Data Mining Project

Problem 1

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.

	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

The Data has 7 variables

Information on data

```
Data columns (total 7 columns):
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
dtypes: float64(7)
```

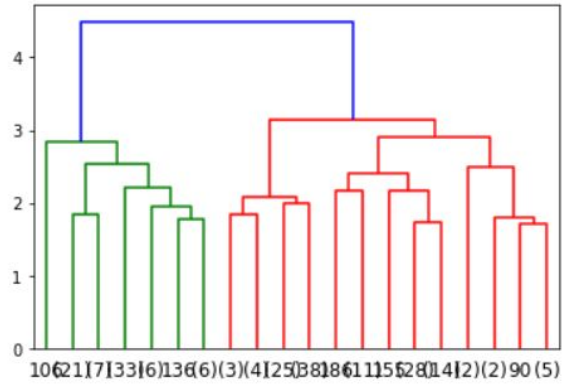
All variables are of float type

There are no null values or duplicates in the dataset

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

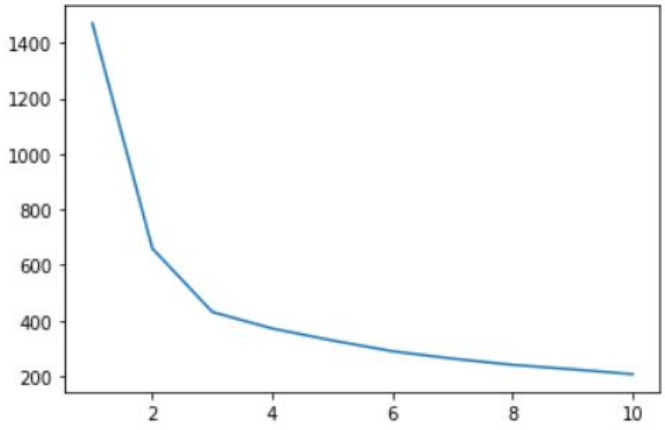
Yes, Scaling is necessary and has been applied to this data as the observations are of varying scales which may affect the results of clustering

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



- Hierarchical clustering was applied and 2 was identified as the optimum number of clusters
- Cluster 1 contains observations of high credit card usage
- cluster 2 contains observations of low credit card usage

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



Elbow Curve

- the optimum number of clusters has been identified as 2 using elbow curve and silhouette score
- Cluster 1 consists of high Credit card usage
- Cluster 2 consists of Low Credit card usage

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

Hierarchical Clustering

- Cluster 1 : High Credit Card Usage (spending high, larger credit limit and current balance)
- cluster 2 : Low credit card usage (spending, payment, current balance are all relatively low)

K Means Clustering

- Cluster 1 : Low Credit Card Usage (except for Minimum payment amount all others are low)
- Cluster 2 : high credit card usage (all the variables are high except for minimum payment amount)

Recommendations

1. Customers in Low credit Card usage can be offered promotions which require smaller credit limit and spending
2. Customers in High Credit Card usage can be offered promotions which are high spending and needs a larger credit limit

Problem 2

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 provides 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.

	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

The dataset has 10 variables

Information about dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
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
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

There are 2 float type, 2 integer type and 6 object type variables in the dataset. There were 139 duplicates in the dataset which were removed. There are no null values in the dataset.

The dataset contains many outliers. CART and Random Forest are not sensitive to outliers but Artificial Neural Networks are sensitive. Since the outliers are not too far out it may not affect ANN

The object type variables have been converted to categorical.

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

The dataset was split into test and training data and classification models CART, RF, and ANN were built on it to predict the Target variable Claim Status

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

The following are the performance measures for all three Classification models

CART

Training Data

- AUC: 81%
- Accuracy: 76%
- Sensitivity: 58%
- precision: 65%
- f1-score: 61%

Test Data

- AUC: 79%
- Accuracy: 78%
- Sensitivity: 60%
- Precision: 68%
- f1-Score: 63%

Training and Test set results are nearly similar indicating no overfitting or underfitting, overall measures are moderate but good enough to predict

Random Forest

Train Data:

- AUC: 70%
- Accuracy: 77%
- Sensitivity: 53%
- Precision: 68%
- f1-Score: 60%

Test Data:

- AUC: 70%
- Accuracy: 77%
- Sensitivity: 52%
- Precision: 69%
- f1-Score: 59%

Training and Test set results are nearly similar, overall measures are moderate, but good enough for predictions

Artificial Neural networks

Training Data

- AUC: 69%
- Accuracy: 75%
- Sensitivity: 49%
- precision: 66%
- f1-score: 56%

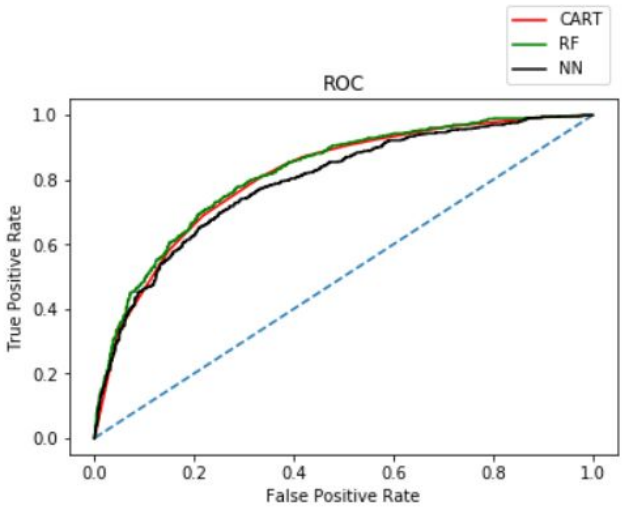
Test Data

- AUC: 70%
- Accuracy: 78%
- Sensitivity: 51%
- Precision: 70%
- f1-Score: 59%

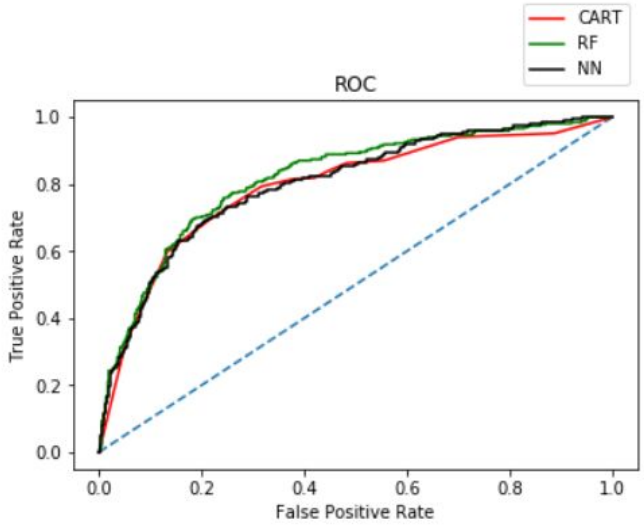
Training and Test set results are almost similar, overall measures are moderate but good for predictions

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

Training Data



Test Data



Random Forest has better performance when compared with CART and Neural network

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

- The model is stable enough to be used for future predictions
- Variable Importance is a feature which can be used to find out which variable is needed to take decisions on claim status
- this model can be used to deduce claim status and prepare in advance according to the customer's data
- I would recommend using the Random Forest Model for future predictions