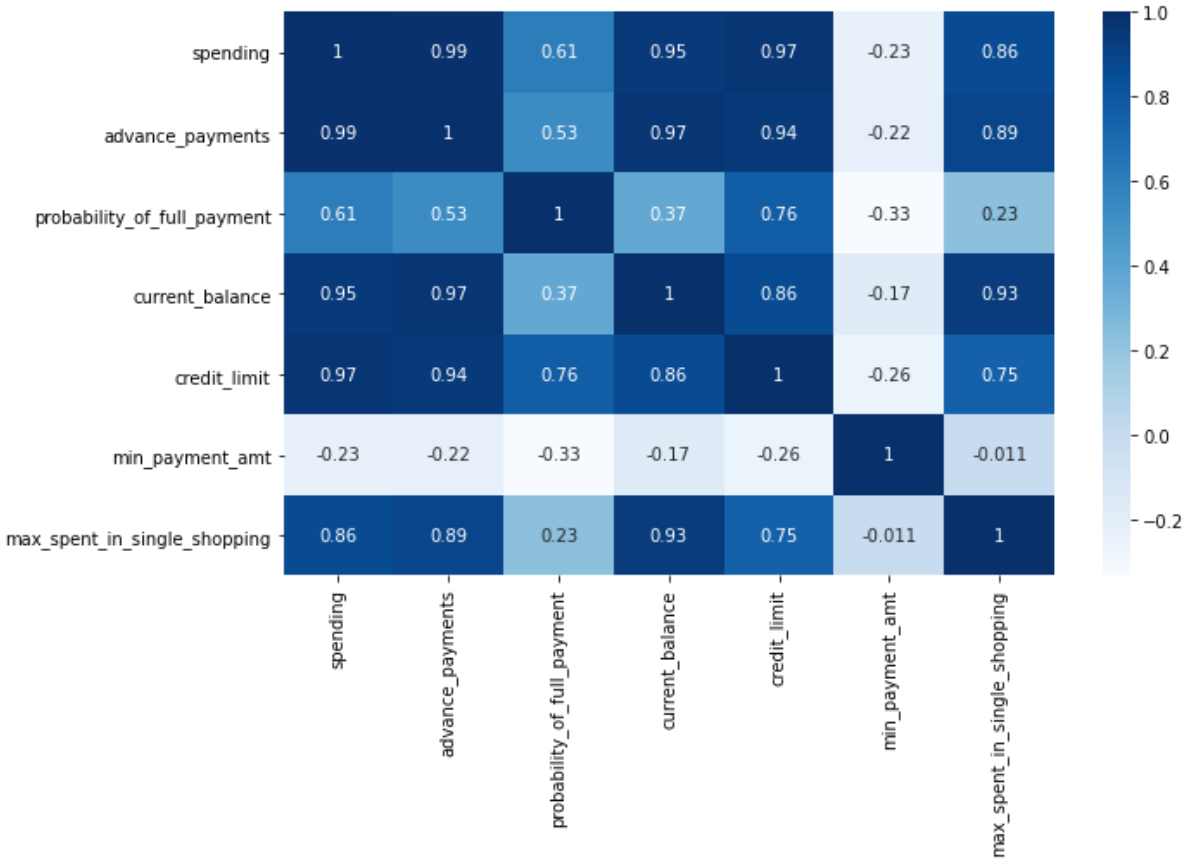


Figure 9- correlation plot

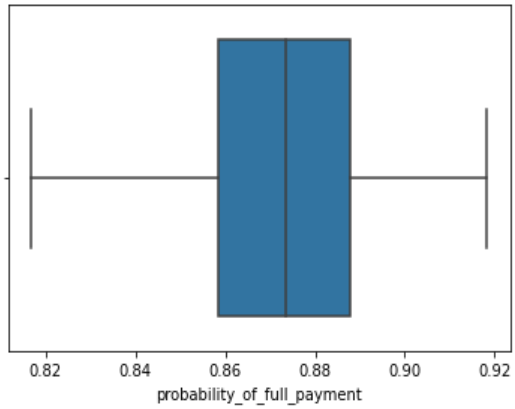
Inference from correlation plot:(Note correlation value ranges from -1 to +1, any value irrespective of sign greater than .8 is considered strong correlation)

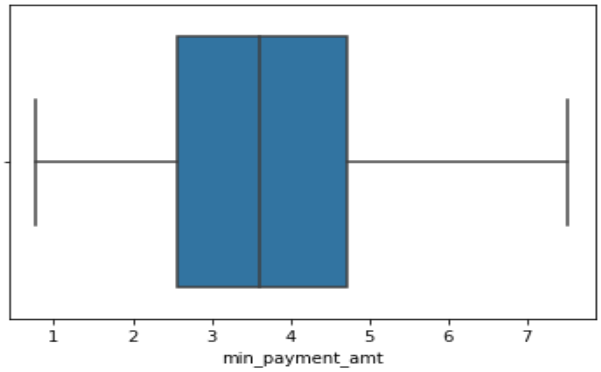
- Advance payments and spending has highest correlation of 0.99

- Current balance and spending, current balance and advance payments has correlation of 0.95 and 0.97 respectively.
- Credit limit and spending, credit limit and advance payments, credit limit and current balance has a correlation of 0.97, 0.94, 0.86 respectively.
- Max spent in single shopping has strong correlation with spending advance payment and current balance with values of 0.86, 0.89, 0.93 respectively.

Removing Outliers:

As confirmed from Box plots, attributes probability of full payment and Min payment has outliers present. As Clustering is sensitive to outliers, they need to be imputed with Median or Mean. We have gone ahead with median to replace the outliers. Following are the boxplot for respective attributes after removing the outliers.





When compared to Figure3 and Figure6 we can observe that there are no observations outside the 1.5 times the inter quartile range. Which indicates that the outliers have been successfully replaced with the attributes median value.

1.2 Scaling

Yes, scaling is necessary for clustering in this case, because all the attributes in the data are not recorded on the same scale. For example, attributes like Min payment are in 100s and probability of full payment takes only values from 0 to 1. Clustering will aim at bringing Homogeneity in group by reducing WSS (within sum of squares). So, in-order to avoid model giving higher weightage to attributes with higher values, scaling should be applied to bring all the attributes to single scale. In this case Z- scale is being preferred which will transform all the attributes to Z-scale values, which will have Mean of 0 and a standard deviation of 1. Following are descriptive statistics of scaled data

Table 1-scaled data

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	-1.771599e-15	-1.089076e-16	-2.994298e-16	1.104936e-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	-2.530009e+00	-1.650501e+00	-1.668209e+00	-2.023649e+00	-1.813288e+00
25%	-8.879552e-01	-8.514330e-01	-6.252806e-01	-8.288816e-01	-8.349072e-01	-7.655544e-01	-7.404953e-01
50%	-1.696741e-01	-1.836639e-01	5.766480e-02	-2.376280e-01	-5.733534e-02	-3.929965e-02	-3.774588e-01
75%	8.465989e-01	8.670693e-01	7.087128e-01	7.945947e-01	8.044956e-01	7.382548e-01	9.563941e-01
max	2.181534e+00	2.065260e+00	2.099666e+00	2.367533e+00	2.055112e+00	2.709892e+00	2.328996e+00

From the above table it can be observed that all the attributes have Mean 0(approximately) and standard deviation of 1(approximately)

1.3Hierarchal clustering.

Following are the dendrograms when applied Euclidean distance with different linkage methods

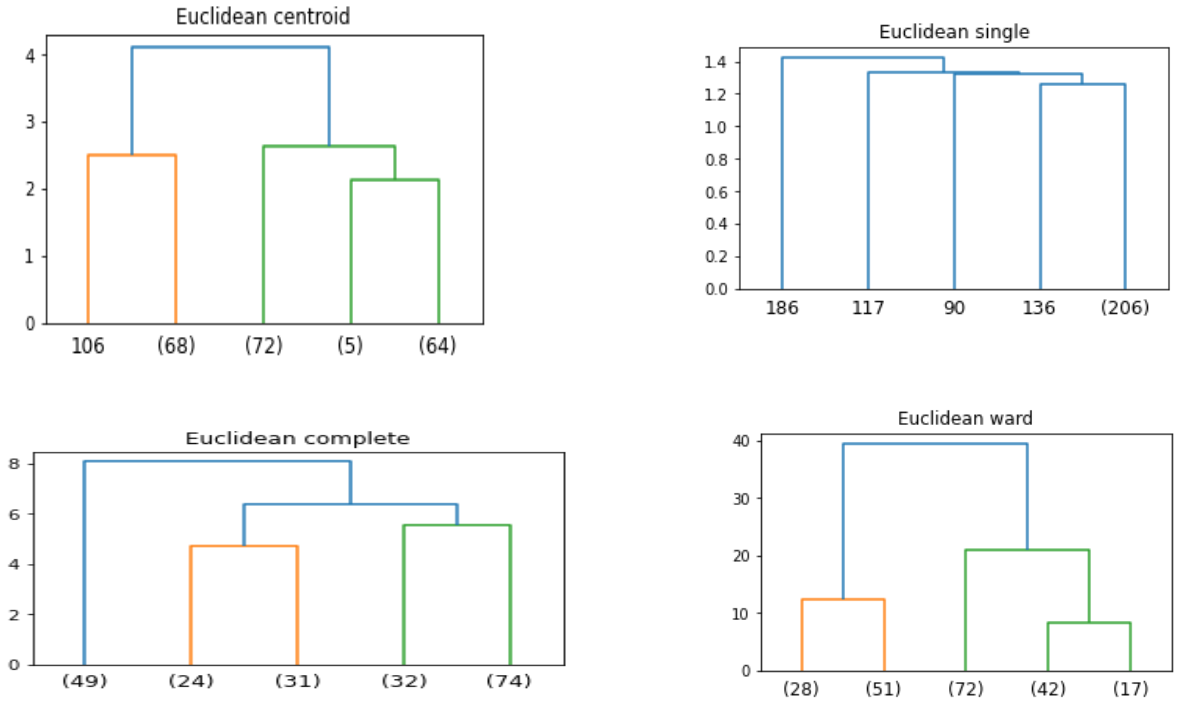


Figure 10- dendrogram

The X-axis represents the clusters in and Y-axis represents the dissimilarity between clusters.

Above dendrograms have been visualised with last 5 clusters, as it can be observed from above dendrograms, most of them are suggesting two clusters. However, as business requirement we need more than 2 clusters therefore, we choose 3 clusters. Wards method will be preferred in clustering the observations as it aims to obtain clusters with minimum WSS error.

Therefore, we choose following methods and metrics for hierarchal clustering

- Euclidean distance
- Wards linkage
- Total clusters= 3

F-cluster is used with linkage ward and maxclust = 3 to obtain the clusters array. Clusters array has been attached to the original data frame.

Following are the proportion of each class:

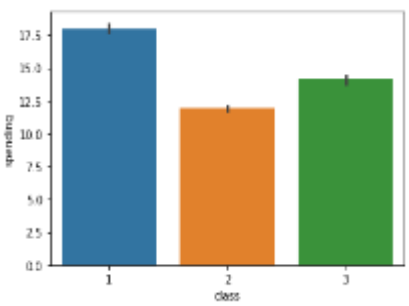
```
1    0.376190
2    0.342857
3    0.280952
Name: class, dtype: float64
```

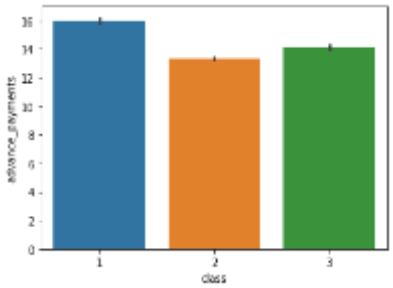
Class 3 has lowest proportion with 28.09% of customers and other classes 1, 2 have 37% and 34% respectively.

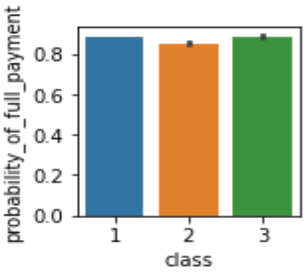
Following are the mean values of each attribute with respect to assigned class:

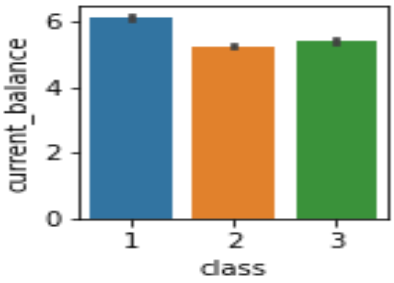
class	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1	18.018861	16.004810	0.881948	6.117266	3.638114	3.595342	5.957772
2	11.960972	13.311111	0.850853	5.259375	2.855917	4.619861	5.104653
3	14.123729	14.146949	0.885276	5.424627	3.241864	2.556054	5.042305

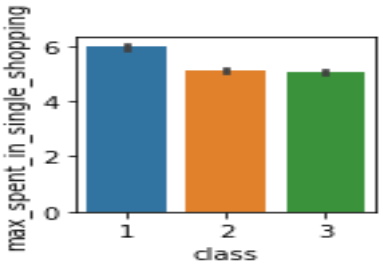
Except probability of full payments all other attributes show significant difference between classes.

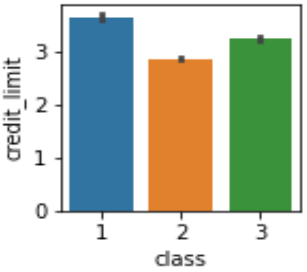




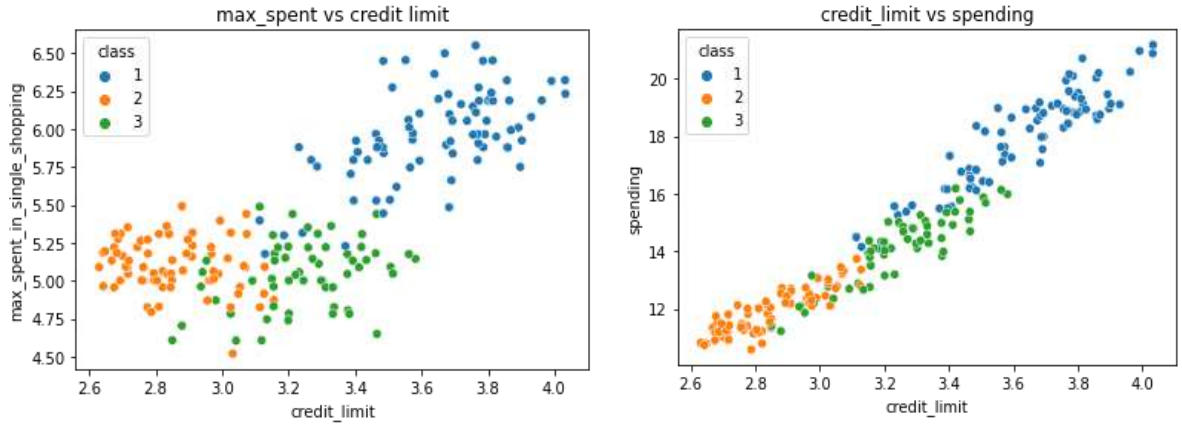








From the above bar charts its evident that higher difference between each class is observed on spending and credit limit and also in max spent in single shopping especially between class 1 and other classes. Further analyse the behaviour of attributes spending, credit limit, max spent in single shopping with the help of scatter plot.



From the above max spent vs credit limit plot it can observed that class 2 and 3 do not show linear relationship, however class 1 exhibit linear relationship between max_spent_in_single_shopping and credit limit. Lets confirm the inference by checking correlation between two variables for class 1 with respect to other two classes.

1.4 K-means Clustering

K-means clustering technique has been applied to scaled data (scaled with Z score) refer to Table 1 for scaled descriptive information. When compared to hierarchal clustering in K means the number of clusters have to be pre-determined. Within sum of squares for each number of clusters will help us understand the optimum number of clusters.

Following plot has been plotted for WSS vs Number of clusters (Elbow curve)

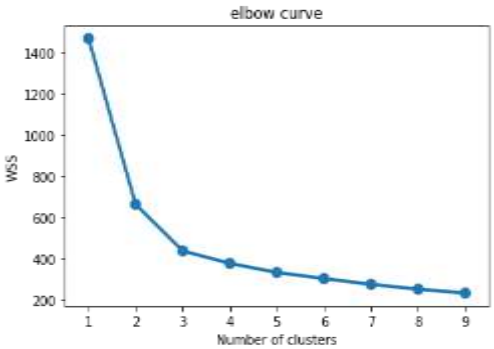


Figure 11- wss plot

From above plot it can be inferred that there is a significant fall in WSS till 3 clusters, but after 3 clusters the reduction rate is not significant, therefore Elbow cure suggest an optimum number of 3.

Let's also check for silhouette score, which check if each observation is correctly mapped to its class. Silhouette score ranges from -1 to +1

Silhouette score for 3 clusters is 0.39870739416187406

A positive silhouette score indicates on average the observations are mapped correctly, but lets also check e ach individual silhouette width.

There are 4 values with negative sign which indicates 4 observations with being have wrongly assigned to c lusters. Two clusters will give better results but from business perspective we need at least 3 to provide mo re customised approach. Therefore, we go ahead with 3 clusters as suggested by elbow curve.

The following tables depicts the first five observation of data frame after the Kmeans clusters have been ap pended to the data frame

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	class
0	19.94	16.92	0.87520	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.90640	5.363	3.582	3.336	5.144	0
2	18.95	16.42	0.88290	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.87345	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.89920	5.890	3.694	2.068	5.837	1

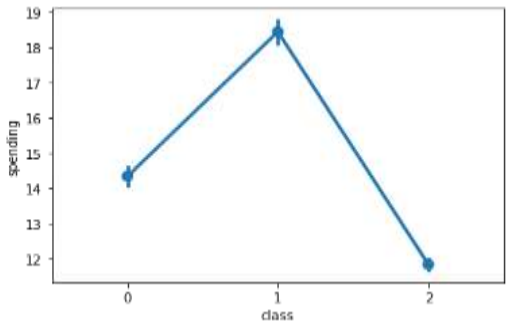
1.5Cluster Profiling and Recommendations.

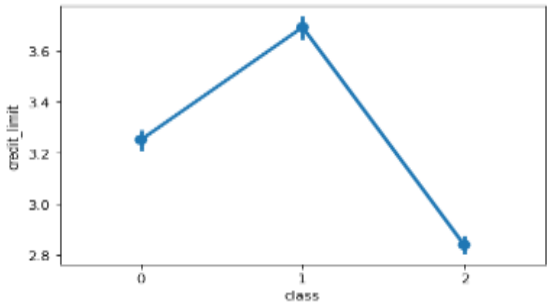
Following represent the mean values of attributes for each class.

Table 2 class wise descriptive

class	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	14.340274	14.284658	0.881952	5.492767	3.251219	2.701729	5.103301
1	18.449559	16.184265	0.884010	6.171397	3.691221	3.644574	6.038074
2	11.834348	13.248406	0.850311	5.237174	2.840072	4.672638	5.109638

We can observe from the above table that cluster 1 has highest average spend and credit limit and max spent is also high compared to other clusters. The below graph visualizes the spending for each cluster.





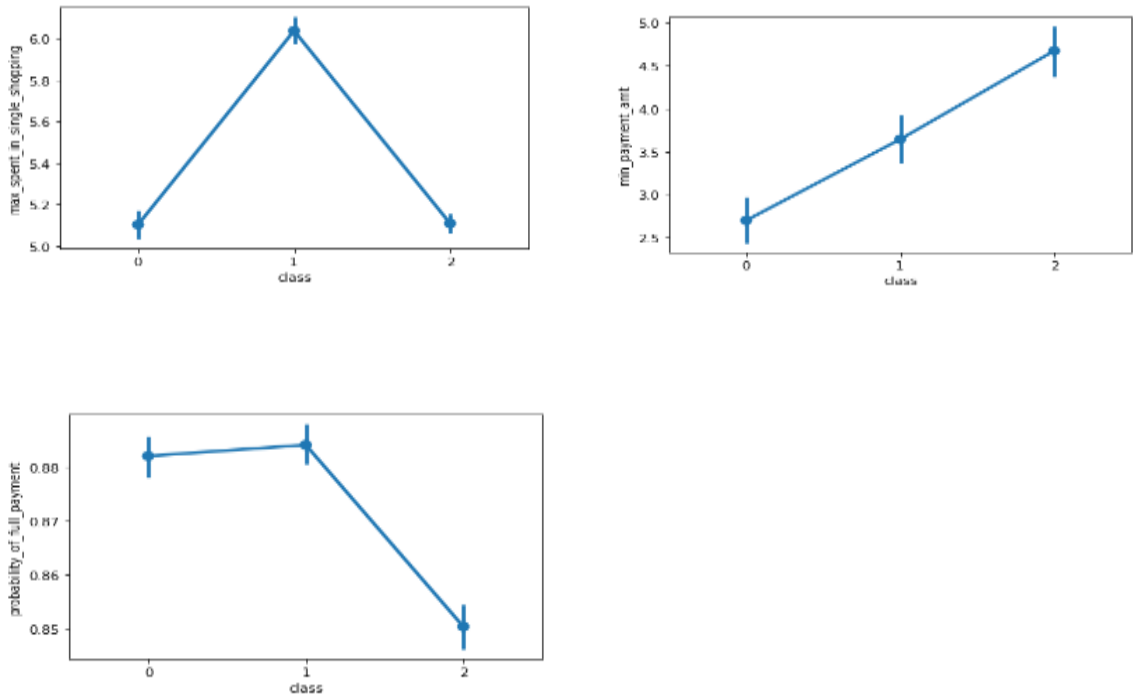


Figure 12- analysis of variable on class

- Class 0 has almost equal probability of full payment when compared to class 1. And also, class 0 has lowest minimum payments which imply that they use credit card even for small purchases and regular payments.
- Class 1 has highest spending, and credit limit, therefore this clusters can be considered as Spenders and also premium customers.
- Class 2 has lowest probability of full payment when compared to other two classes. And also, they have high average on minimum payment amount. And class two has lowest credit limit and spending, but their max spent on single shopping is close to class 1(almost 50% of their spending), which imply that they spend their money mostly in bulk. And also, they have lowest minimum amount payment which imply that they are not regular users or do not use credit card for making small purchases.
- Therefore class 0 can be defined as regular users of credit card and potential high spenders, Class 1 can be defined as premium customers and whereas class 2 can be defined as Bulk purchase users.

Recommendations:

- As class 2 use credit cards for bulk payments, promotional offers like discounts should be provided to them during sales like Amazon big billion-day, flip-kart -sale events. And as their probability of full payment is less, EMI options can be made available to them.
- As class 0 has least min payment amount, cashback offers can be provided to them to increase their frequency of usage. As their probability of full payment is also high, options to increase the credit limit should be provided if available.
- Class 1 has highest spending and highest credit limit, therefore offers like subscribing to premium services (like gold card or platinum card) should be offered to them. And also, premium customer services can also be provided to this customer.

Problem-2

Introduction: An insurance company is facing higher claim rates. Therefore, the Management wants to understand the reasons for the claim rate. Therefore, they have collected the claim data for past years and wants to study the data to ascertain the factors influencing the claim rate.

Objective: The objective the study is to understand the data and Apply CART , Random forest, ANN techniques to ascertain the factors influencing claimed tar get variable.

2.1 Exploratory data analysis

Data description:

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

Data dimension: The data contain total of 3000 observations with total of 10 variables or attributes.

Data info: following tables provides information about data types and non-null values.

Table 3 data info

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Age           3000 non-null   int64
 1   Agency_Code   3000 non-null   object
 2   Type          3000 non-null   object
 3   Claimed       3000 non-null   object
 4   Commision     3000 non-null   float64
 5   Channel       3000 non-null   object
 6   Duration      3000 non-null   int64
 7   Sales         3000 non-null   float64
 8   Product Name  3000 non-null   object
 9   Destination   3000 non-null   object
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

Numerical variables include: Age, commission, duration, sales

Categorical variables include: Agency code, Type, claimed, channel, product name, Destination. Also the 3000 non-null values corresponding to all variables confirms that there are no null values in the data.

Data head: Data head represents the first five observations of the data frame:

Table 4 data head

	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

Checking for null values: The following output confirms the absence of null values in the data:

Table 5 null values

```
Age             0
Agency_Code     0
Type            0
Claimed         0
Commision       0
Channel         0
Duration        0
Sales           0
Product Name    0
Destination     0
dtype: int64
```

Univariate analysis:

1)Age

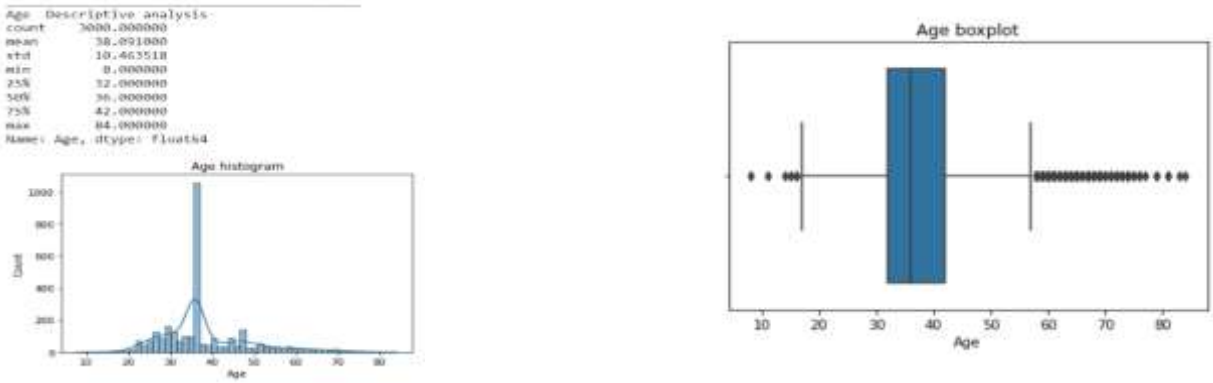


Figure 13-Age analysis

- Average age is 38.091, with 50 % of customers having less than 36, higher mean than median indicates a right skewness. And we can observe high proportion of customers at 36-38.
- Boxplot confirms that there are high proportion of outliers in the variable.

2)Agency code

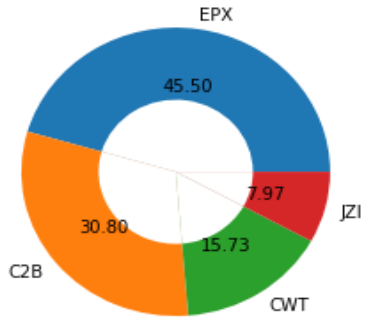


Figure 14- Agency code analysis

- The above pie plot shows the proportion of each tour firm, as we can see tour firm with code epx has highest proportion and tour firm with code JZI has lowest proportion.

3)Type

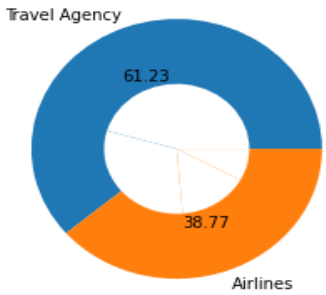


Figure 15- Type analysis

- Above plot shows the proportion of each type of insurance firms, there are 2 types Travel agency and Airline. With travel agency being higher in proportion.

4)Claimed

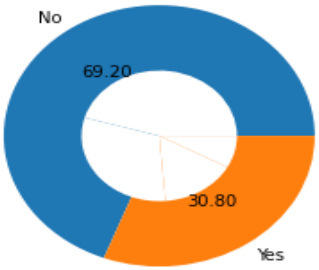


Figure 16-claimed analysis

- There two level in target Variable. 1)No 2)Yes with no being around 69% of total observations.

5)Commission

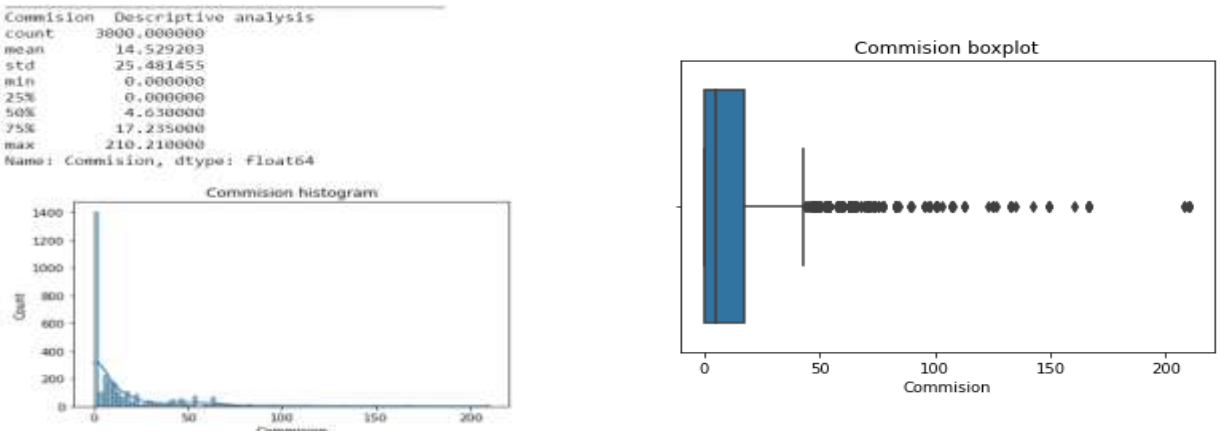


Figure 17- commission analysis

- On average commission paid is 14.52 percent, whereas 50% times commission paid is less than 4.63%. The data is highly right skewed. Max commission paid is 210 percent which implies that there are anomalies or erroneous data present in commission variable. As a logic the commission cannot be higher than 100% sales.
- Boxplot confirms there are high proportion of outliers in the data.

6)Channel

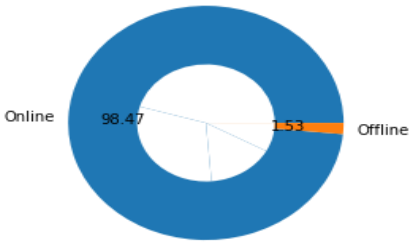


Figure 18 channel analysis

- We can observe that offline channel has very small proportion of observation only 1.53%, We will further see the importance of the variable while Model building.

7)Duration

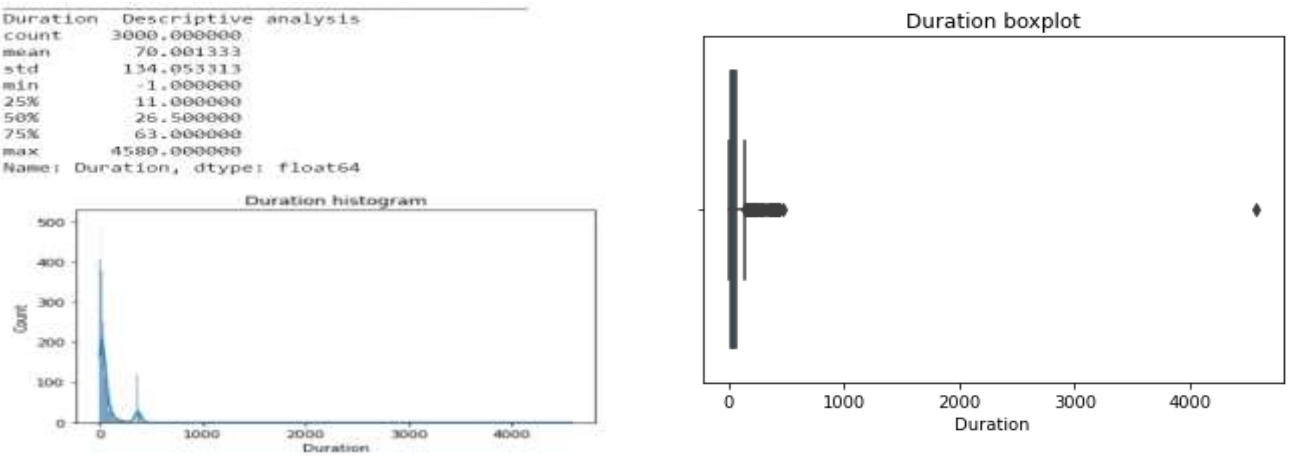


Figure 19- Duration analysis

- We can see that Mean duration is 70 days with median being 26, the values tell us that the variable is extremely Right skewed. And with max being 4580(almost 13 years), there is erroneous data in the variable, the sales value for observation is also less therefore it vam be considered as long term insurance.
- And box plot also confirms that there are outliers in the data.
- The minimum value is also in negative, therefore the two values will be imputed with the median.

8)Sales

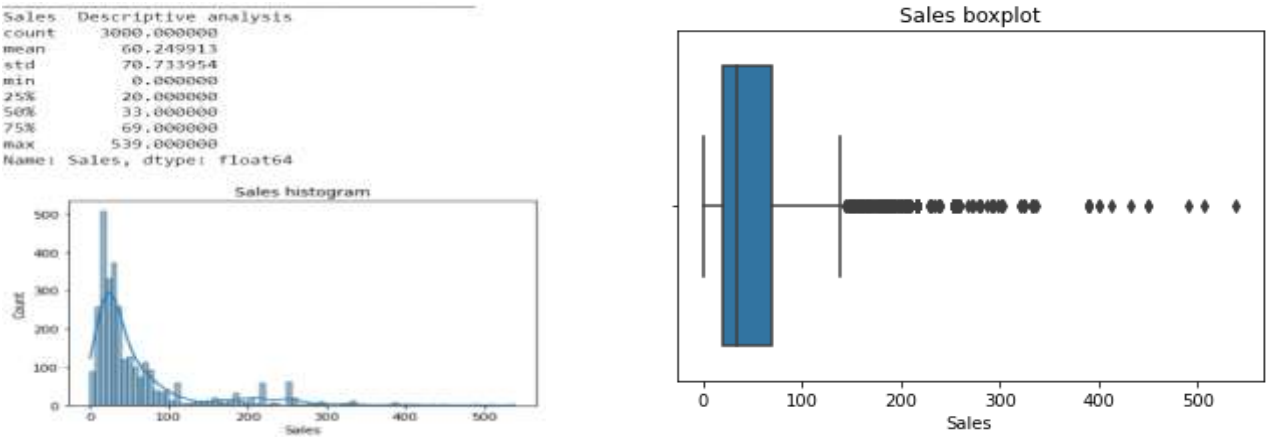


Figure 20-Sales analysis

- Average sales are 6024 where as 50% sales are less than 3300, indicating right skewness in the variable.
- Box plot also confirms the presence significant number of outliers in the variable.

9)Product Name

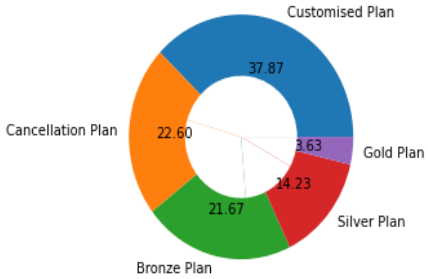


Figure 21-Product Name analysis

- From the above figure, Customised Plan has highest proportion with 37.87 observation, whereas Gold plan has lowest with 3.63% of observations.

10)Destination

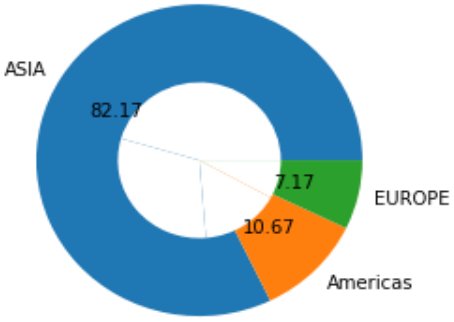


Figure 22-Destination analysis

- Destination Asia has highest proportion with 82.17% with Europe being the lowest with 7.17%

Multivariate analysis:

Pair plot helps us understand the relationship between the numerical variables using scatter plot.

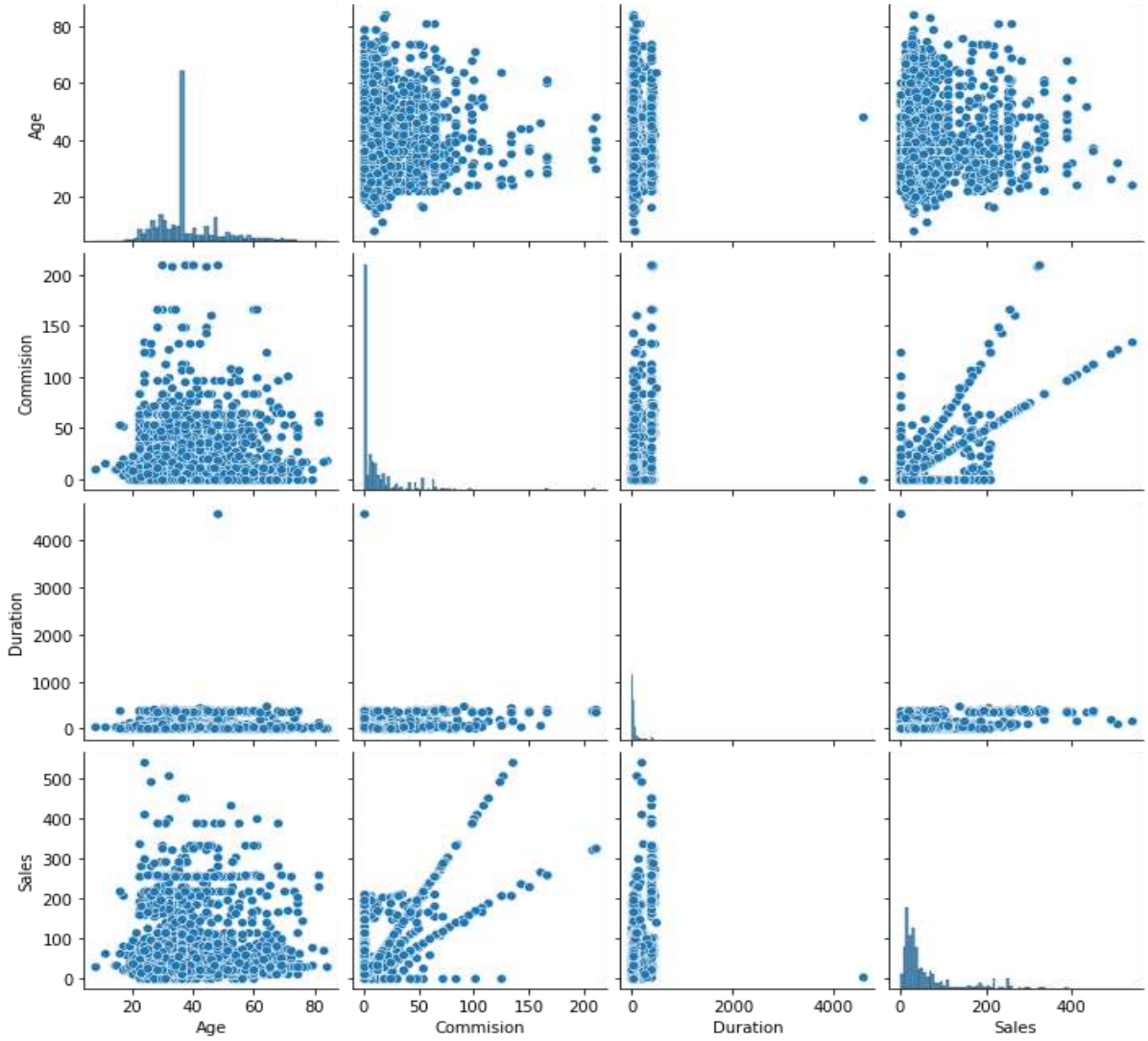


Figure 23-Pair plot insurance

- From the above plot, It can be observed that only commission and sales seems to be having linear relationship, however the magnitude of correlation can be checked with the help of correlation plot.

Correlation plot:

Figure 24-Correlation plot insurance

- From the above figure, it's evident that commission and sales show a correlation of .77(This correlation can be impacted due to presence of outliers.)

2.2 Train-Test split and Application of CART, ANN, Random Forest.

First the categorical variables data is coded into numerical type using pd.Categorical.codes, And the outliers are treated for Cart and Random Forest, but they have been treated for ANN as its sensitive to the outliers. The duplicate observations have not been dropped as there is no unique identification given to each observation.

The data has been split, with train size of 70% and with test data of 30 percent, with random state=0

(Random state specification would help us get same split every time execute the split)

Application of Random Forest:

First the Gridsearch CV is run to check the best hyper paramenters. The following levels are used in Param_grid

{"min_samples_split":[100,200],

"min_samples_leaf":[50,100],

"max_depth":[6,7]}

And following are results given when checked for Best parameters

{'max_depth': 6, 'min_samples_leaf': 100, 'min_samples_split': 100}

Therefore, the above suggested hyperparameters will be used to build the model.

The model has been fit into the Train data set and following are feature importance produced

	imp
Agency_Code	0.645605
Sales	0.239642
Product Name	0.055257
Duration	0.035850
Commision	0.023646
Age	0.000000
Type	0.000000
Channel	0.000000
Destination	0.000000

Table 6-CART feature importance

As we can see agency code, sales have highest importance in CART model in predicting the target variable, and variables like Age, Type, channel, Destination do not have any importance with respect to Cart model

Random Forest Model:

GRID search CV is also used for random forest in ascertaining the best hyperparameters.

Following are the levels in hyperparameters considered.

{"min_samples_split":[100,200],

"min_samples_leaf":[50,100],

"max_depth":[6,7],

"n_estimators":[100,200,300]}

The GridSearchcv has returned the following parameters as best params.

max_depth': 7,
'min_samples_leaf': 50,
'min_samples_split': 100,
'n_estimators': 200

We will go ahead with above parameters for building our random forest model, Please Note, Random State 0 is used in the model to produce same randomisation of samples from data, every time we run the model.

Following are best feature results for Random Forest model:

Table 7-feature importance

Agency_Code	0.295018
Product Name	0.255721
Sales	0.177598
Commision	0.123391
Duration	0.065420
Type	0.058200
Age	0.016563
Destination	0.008088
Channel	0.000000

From the above table, its evident that Agency code is the most important feature followed by Product name and sales, Variables like Type, Age, Destination, Channel have no significant importance as per model.

ANN model:

We have chosen to scale the data before proceeding with ANN, however we choose not to treat outliers for this model, and the erroneous data has been replaced with the median values.

Following are the levels considered while running GRIDSEARCHCV

{"solver":["adam","sgd"],

"hidden_layer_sizes":[100,200,300],"activation":["relu","logistic"]}

GRIDSEARCHCV has given the following best parameters:-

'activation': 'logistic', 'hidden_layer_sizes': 100, 'solver': 'adam'

Therefore, Model will be built using above hyperparameters. The ANN model does not provide us the feature importance. However, the performance of the model will be analysed with help of Model performance metrics in further analysis.

2.3 Performance Metrics Using confusion matrix, classification Report, Roc_score,Roc_curve

The performance of each model will be analysed with the help of above mentioned metrics. Lets analyse the models each by each. The results will be shown both on Test data and train data

Confusion matrix= confusion matrix provides cross tabulation of predicted value to actual value.

Classification report = calculates the precision, F1 score, recall ratio for the model.

Roc_auc_score- it calculates the total area under the Roc curve

CART model:

Table 8-classification report -cart

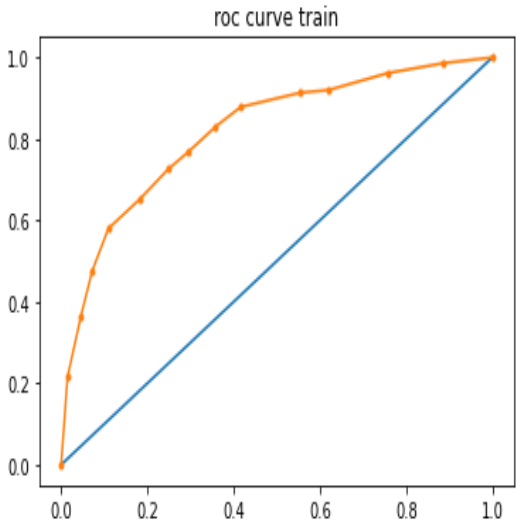
	precision	recall	f1-score	support
0	0.80	0.91	0.85	1464
1	0.68	0.47	0.55	636
accuracy			0.77	2100
macro avg	0.74	0.69	0.70	2100
weighted avg	0.76	0.77	0.76	2100

```
array([[568,  44],
       [151, 137]], dtype=int64)
```

	precision	recall	f1-score	support
0	0.79	0.93	0.85	612
1	0.76	0.48	0.58	288
accuracy			0.78	900
macro avg	0.77	0.70	0.72	900
weighted avg	0.78	0.78	0.77	900

```
array([[568,  44],
       [151, 137]], dtype=int64)
```

As we can observe from classification report that the precision for class 1 on train and test data(as we need to predict the claimed , class 1 si more important here) is .68 and 0.76 respectively i.e around 70 % of predicted class 1 are correct, but the recall which tells the actual 1 which predicted is less only at 48% is not a efficient result, however lets check for other models as well and lets also check ROC score and Roc curve for train and test.



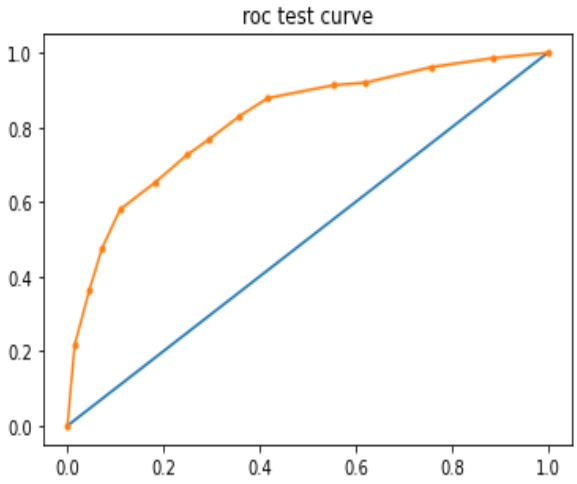


Figure 25-roc curve cart

Roc score for train and test data is 0.809, 0.818 respectively.

Random forest:

The following are classification report, confusion matrix, roc_curve for the train and test data respectively.

Table 9- classification report Random- forest for train and test

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1464
1	0.69	0.52	0.59	636
accuracy			0.78	2100
macro avg	0.75	0.71	0.72	2100
weighted avg	0.78	0.78	0.77	2100

	precision	recall	f1-score	support
0	0.80	0.92	0.85	612
1	0.74	0.51	0.60	288
accuracy			0.79	900
macro avg	0.77	0.71	0.73	900
weighted avg	0.78	0.79	0.77	900

Table 10-confusion matrix- random forest for train and test

```
array([[1318,  146],
       [ 307,  329]], dtype=int64)
```

```
array([[562,  50],
       [142, 146]], dtype=int64)
```

From the above table it can be observed that the precision for the positive class 1 is 69% for train and 74% for test data, and sensitivity for train and test is 53% and 51% respectively, has shown similar performance both on test and train data.

From confusion matrix, out of 636 positive cases 329 have been predicted correctly in train data and in test out of 288 positive cases 146 have been predicted as positive cases. Though the model has good precision, the model has shown moderate results on sensitivity. This can be due to imbalance in the target variable as we have more negative cases than positive cases.

Let's also plot the roc curve and roc auc score for the model. The following figures are roc curves for train and test data respectively.

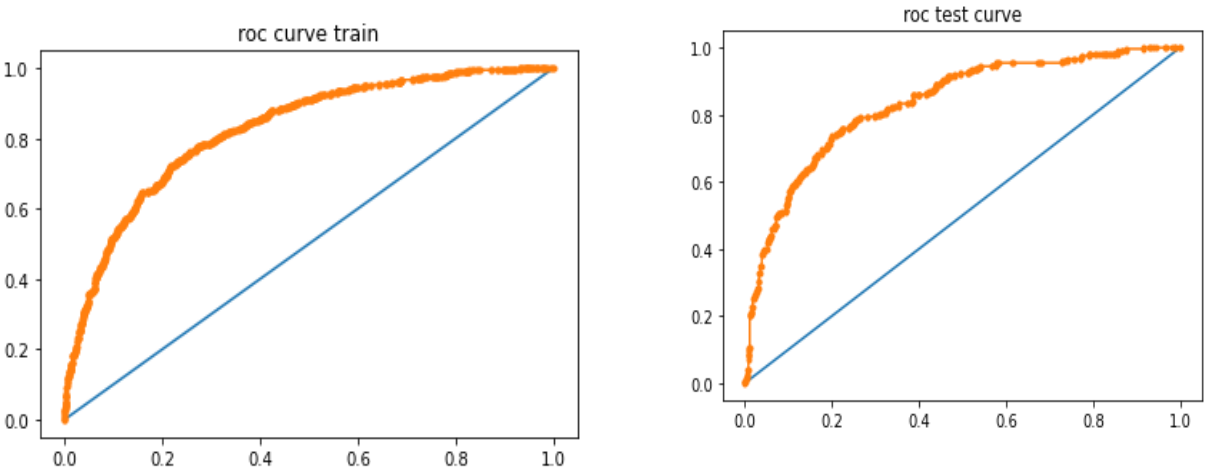


Figure 26-roc curve for random forest

The x-axis is for FPR and y-axis is of TPR, and the roc score for above train data and test data are 0.82 and 0.83 respectively. Has shown similar performance on both train and test data.

ANN model performance: The following are classification report , confusion matrix , roc_curve for the train and test data respectively

Table 11-classification report for ANN

	precision	recall	f1-score	support
0	0.78	0.94	0.85	1464
1	0.74	0.38	0.50	636
accuracy			0.77	2100
macro avg	0.76	0.66	0.68	2100
weighted avg	0.77	0.77	0.74	2100

	precision	recall	f1-score	support
0	0.76	0.96	0.85	612
1	0.82	0.35	0.50	288
accuracy			0.77	900
macro avg	0.79	0.66	0.67	900
weighted avg	0.78	0.77	0.74	900

Precision for the positive class in train and test is .74 and .82 respectively. However, the sensitivity for the positive class is very less .38, therefore the ANN model has not predicted accurately. This can be due to imbalanced data. Lets also plot the ROC curve and ROC score for the train and test.

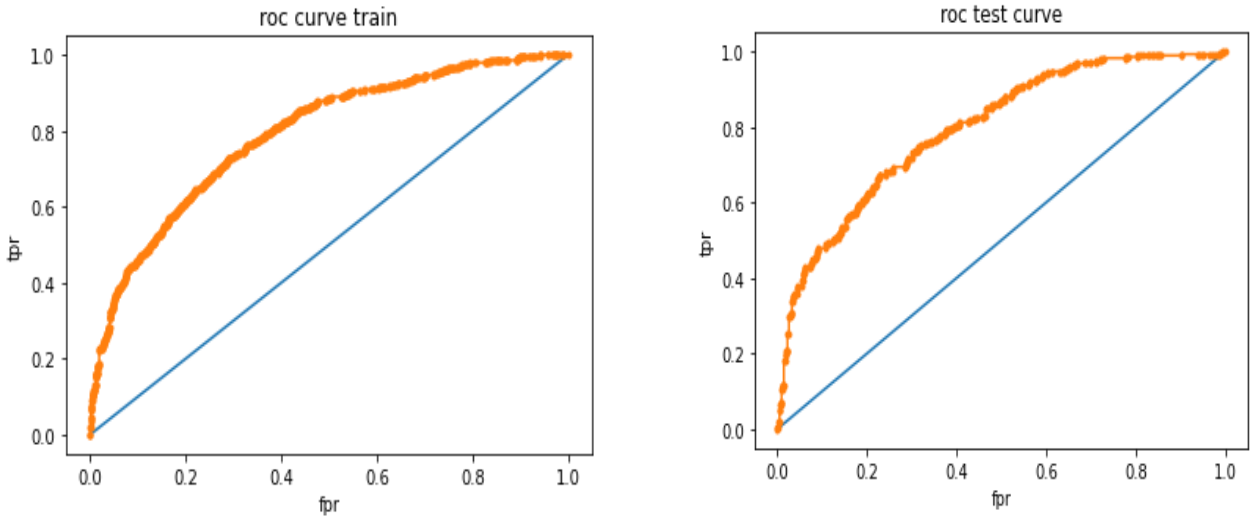


Figure 27-Roc curve for Train and test for ANN

The roc curve for train and test data are .79 and .79 respectively. The roc score is also less compared the other three models.

2.4 Final Model

The below table summarizes the Precision, recall, f1 score of each model for test data for positive class.

Model/metrics	Precision	recall	F1 score	ROC score
CART	.78	.48	.58	0.818
Random Forest	.70	.51	.60	0.83
ANN	.82	.35	.50	0.79

As we can observe form the above table, that precision is highest in ANN model, however the recall is lowest in ANN, therefore we can go ahead with random forest as it has given more balanced results. And also, the roc score, is also highest for random forest. Therefore, we can say that random forest is more optimized for this data.

Recommendations:

as per Random Forest, agency code and product are given highest importance, Lets further understand using the following count plot

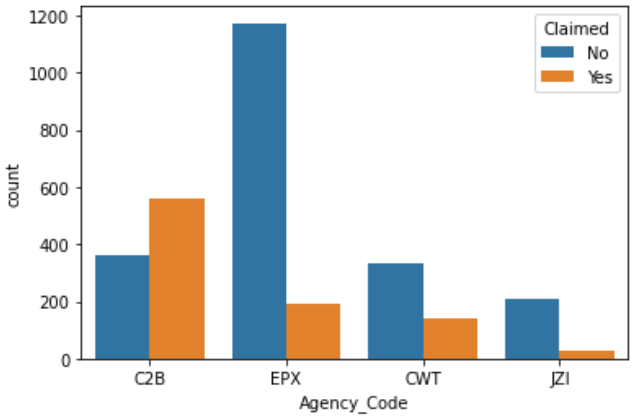


Figure 28-agency code vs claimed

- We can observe that Agency with code EPX has very lower proportion of claims there fore the insurance company should focus selling Insurance policies through that EPX agency, by incentivising with high commission.
- Also investigate further why C2b agency code has higher claim rate
- Further understand claims with product name

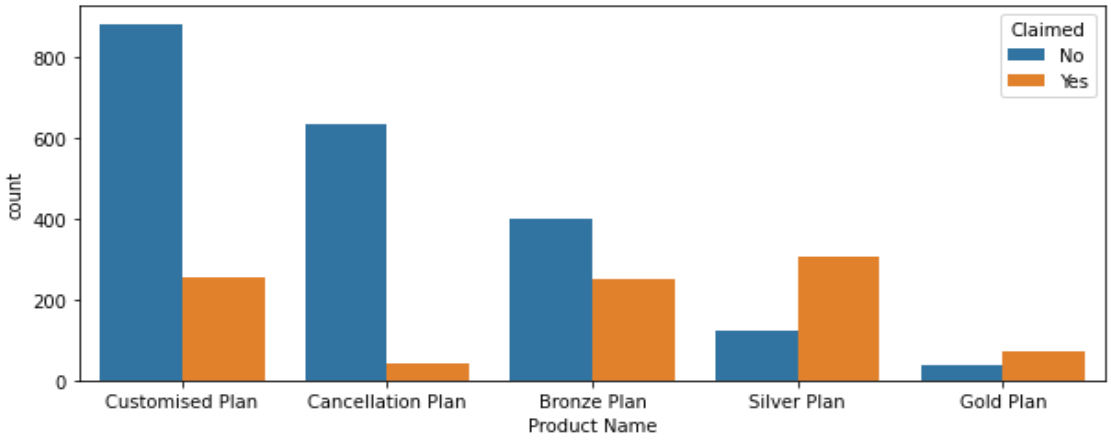


Figure 29- product vs claim

- We can observe that customised plan and cancellation plan has lowest rate of claims, therefore insurance company should further push this product through promotion, maybe increase the premiums of silver and gold plan as the claim rates are high.
- The insurance company should try to push the combination of customised plan, cancellation plan and EPX agency code to further decrease the chance of claim.