**Business Report**

**Data Mining Project**

**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.

**1.1** Read the data and do exploratory data analysis. Describe the data briefly.

1. All column are integers
2. No null values
3. No duplicate values
4. All the variables are almost normally distributed
5. Only 2 variables have outliers - Probability of full payment and min\_payment\_amt

**Measure of Central Tendency**

Mean value of spending : 14.847523809523818

Median value spending : 14.355

Mean value of advance\_payments : 14.559285714285727

Median value advance\_payments : 14.32

Mean value of probability\_of\_full\_payment : 0.8709985714285714

Median value probability\_of\_full\_payment : 0.8734500000000001

Mean value of current\_balance : 5.628533333333334

Median value current\_balance : 5.5235

Mean value of credit\_limit : 3.258604761904763

Median value credit\_limit : 3.237

Mean value of min\_payment\_amt : 3.7002009523809507

Median value min\_payment\_amt : 3.599

Mean value of max\_spent\_in\_single\_shopping : 5.408071428571429

Median value max\_spent\_in\_single\_shopping : 5.223000000000001

**Measure of Dispersion**

Minimum value of spending : 10.59

Maximum Age of spending : 21.18

Range of values of spending : 10.59

Standard deviation of spending : 2.909699430687361

Minimum value of advance\_payments : 12.41

Maximum Age of advance\_payments : 17.25

Range of values of advance\_payments : 4.84

Standard deviation of advance\_payments : 1.305958726564022

Minimum value of probability\_of\_full\_payment : 0.8081

Maximum Age of probability\_of\_full\_payment : 0.9183

Range of values of probability\_of\_full\_payment : 0.11019999999999996

Standard deviation of probability\_of\_full\_payment : 0.023629416583846496

Minimum value of current\_balance : 4.899

Maximum Age of current\_balance : 6.675

Range of values of current\_balance : 1.7759999999999998

Standard deviation of current\_balance : 0.4430634777264493

Minimum value of credit\_limit : 2.63

Maximum Age of credit\_limit : 4.033

Range of values of credit\_limit : 1.4030000000000005

Standard deviation of credit\_limit : 0.3777144449065874

Minimum value of min\_payment\_amt : 0.7651

Maximum Age of min\_payment\_amt : 8.456

Range of values of min\_payment\_amt : 7.690899999999999

Standard deviation of min\_payment\_amt : 1.5035571308217792

Minimum value of max\_spent\_in\_single\_shopping : 4.519

Maximum Age of max\_spent\_in\_single\_shopping : 6.55

Range of values of max\_spent\_in\_single\_shopping : 2.0309999999999997

Standard deviation of max\_spent\_in\_single\_shopping : 0.4914804991024054

**Quantiles**

Quantile 25% spending : 12.27

Quantile 50% spending : 14.355

Quantile 75% spending : 17.305

IQR spending : 5.035

Quantile 25% advance\_payments : 13.45

Quantile 50% advance\_payments : 14.32

Quantile 75% advance\_payments : 15.715

IQR advance\_payments : 2.2650000000000006

Quantile 25% probability\_of\_full\_payment : 0.8569

Quantile 50% probability\_of\_full\_payment : 0.8734500000000001

Quantile 75% probability\_of\_full\_payment : 0.887775

IQR probability\_of\_full\_payment : 0.030874999999999986

Quantile 25% current\_balance : 5.26225

Quantile 50% current\_balance : 5.5235

Quantile 75% current\_balance : 5.97975

IQR current\_balance : 0.7175000000000002

Quantile 25% credit\_limit : 2.944

Quantile 50% credit\_limit : 3.237

Quantile 75% credit\_limit : 3.56175

IQR credit\_limit : 0.61775

Quantile 25% min\_payment\_amt : 2.5614999999999997

Quantile 50% min\_payment\_amt : 3.599

Quantile 75% min\_payment\_amt : 4.76875

IQR min\_payment\_amt : 2.20725

Quantile 25% max\_spent\_in\_single\_shopping : 5.045

Quantile 50% max\_spent\_in\_single\_shopping : 5.223000000000001

Quantile 75% max\_spent\_in\_single\_shopping : 5.877000000000001

IQR max\_spent\_in\_single\_shopping : 0.8320000000000007

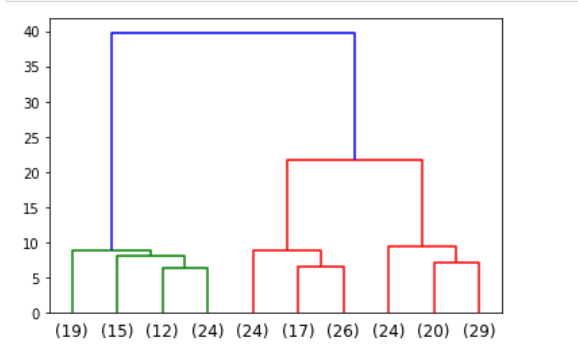
**1.2** Do you think scaling is necessary for clustering in this case? Justify

Yes scaling is important for clustering. Any distance based methods require scaling so that they are not biased towards a variable because of high numeric value

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

The number of optimum clusters using Dendrogram = 3

Dendogram for last 10 merges



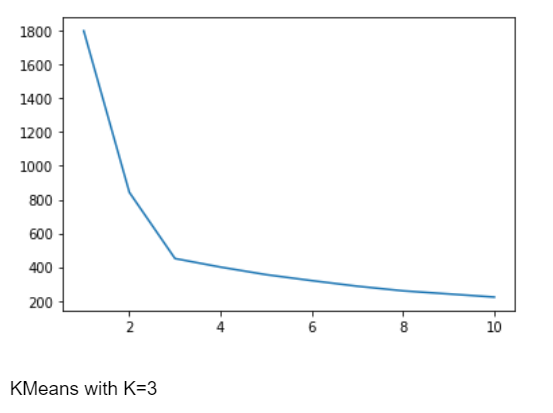
Hierarchical clustering file



**1.4** Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

Optimum clusters = 3

Elbow curve



Silhoutte score = 0.5019103623310144

Kmeans clustered file



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

Cluster 1 : Financially good, clears dues in time, high income, use credit limit at once

Cluster 2 : Financilly unstable, do no clear dues on time, low income, do not use their credit limit at once

Cluster 3 : Moderate spenders, clearing dues and using credit limit lies between cluster 1 cluster 2.

**Promotional strategies**

Cluster 1 :

As these are high income people with high spending power, they can be targeted to for expensive assets,etc.

Cluster 2 :

This groups does not use their credit limit that much , so they can given promotional offers, emi schemes so that it becomes easier for them to pay off their credits. This will also make them more frequent customers

Cluster 3:

This group is at moderate level neither do they spend high do they spend less.

They can be offered cashbacks so that try varied services.

**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** Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.

1. 10 variables
2. Age, Commision, Duration, Sales are numeric variable
3. Rest are categorial variables
4. 3000 records, no missing one
5. 9 independant variable and one target variable – Claimed
6. No missing values
7. Number of duplicate rows = 139

**Unique values in each column**

Agency\_Code : 4

JZI 239

CWT 471

C2B 913

EPX 1238

Type : 2

Airlines 1152

Travel Agency 1709

Claimed : 2

Yes 914

No 1947

Channel : 2

Offline 46

Online 2815

Product Name : 5

Gold Plan 109

Silver Plan 421

Cancellation Plan 615

Bronze Plan 645

Customised Plan 1071

Destination : 3

EUROPE 215

Americas 319

ASIA 2327

**Measure of Central Tendency**

Mean value of Age : 38.204124432016776

Median value Age : 36.0

Mean value of Commision : 15.080996155190423

Median value Commision : 5.63

Mean value of Duration : 72.12023767913317

Median value Duration : 28.0

Mean value of Sales : 61.757878364208416

Median value Sales : 33.5

**Measure of Dispersion**

Minimum value of Age : 8

Maximum Age of Age : 84

Range of values of Age : 76

Standard deviation of Age : 10.678106340128144

Minimum value of Commision : 0.0

Maximum Age of Commision : 210.21

Range of values of Commision : 210.21

Standard deviation of Commision : 25.826834110304418

Minimum value of Duration : -1

Maximum Age of Duration : 4580

Range of values of Duration : 4581

Standard deviation of Duration : 135.9771996496429

Minimum value of Sales : 0.0

Maximum Age of Sales : 539.0

Range of values of Sales : 539.0

Standard deviation of Sales : 71.39974021475756

**Quantiles**

Quantile 25% Age : 31.0

Quantile 50% Age : 36.0

Quantile 75% Age : 43.0

IQR Age : 12.0

Quantile 25% Commision : 0.0

Quantile 50% Commision : 5.63

Quantile 75% Commision : 17.82

IQR Commision : 17.82

Quantile 25% Duration : 12.0

Quantile 50% Duration : 28.0

Quantile 75% Duration : 66.0

IQR Duration : 54.0

Quantile 25% Sales : 20.0

Quantile 50% Sales : 33.5

Quantile 75% Sales : 69.3

IQR Sales : 49.3

**Skewness**

Duration is highly skewed

Commission is also skewed

Age and Sales is slightly closer to normal distribution

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

**Dimensions after splitting data**

x\_train (2002, 9)

x\_test (859, 9)

train\_labels (2002,)

test\_labels (859,)

**Decision Tree Classifier**

**Best** **grid**

DecisionTreeClassifier(max\_depth=4, min\_samples\_leaf=30, min\_samples\_split=200)

**Dot file**



**Variable Importance**

Imp

Agency\_Code 0.602603

Sales 0.306060

Product Name 0.047526

Commision 0.035850

Duration 0.007961

Age 0.000000

Type 0.000000

Channel 0.000000

Destination 0.000000

**Random Forest Classifier**

**Best grid**

RandomForestClassifier(max\_depth=7, max\_features=5, min\_samples\_leaf=7,

min\_samples\_split=55, n\_estimators=300, random\_state=1)

**Variable Importance**

Imp

Agency\_Code 0.326269

Sales 0.199466

Product Name 0.198487

Commision 0.094315

Duration 0.079026

Age 0.063554

Type 0.022940

Destination 0.013399

Channel 0.002545

**Neural Network Classifier**

**Best grid**

'activation': 'relu',

'hidden\_layer\_sizes': 230,

'max\_iter': 2500,

'solver': 'adam',

'tol': 0.01

MLPClassifier(hidden\_layer\_sizes=230, max\_iter=2500, random\_state=1, tol=0.01)

**2.3** Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model

**Cart Performance Metrics**

**Train Data:**

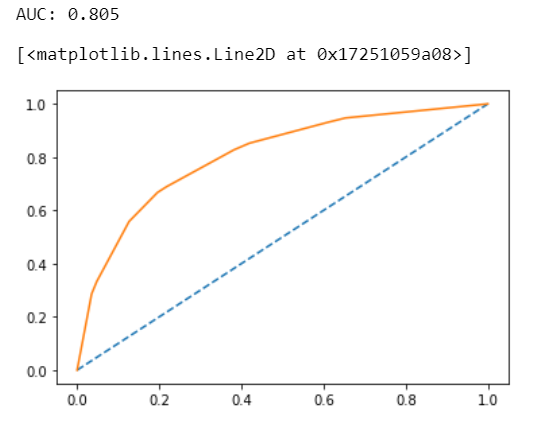
AUC: 80.5%

Accuracy: 77.22%

Precision: 68%

f1-Score: 61%

ROC\_AUC Graph



**Test Data:**

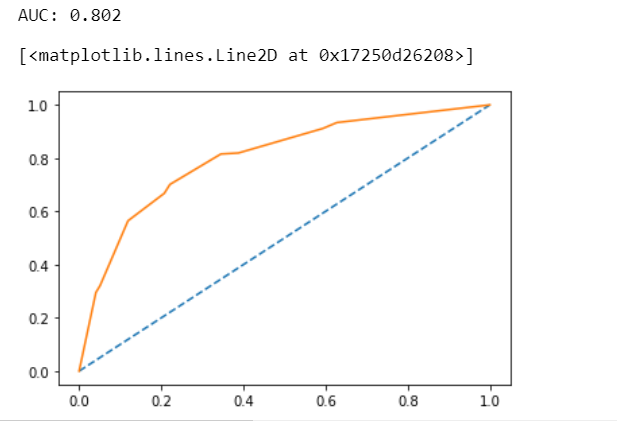
AUC: 80.2%

Accuracy: 78.11%

Precision: 69%

f1-Score: 62%

ROC\_AUC Graph



**Random Forest Classifier Performance Metrics**

**Train Data:**

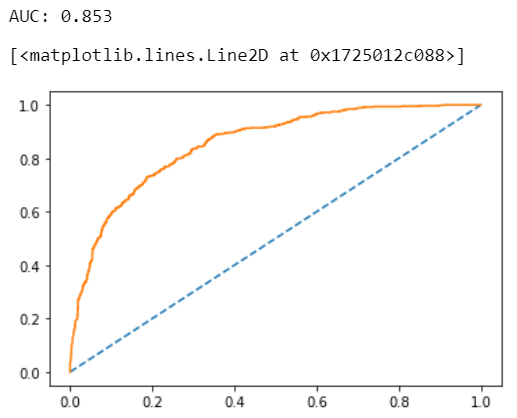
AUC: 85.3%

Accuracy: 80.02%

Precision: 68%

f1-Score: 61%

ROC\_AUC Graph



**Test Data:**

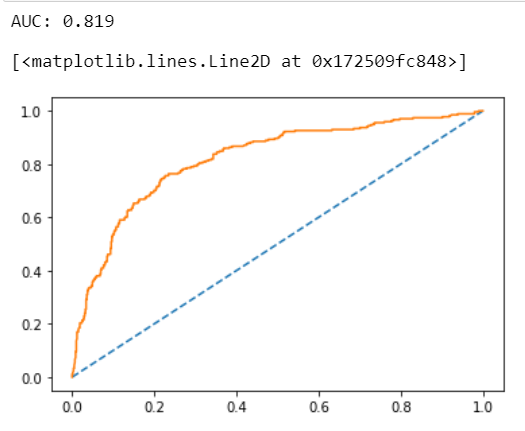
AUC: 81.9%

Accuracy: 78%

Precision: 69%

f1-Score: 62%

ROC\_AUC Graph



**Neural Network Classifier Performance Metrics**

**Train Data:**

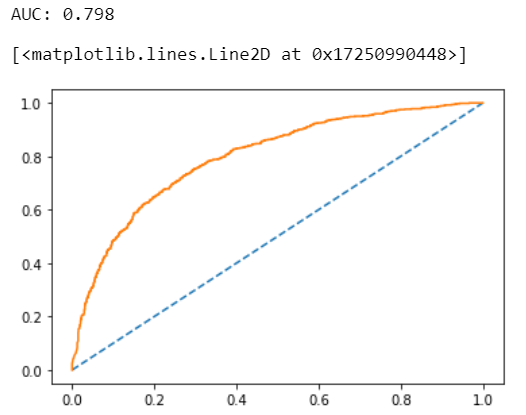
AUC: 79%

Accuracy: 77%

Precision: 68%

f1-Score: 61%

ROC\_AUC Graph



**Test Data:**

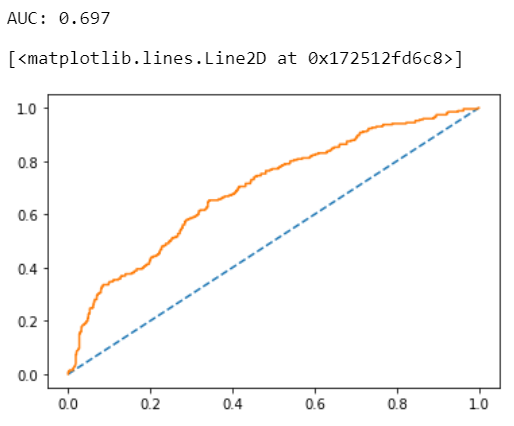
AUC: 69.7%

Accuracy: 78%

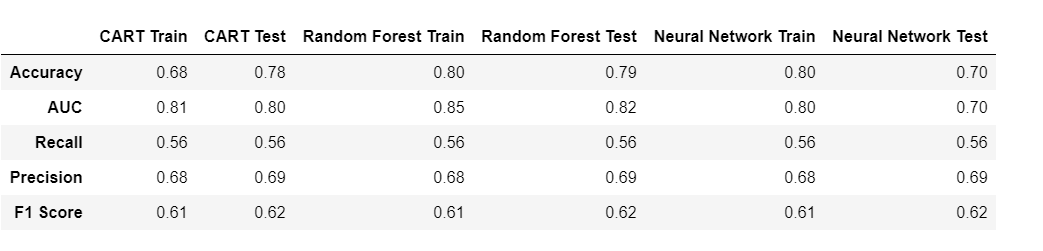
Precision: 69%

f1-Score: 62%

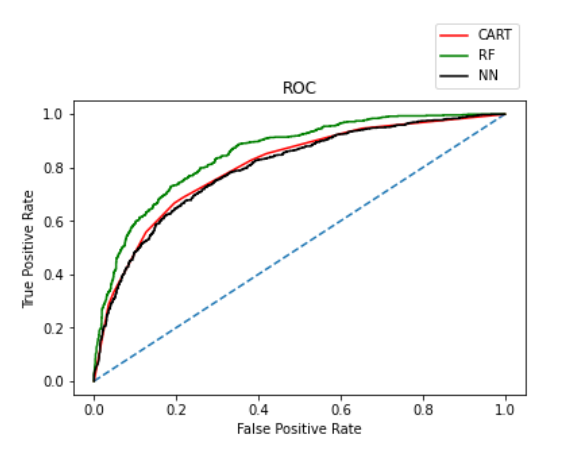
ROC\_AUC Graph



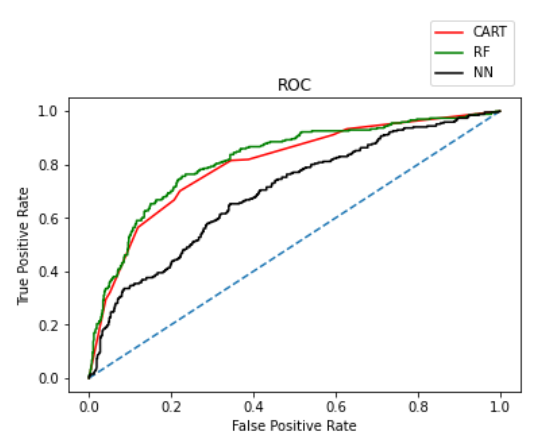
**2.4** Final Model: Compare all the model and write an inference which model is best/optimized.



ROC \_AUC graph for training data



**ROC \_AUC graph for testing data**



I am selecting the RF model, as it has better accuracy, precision, recall, f1 score better than other two CART & NN

**2.5** Inference: Basis on these predictions, what are the business insights and recommendations

Online experiences benefitted customers

90% of insurance is done by online channel

The JZI agency resources to pick up sales as they are in bottom

Also based on the model we are getting 80%accuracy, so we need customer books airline tickets based on the claim data pattern.

More sales happen via Agency than Airlines and the trend shows the claim are processed more at Airline

Increase customer satisfaction