Data Mining Project

PGP – DSBA Online

May ‘22

Date: 02-10-2022

**Contents**

1 Problem Statement 1: Clustering………………………………………………………………………………………….5

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)…………………………………………………………………………………………..5

1.2 Do you think scaling is necessary for clustering in this case? Justify…………………………………11

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them………………………………………………………………………………...12

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………………………………………………………………………………………………………………………………...13

1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters……………………………………………………………………………………………………………..15

2 Problem Statement 2: CART-RF-ANN..………………………………………………………………………………………….16

2.1Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)..…………………………………………………………………………………………………16

2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network..……………………………………………………………………………………………………………….23

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...…………………………………………………………………………………………………………………………………..26

2.4 Final Model: Compare all the models and write an inference which model is best/optimized…...32

2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations..33

**List of Figures**

Fig 1. Information of bank dataset………………………………………………………………….5

Fig 2. Univariate Analysis………………………………………………………………………………..6

Fig 3. Bivariate Analysis…………………………………………………………………………………..8

Fig 4. Multivariate analysis……………………………………………………………………………..9

Fig 5. Outlier treatment………………………………………………………………………………...10

Fig 6. Before scaling……………………………………………………………………………………….11

Fig 7. After scaling………………………………………………………………………………………….11

Fig 8. Cluster Frequencies……………………………………………………………………………...12

Fig 9. Within sum of squares and wss plot……………………………………………………..13

Fig 10. Silhouette score and silhouette coefficient plot…………………………………..13

Fig 11. Appending cluster labels to original dataframe…………………………………...14

Fig 12. ElbowVisualiser……………………………………………………………………………………14

Fig 13. Cluster Profiling…………………………………………………………………………………..15

Fig 14. Description of dataset 2……………………………………………………………………...17

Fig 15. Univariate Analysis of numeric variables……………………………………………..18

Fig 16. Univariate Analysis of Categorical Variables………………………………………..19

Fig 17. Pair plot for bivariate analysis……………………………………………………………..20

Fig 18. Correlation Heat Map for Multivariate Analysis…………………………………..21

Fig 19. Numeric datatypes………………………………………………………………………………23

Fig 20. Decision Tree Classifier (CART)…………………………………………………………….23

Fig 21. Area Under the Curve for CART Train data…………………………………………..26

Fig 22. Area Under the Curve for CART Test data…………………………………………….27

Fig 23. Area under the curve for Random Forest Train data…………………………….28

Fig 24. Area under the curve for Random Forest Test data ……………………………..29

Fig 25. Area under the curve for ANN Train data …………………………………………….30

Fig 26. Area under the curve for ANN Test data……………………………………………….31

Fig 27. DataFrame CART-RF-ANN values…………………………………………………………..32

Fig 28. ROC Curve for the 3 models………………………………………………………………….32

Fig 29. ROC Curve for the 3 models on test data……………………………………………...32

**Problem Statement 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.

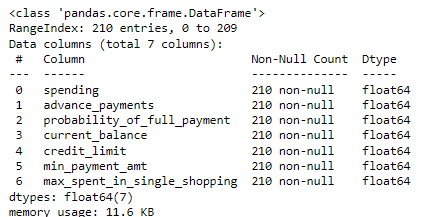
**1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).**

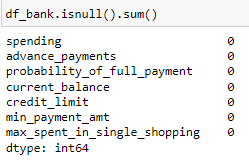
From the initial validation of the dataset it is observed that:

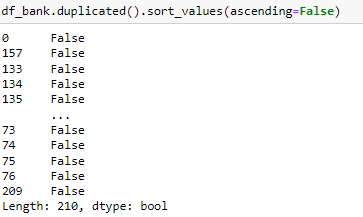
a. The dataset contains 210 non-null columns across 7 features, all of numeric datatype. So we need not change the datatype for further analysis.

b. We do not have any null values or missing records

c. We do not have any duplicated values

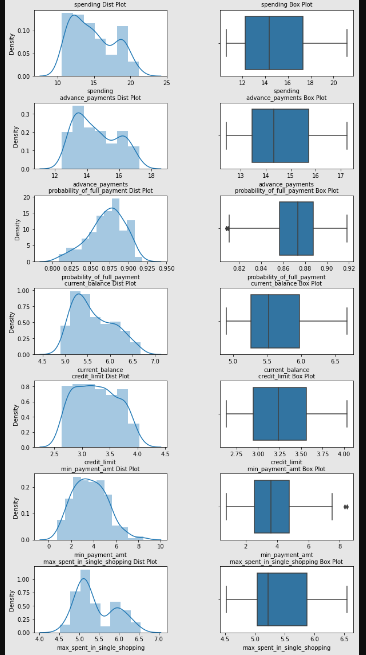


**

**

*Fig 1: Information of bank dataset*

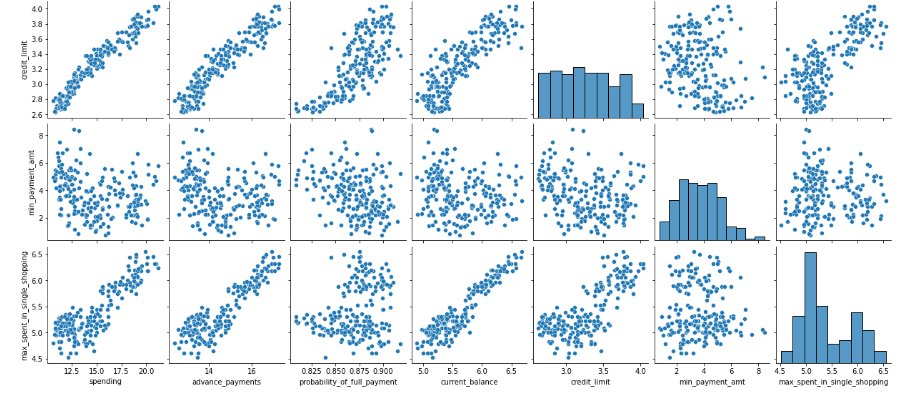
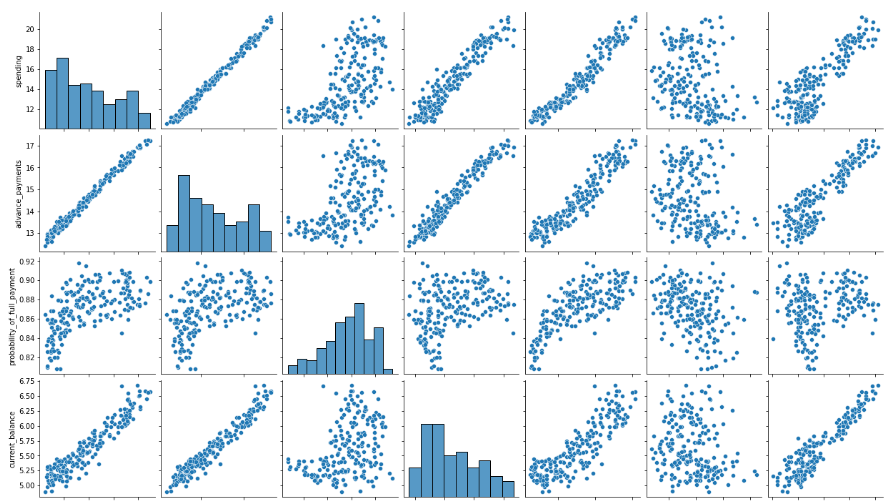
Univariate Analysis – Histogram and Box Plot

****

*Fig 2: Univariate Analysis*

Observation: From univariate analysis it is observed that all the feature distributions are right-skewed except probability\_of\_full\_payment which is left-skewed.

Bivariate Analysis – Pair Plot

*Fig 3. Bivariate Analysis*

Observation:

- spending and advance\_payments are highly correlated

- spending and credit\_limit are highly correlated

- spending and current\_balance are highly correlated

- current\_balance and max\_spent\_in\_single\_payment are highly correlated

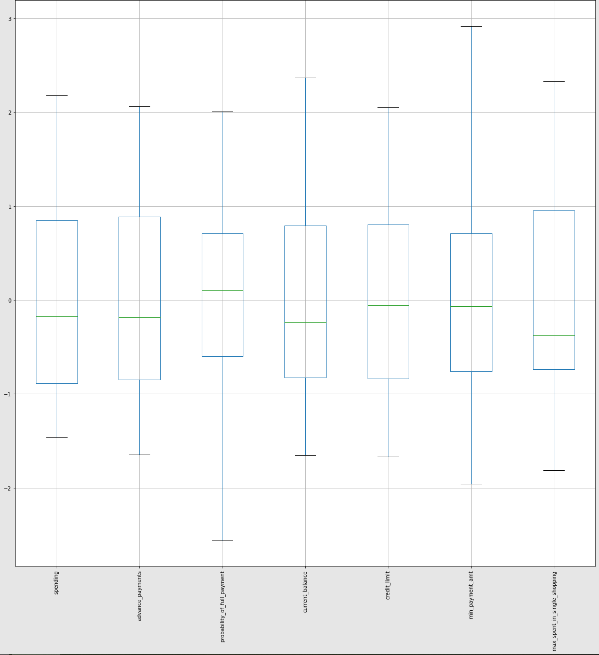
- current\_balance and advance\_payments are highly correlated

- credit\_limit and advance\_payments are highly correlated

Multivariate Analysis – Correlation Heat Map

*Fig 4. Multivariate analysis*

Observation: Mostly positive correlations are observed between the variables.

*Fig.5 Outlier treatment*

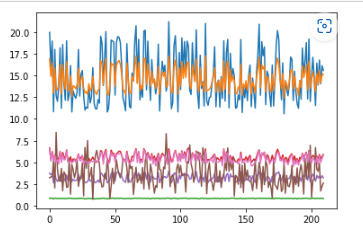
Outliers may lead to misinterpretation of analysis data due to their extreme values. Hence they have been treated to proceed with further analysis.

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

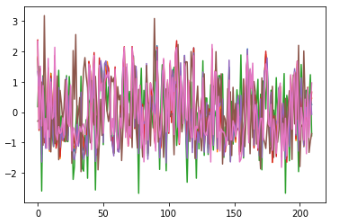
Scaling is necessary in this case because the variables are not of the same scale. Scaling needs to be preformed to standardize the range of features. It is a very important step in data pre-processing while using machine learning algorithms.

Spending and advance\_payments features are in higher scales than the other features, so scaling them using z-score technique to bring them within -3 and +3 values.

This can be understood from the figures below.



*Fig 6: Before scaling*



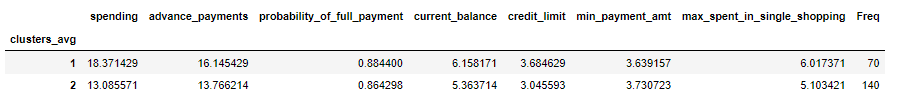
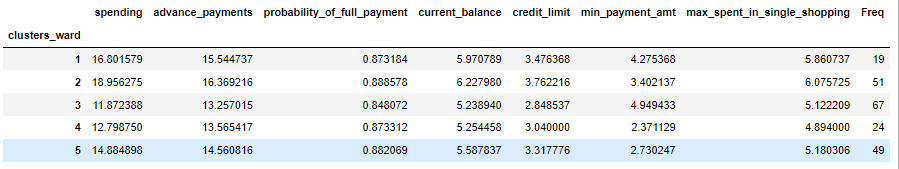
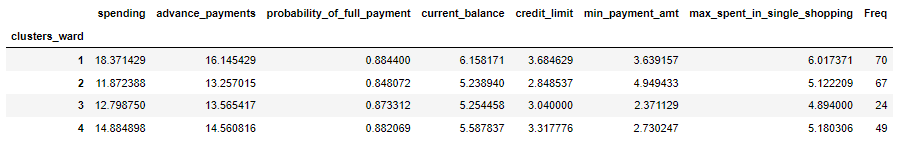
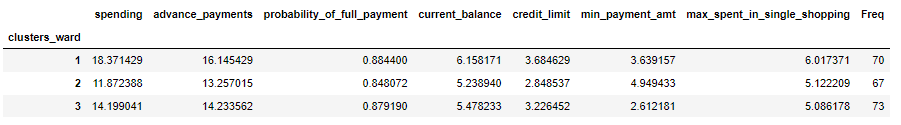
*Fig 7: After scaling*

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

Hierarchical clustering is performed to create clusters of sequential records based on the distance between records and distance between clusters by forming a graphical representation called dendrogram.

Agglomerative clustering is used when we have smaller datasets. For larger datasets it is preffered to use divisive clustering.

Divisive clustering – Ward’s method has been used to derive the results. This is the most optimal method as we have a big dataset and Ward’s method joins the records progressively to produce large clusters. Average method has also been carried out, but the results of both these methods are similar.

*Fig 8. Cluster Frequencies*

2

3

4

5

Above we can see the cluster frequencies for 2, 3, 4 and 5 clusters. The difference between the frequencies for 3 clusters is minimum when compared to 2, 4 and 5 clusters. Hence we choose 3 clusters for hierarchical clustering.

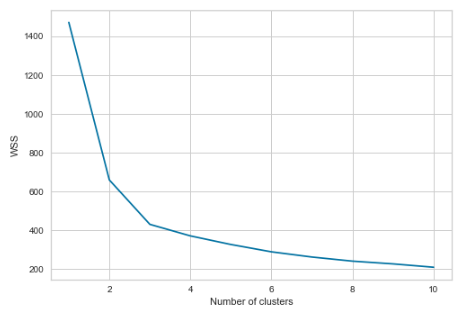
Observation:

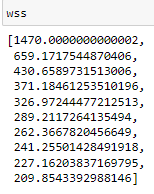
This 3 cluster result set can be assumed as a patter for high, medium and low spending.

High spending customers probability to do the full payment to the bank is 88%, pay an advance of approx. Rs. 1600 to the bank and spend around Rs.18000 per month.

Low spending customers have a credit limit of Rs.28000 and spend around Rs.11000 per month.

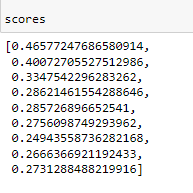
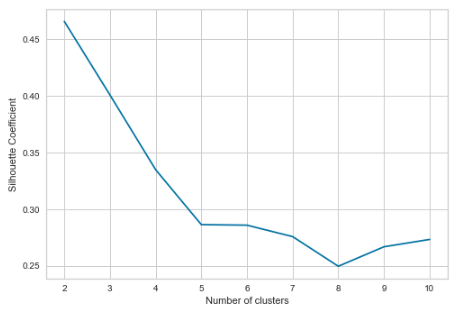
**1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.**

****

**

*Fig 9. Within sum of squares and WSS plot*

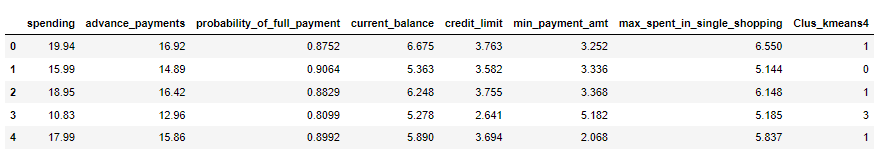
Wss plot helps to identify the optimal number of clusters for this dataset. The Elbow point cuts at 3, hence the optimal number of clusters for further analysis would be 3 in this case. The plot looks like an elbow at k = 3 (Fig 9).

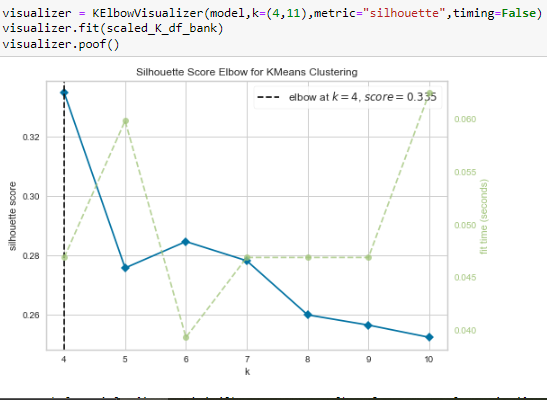
**

*Fig 10. Silhouette scores and Silhouette coefficient plot*

Silhouette score is another technique to determine the number of clusters. The range of this coefficient varies between -1 and 1. It tells us the distance between the neighbouring cluster (1) and the distance within the same cluster (-1). Fig 10 shows a plot of these scores, from which we can determine the optimal number of clusters to be 3 or 4.

From both these above methods we can determine the optimal number of clusters to be 3 or 4.

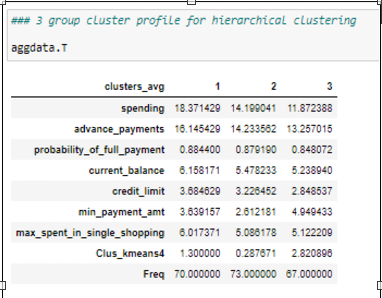
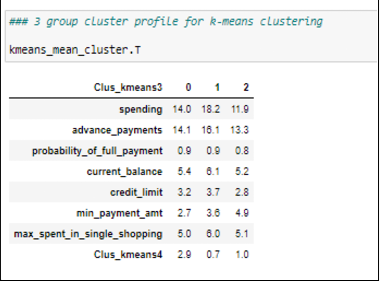
*Fig 11. Appending cluster labels to original dataframe*

**

*Fig 12. Elbow Visualiser*

The elbow visualiser above can also be used to determine the k value. The elbow is annotated with a dashed line (---).

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

****

*Fig 13. Cluster Profiling*

For hierarchical clustering: For k-means clustering:

Group 1: High Spending Group 1: High Spending

Group 2: Medium Spending Group 0: Medium Spending

Group 3: Low Spending Group 2: Low Spending

Promotional Strategies

High Spending Group:

The bank can suggest that for the customers who pay in full, the bank can increase their credit limit.

The max amount spent in single shopping is around Rs.6000 and hence for customers spending Rs.10000 in a single shopping can be offered discounts on their next purchase.

The high spending customers are always valuable assets for the bank, hence the bank can offer loans to them.

Medium spending group:

These are customers who try to pay bills on time and also maintain a decent credit balance, hence by increasing their credit limit the bank can hope to improve their spending.

Their probability of full payment and maximum amount spent in single shopping is also high, hence these customers can be offered premium benefits.

By providing a better interest rate to these customers, the bank can aim at prolonging the tenure of deposits made by them.

Low spending group:

These customers have low spending and low probability of full payment. The bank can provide them with festive offers / occasional offers to improve their interest on spending.

The bank can target and improve business with these customers by decreasing their minimum payment amount and by increasing their credit limit.

**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.

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

The data has 3000 rows and 10 columns.

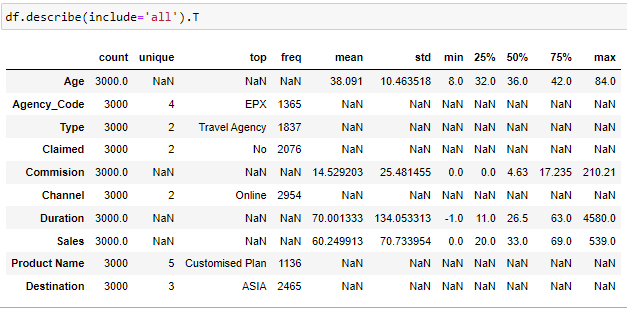
There are no missing or null values.

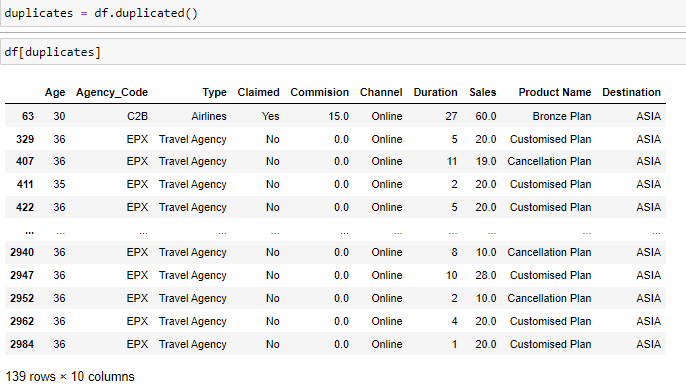
There is a minimum value of -1 for duration, which needs to be replaced.

Age, commission, Duration and Sales are numeric variables, and the rest are categorical variables.

Claimed is the target variable.

There are 139 duplicates, but removing them is not wise, as we do not have any unique identifier. They may be valid customers. Hence not removing the duplicates.

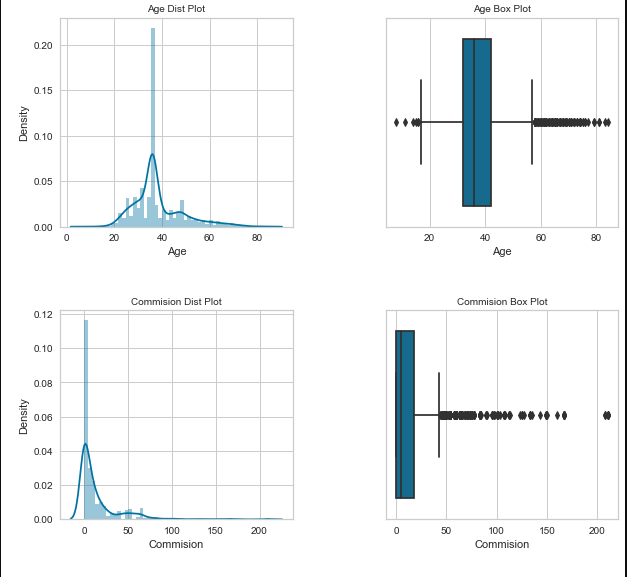


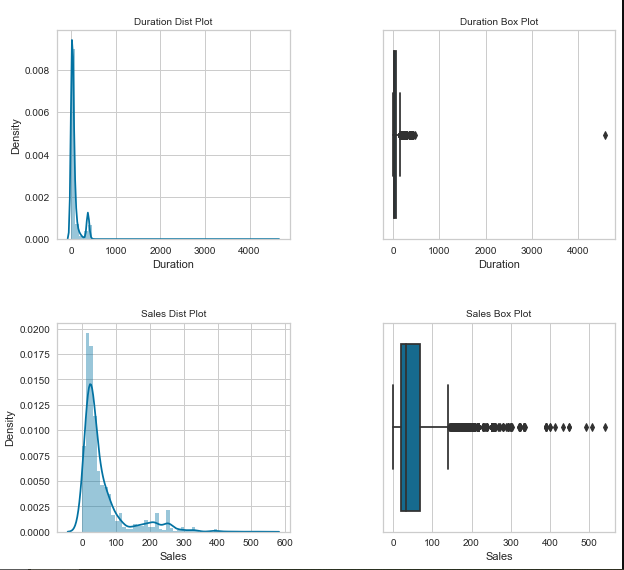
**

*Fig 14.Description of dataset 2*

Univariate, Bivariate and Multivariate Analysis will be discussed further.

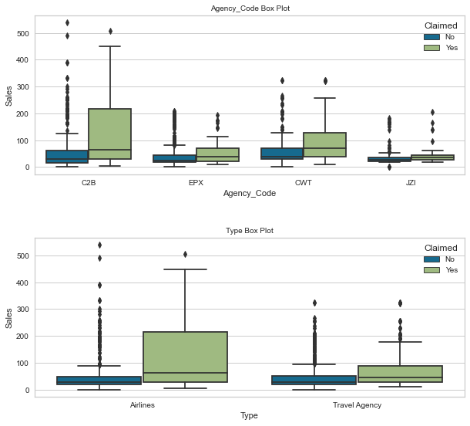
Univariate Analysis – Histogram and Box Plot

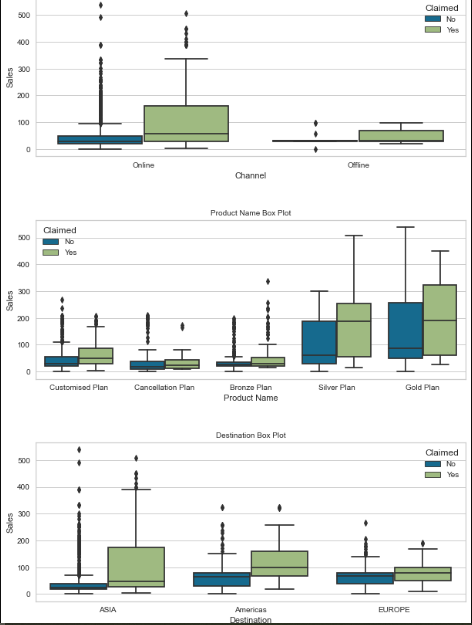
****

****

*Fig 15. Univariate Analysis of numeric variables*

All these variables have many outliers. Random Forest and ANN can handle outliers hence not treating them. They all have a right-tailed distribution.





*Fig 16. Univariate Analysis of Categorical Variables*

There are 4 tour firm agencies (C2B, EPX, CWT, JZI), 2 types of insurance firms (Airlines and Travel Agency), 2 distribution channels (Online and Offline), 5 insurance products (Customized, Cancellation, Bronze, Silver, Gold) and 3 Destinations (Americas, ASIA and Europe). These Box plots above explains the sales under each category and hue describes the claimed status.

Observations:

C2B agency has maximum sales and more claims than the other agencies

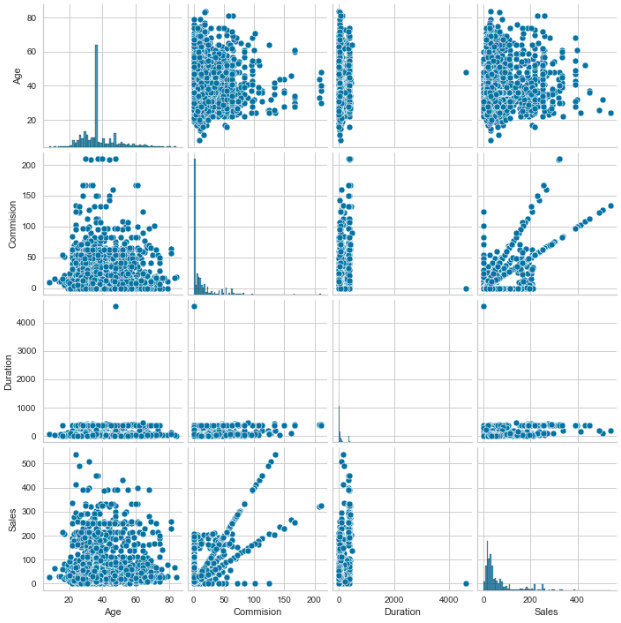
Airlines type of tour insurance providers have more sales and also more claims than that of Travel agencies.

Online transactions are preferred over offline transactions.

Gold Plan has higher sales than any other plans.

Asia shows more claims and sales than that of Europe or Americas.

Bivariate Analysis – Pair Plot

****

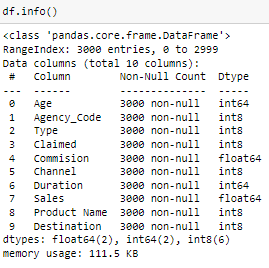
*Fig 17. Pair Plot for bivariate analysis*

Multivariate Analysis – Heat Map



*Fig 18. Correlation Heat Map for Multivariate Analysis*

We can infer from the above two plots that all variables are positively correlated. Commission and sales are highly correlated with each other.

Converting all objects to categorical codes

feature: Agency\_Code

Categories (4, object): ['C2B', 'CWT', 'EPX', 'JZI']

[0 2 1 3]

feature: Type

Categories (2, object): ['Airlines', 'Travel Agency']

[0 1]

feature: Claimed

Categories (2, object): ['No', 'Yes']

[0 1]

feature: Channel

Categories (2, object): ['Offline', 'Online']

[1 0]

feature: Product Name

Categories (5, object): ['Bronze Plan', 'Cancellation Plan', 'Customised Plan', 'Gold Plan', 'Silver Plan']

[2 1 0 4 3]

feature: Destination

Categories (3, object): ['ASIA', 'Americas', 'EUROPE']

[0 1 2]

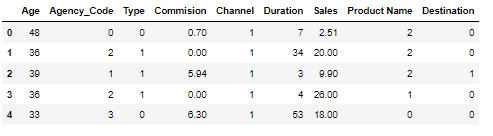
Proportion of 1's and 0's Claimed

0 0.692

1 0.308

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

Data for Testing and Training: Object datatypes have been converted to categorical variables



*Fig 19. Numeric datatypes*

**CART – Decision Tree Classifier**

Train and test labels: test\_size = 0.30, random\_state = 1

X\_train (2100, 9)

X\_test (900, 9)

train\_labels (2100,)

test\_labels (900,)

Grid\_search:

GridSearchCV(cv=10, estimator=DecisionTreeClassifier(random\_state=1),

param\_grid={'criterion': ['gini'],

'max\_depth': [4.85, 4.9, 4.95, 5.0, 5.05, 5.1, 5.15],

'min\_samples\_leaf': [40, 41, 42, 43, 44],

'min\_samples\_split': [150, 175, 200, 210, 220, 230,

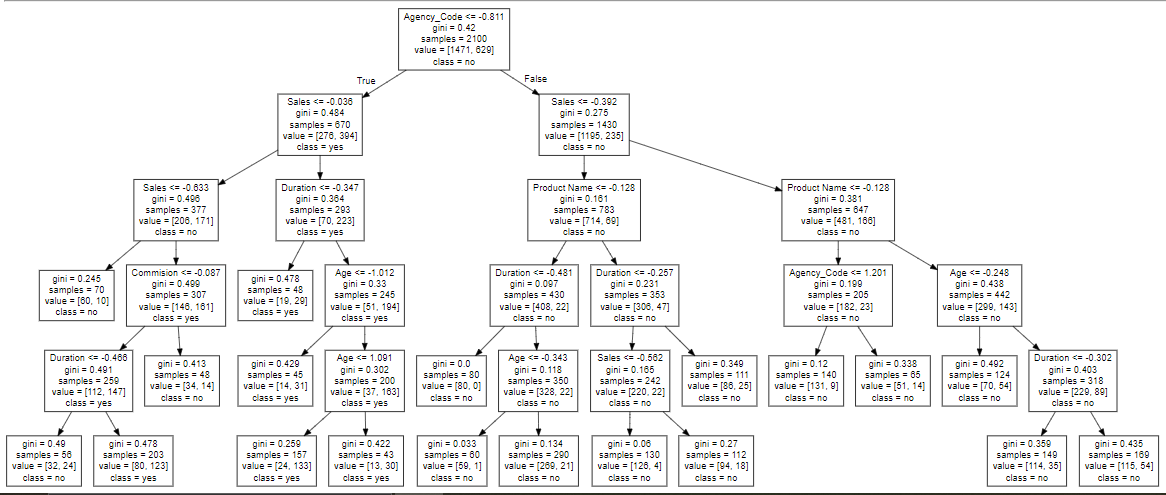
240, 250, 260, 270]})

Best Params: {'criterion': 'gini', 'max\_depth': 5.0, 'min\_samples\_leaf': 42, 'min\_samples\_split': 200}

Best Estimator: DecisionTreeClassifier(max\_depth=5.0, min\_samples\_leaf=42,

min\_samples\_split=200, random\_state=1)

Decision Tree:

**

*Fig 20. Decision Tree Classifier (CART)*

Variable Importance:

Imp

Agency\_Code 0.634112

Sales 0.220899

Product Name 0.086632

Commision 0.021881

Age 0.019940

Duration 0.016536

Type 0.000000

Channel 0.000000

Destination 0.000000

Predicted classes and probabilities:

| **0** | **1** |
| --- | --- |
| 0.935714 | 0.064286 |
| 0.394089 | 0.605911 |
| 0.394089 | 0.605911 |
| 0.311111 | 0.688889 |
| 0.927586 | 0.072414 |

**Random Forest Classifier – RF**

Grid Search:

GridSearchCV(cv=5, estimator=RandomForestClassifier(random\_state=1),

param\_grid={'max\_depth': [4, 5, 6], 'max\_features': [2, 3, 4, 5],

'min\_samples\_leaf': [8, 9, 11, 15],

'min\_samples\_split': [46, 50, 55],

'n\_estimators': [290, 350, 400]})

Best Params:

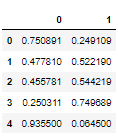
**{'max\_depth': 6, 'max\_features': 5, 'min\_samples\_leaf': 9, 'min\_samples\_split': 46, 'n\_estimators': 290}**

Best Estimator:

RandomForestClassifier(max\_depth=6, max\_features=5, min\_samples\_leaf=9,

min\_samples\_split=46, n\_estimators=290, random\_state=1)

Prediction classes and probabilities:



Variable Importance:

Imp

Agency\_Code 0.353930

Product Name 0.201006

Sales 0.167986

Commision 0.107705

Duration 0.071311

Age 0.052319

Type 0.038278

Destination 0.006493

Channel 0.000972

## **Artificial Neural Network Classifier – ANN**

Grid Search:

GridSearchCV(cv=10, estimator=MLPClassifier(random\_state=1),

param\_grid={'hidden\_layer\_sizes': [50, 100, 200],

'max\_iter': [2500, 3000, 4000], 'solver': ['adam'],

'tol': [0.01]})

Best Grid:

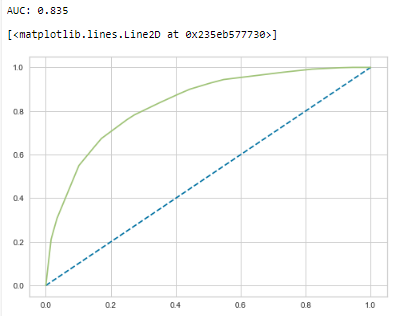
MLPClassifier(hidden\_layer\_sizes=200, max\_iter=2500, random\_state=1, tol=0.01)

Predicted classes and probabilities:

|  | **0** | **1** |
| --- | --- | --- |
| **0** | 0.829737 | 0.170263 |
| **1** | 0.624472 | 0.375528 |
| **2** | 0.527335 | 0.472665 |
| **3** | 0.325167 | 0.674833 |
| **4** | 0.923836 | 0.076164 |

**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.**

**AUC\_ROC\_CART:**

****

*Fig 21. Area Under the Curve for CART Train data*

Accuracy: 0.79

Classification Report:

precision recall f1-score support

0 0.82 0.90 0.86 1471

1 0.70 0.55 0.62 629

accuracy 0.79 2100

macro avg 0.76 0.72 0.74 2100

weighted avg 0.79 0.79 0.79 2100

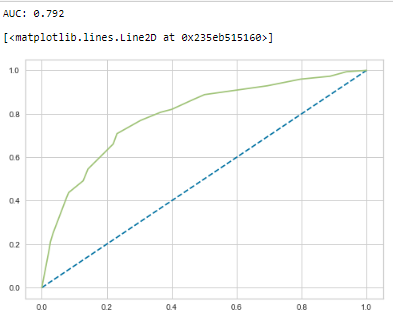


Fig 22. Area Under the Curve for CART Test data

Accuracy: 0.76

Classification Report:

precision recall f1-score support

0 0.77 0.92 0.84 605

1 0.72 0.44 0.54 295

accuracy 0.76 900

macro avg 0.75 0.68 0.69 900

weighted avg 0.75 0.76 0.74 900

# Cart Observation:

### Train Data: Test Data:

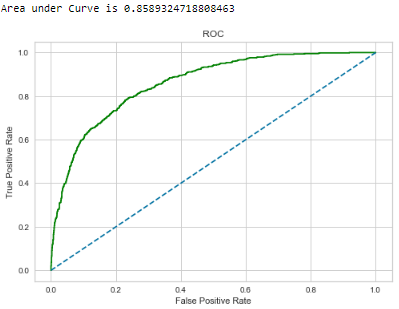
- AUC: 82% - AUC: 80%

- Accuracy: 79% - Accuracy: 77%

- Precision: 70% - Precision: 80%

- f1-Score: 60% - f1-Score: 84%

It is observed that the training and test data are almost similar, and with the overall measures high, the model is a good model.

**AUC\_ROC\_RF:**

*Fig 23. Area under the curve for Random Forest Train data*

Accuracy: 0.82

Classification\_report:

precision recall f1-score support

0 0.84 0.90 0.87 1471

1 0.72 0.61 0.66 629

accuracy 0.81 2100

macro avg 0.78 0.75 0.77 2100

weighted avg 0.81 0.81 0.81 2100

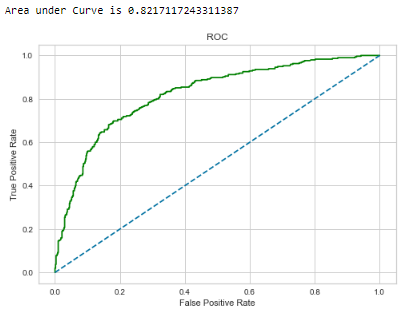


Fig 24. Area under the curve for Random Forest Test data

Accuracy: 0.78

Classification Report:

precision recall f1-score support

0 0.79 0.91 0.85 605

1 0.73 0.51 0.60 295

accuracy 0.78 900

macro avg 0.76 0.71 0.72 900

weighted avg 0.77 0.78 0.76 900

### Train Data: Test Data:

- AUC: 86% - AUC: 82%

- Accuracy: 80% - Accuracy: 78%

- Precision: 72% - Precision: 68%

- f1-Score: 66% - f1-Score: 62%

### The model seems to be a good model as the train and test data results are almost similar.

**AUC\_ROC\_ANN:**

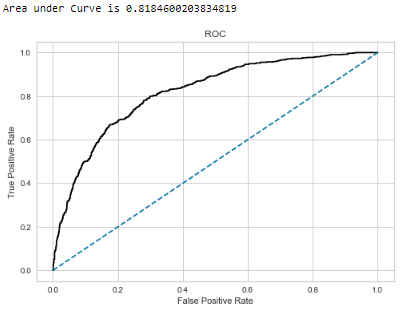


Fig 25. Area under the curve for ANN Train data

Accuracy: 0.78

Classification Report:

precision recall f1-score support

0 0.81 0.89 0.85 1471

1 0.67 0.51 0.58 629

accuracy 0.78 2100

macro avg 0.74 0.70 0.71 2100

weighted avg 0.77 0.78 0.77 2100

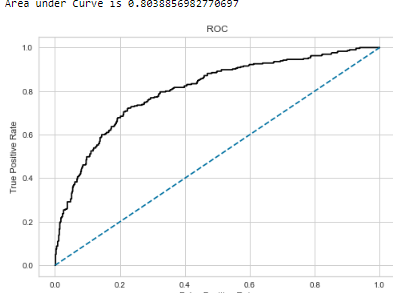


Fig 26. Area under the curve for ANN Test data

Accuracy: 0.76

Classification Report:

precision recall f1-score support

0 0.77 0.92 0.84 605

1 0.72 0.43 0.54 295

accuracy 0.76 900

macro avg 0.74 0.67 0.69 900

weighted avg 0.75 0.76 0.74 900

Observation:

### Train Data: Test Data:

- AUC: 82% - AUC: 80%

- Accuracy: 78% - Accuracy: 77%

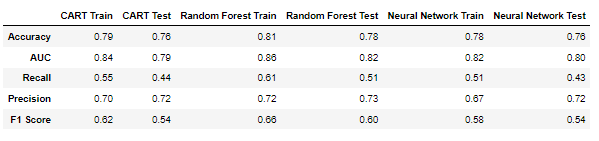
- Precision: 68% - Precision: 67%

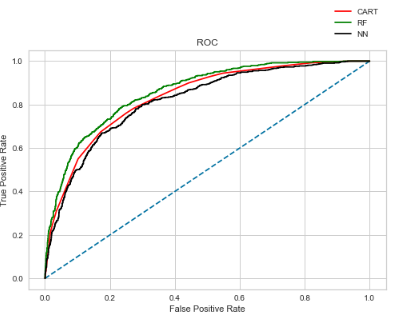
- f1-Score: 59 - f1-Score: 57%

The model seems to be a good model as the training and test result sets are almost similar.

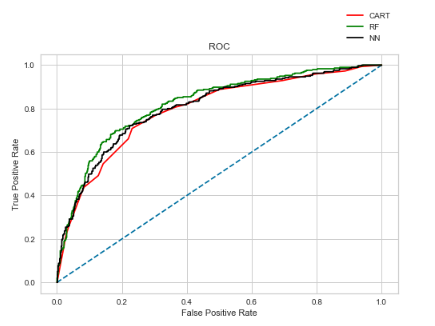
**2.4 Final Model: Compare all the models and write an inference which model is best/optimized.**

Comparing all the values among CART – RF – ANN

*Fig 27. DataFrame CART-RF-ANN values*

ROC Curve:

*Fig 28. ROC Curve for the 3 models on trained data*

**

*Fig 29. ROC Curve for the 3 models on test data*

Random Forest values and curve seem to have better performance on the train and test data when compared to the other two models – Accuracy, AUC, Recall, Precision and f-1 score.

Hence we can conclude that Random Forest is the preferred method for analysing this dataset.

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

Business insights:

Customers preferred online transactions when compared to offline.

JZI agency has very less sales when compared to other agencies.

Airline insurance has more sales than travel agency.

Gold Plan customers have more unclaimed insurance than any other plan.

Asia has more sales when compared to Europe and Americas.

The accuracy is around 80% for the given dataset, so if we include more historical data, then we might arrive at more insights.

Recommendations:

More agencies and many other variables like time, age group, location, vehicle types can be included for analysis.

Customer satisfaction index could be initiated half-yearly to find out opinions from customers.

Claim costs can be reduced to attract more customers.

Increasing advertisements may improve the offline customer base.

More real time data or historical data could help understand and analyse the data better.