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.

1.2 Read the data and do exploratory data analysis. Describe the data briefly.

Let us understand the data with the provided data dictionary

Data Dictionary for Market Segmentation:

spending: Amount spent by the customer per month (in 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)

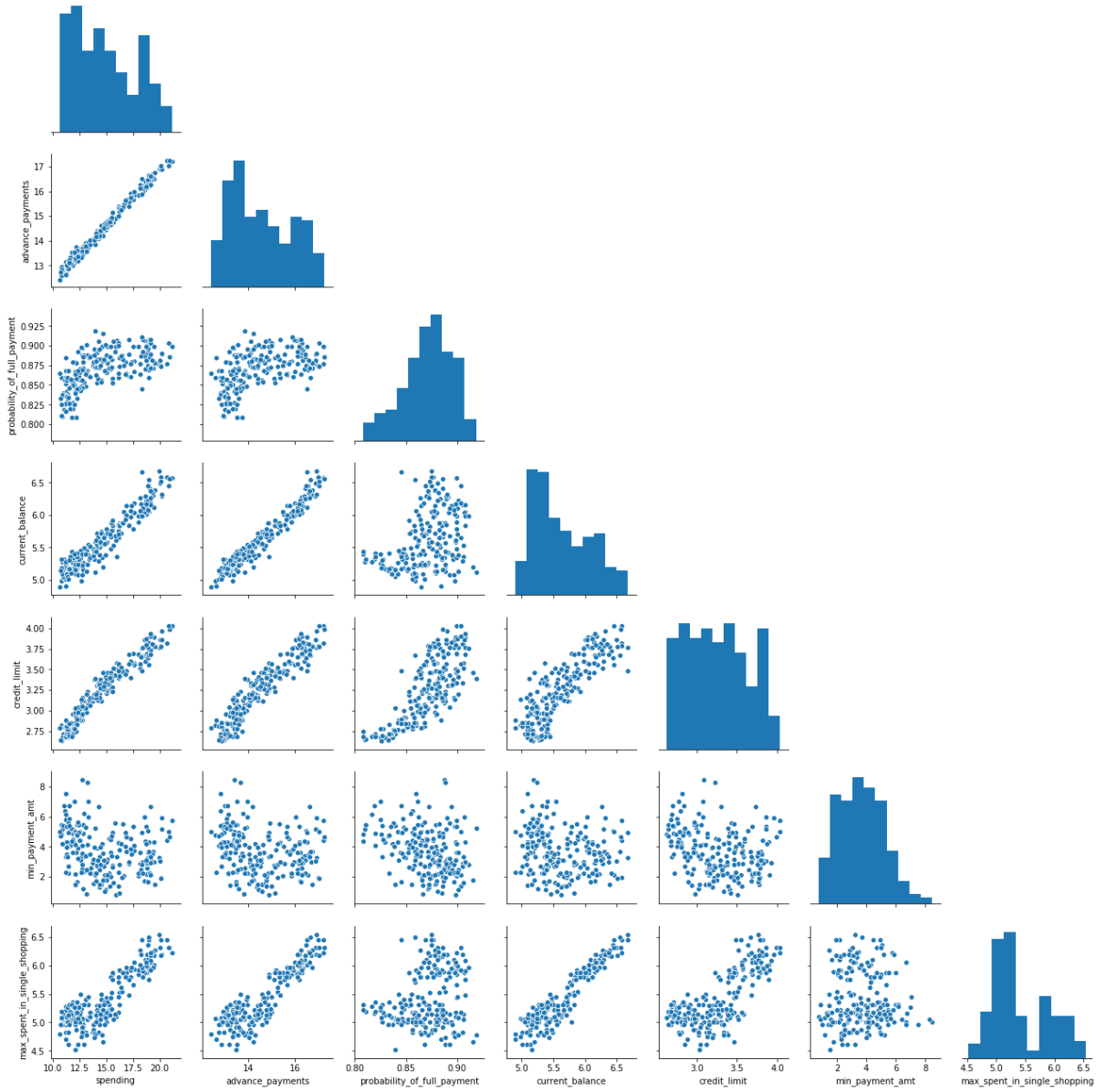
Let us do some exploratory data analysis to get a better sense of our data set.

Get the 5 point statistical summary using the describe function.

	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.5500

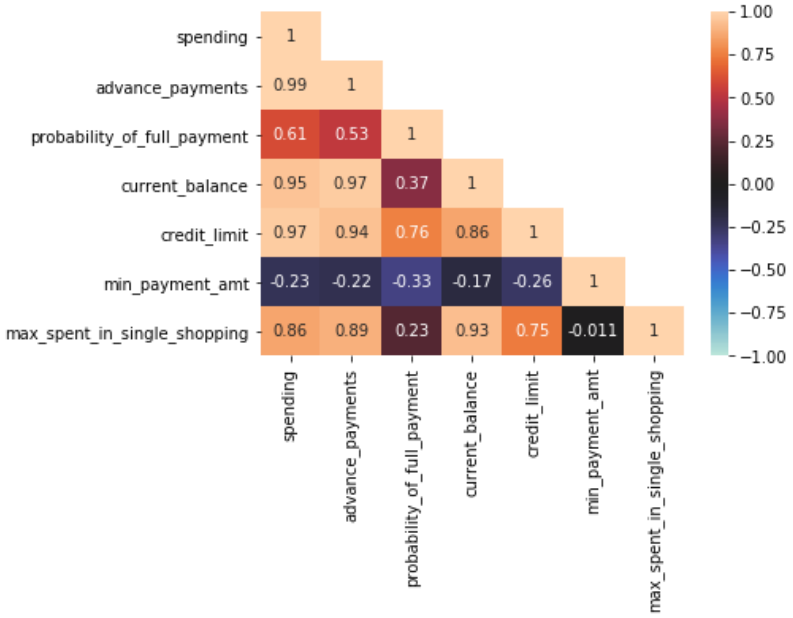
Our dataset set has 210 rows in 7 columns. Each column is of float datatype, so we do not have to perform any type-casting for clustering. There are no null or duplicate rows in our data.

Let us look at the pairplot of our data to get some visual sense of distribution and correlation

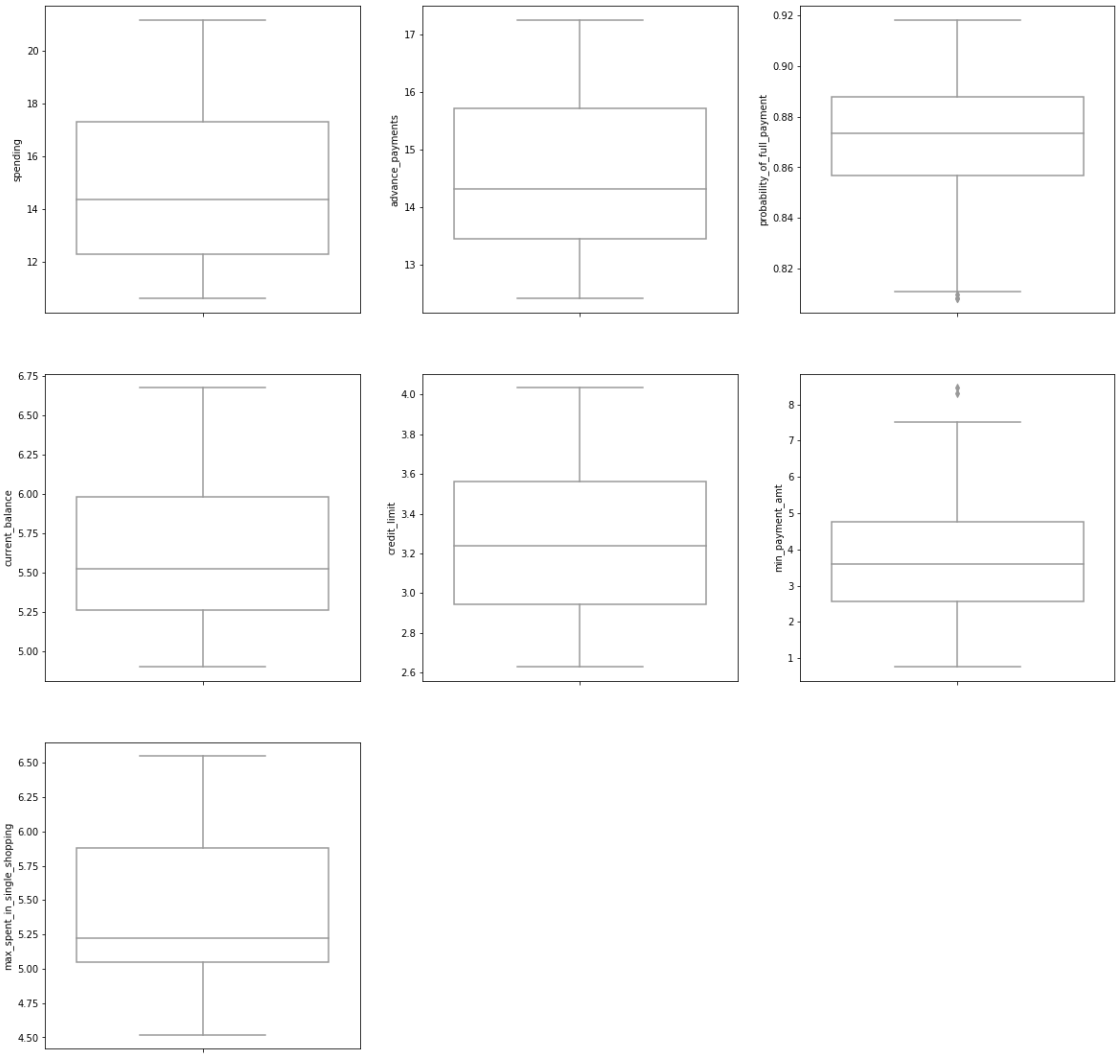


It is clear some variables show strong correlation with each other e.g. **Advance Payments and Spending**, where as a feature like Minimum payment amount does not seem to show any clear correlation with all other variables.

Let us plot the heatmap or correlation to get more statistical understanding about the strength and direction or the correlation.



Let us use boxplots to look for outliers

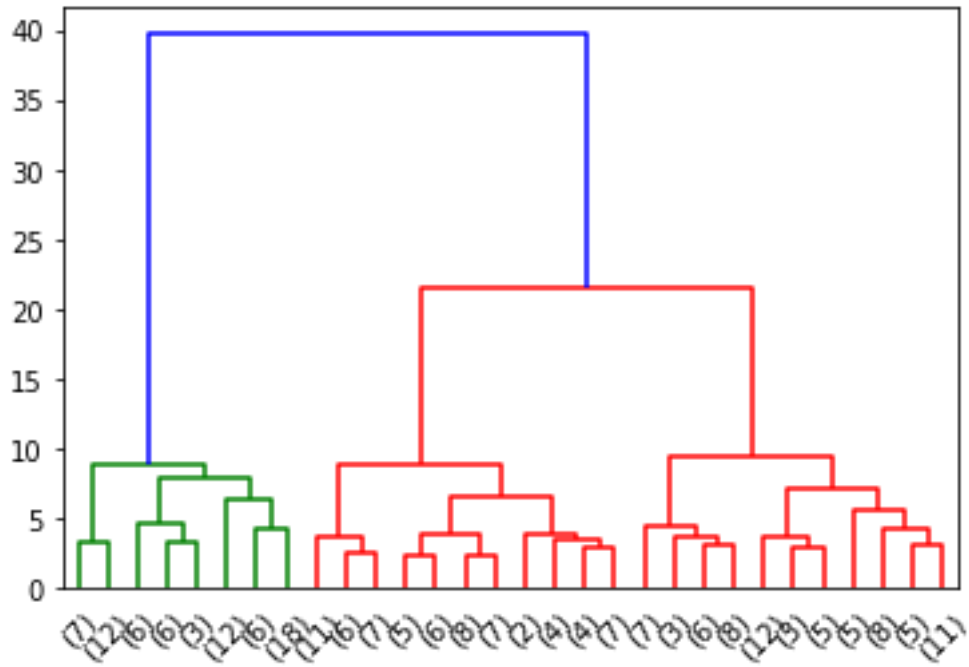


From the plot, we see that 2 features have outliers, **Probability of Full Payment**, and **Minimum payment amount**. Since the outliers are very close to the outlier limit and only a couple of data points are outliers, we will not impute these for now.

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

When working with Clustering, it is important to understand that all clustering algorithms use distance measure of some kind such as Euclidean, Manhattan et al. This distance measure can get impacted significantly when features have different scales. Hence it is important to scale one's data before clustering.

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



Based on this dendrogram, we will use fcluster function and distance criterion to select arrive at the number of clusters. Looking at the figure above, 10 looks a reasonable distance value to arrive at total clusters.

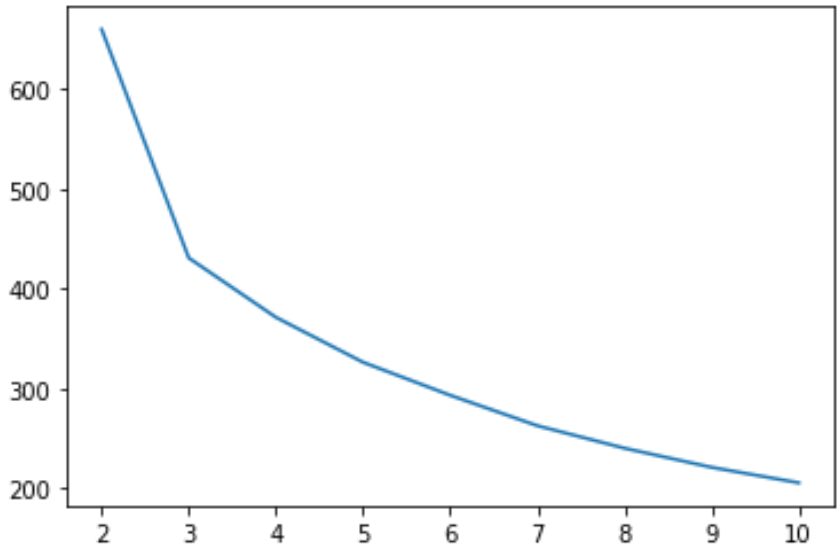
On running the function, we get a total of 3 clusters for our dataset.

We will further validate this when we create clusters using k-means and evaluate for best k.

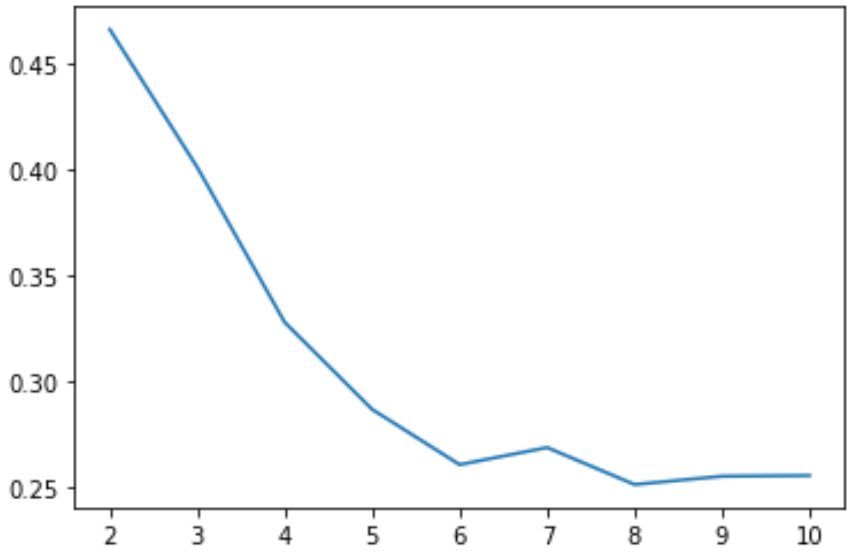
1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

To determine the optimum number of clusters, we ran a loop to capture within-cluster sum-of-squares (WSS) and the corresponding Silhouette Score to check for the ideal number of clusters.

Below is the elbow curve for WSS. We van see a clear plateauing after 3 clusters.

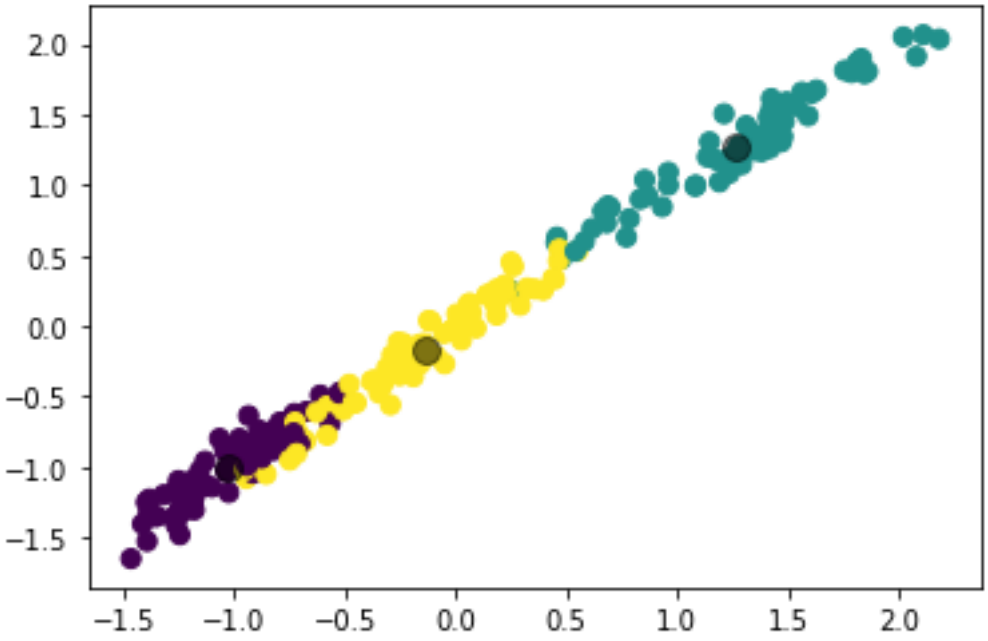


The corresponding plot for the Silhouette Scores is given below.



While the silhouette score for 2 clusters is slightly better, in real word application, 2 cluster may be too few. But this will depend on the actual business application.

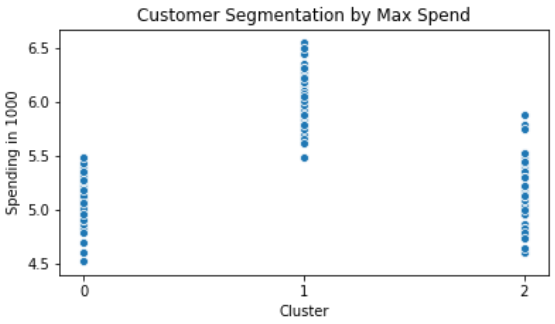
Here is a visualization of the 3 clusters



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

The following four graphs give an interesting insight.

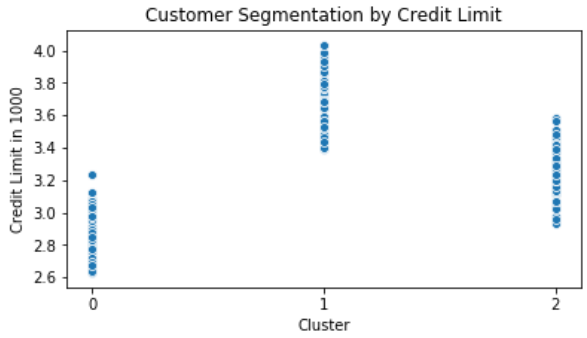
The first two graphs show customer segmentation by overall spending and by maximum spend in a single purchase, respectively.

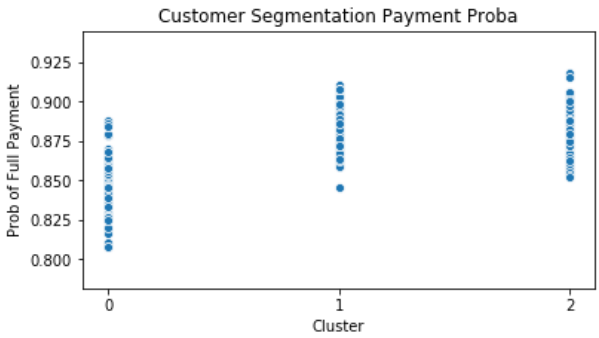


Based on this I can call these clusters as

Low (Cluster 0), High (Cluster 1) and Medium (Cluster 2) spenders.

The next two graphs help analyse the credit limit offered by the bank to these groups and the probability of full payment for each segment.





The mean credit limit (in 10000s) for each segment is as follows

Low Spenders : 2.84
Medium Spenders : 3.25
High Spenders : 3.69

The mean Probability of Full Payment is as below

Low Spenders : 84%
Medium Spenders : 88%
High Spenders : 88%

Based on these insights, the bank should look at increasing the credit limits especially for the segment of medium spenders. Credit limit shows a strong correlation with spending as we've seen in the pair-plot and heatmap.

The probability of full payment for Medium and High Spenders is nearly same, that also shows there is room for credit limit increase without introducing any additional risk.

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.

The insurance dataset has 10 columns and 3000 rows.

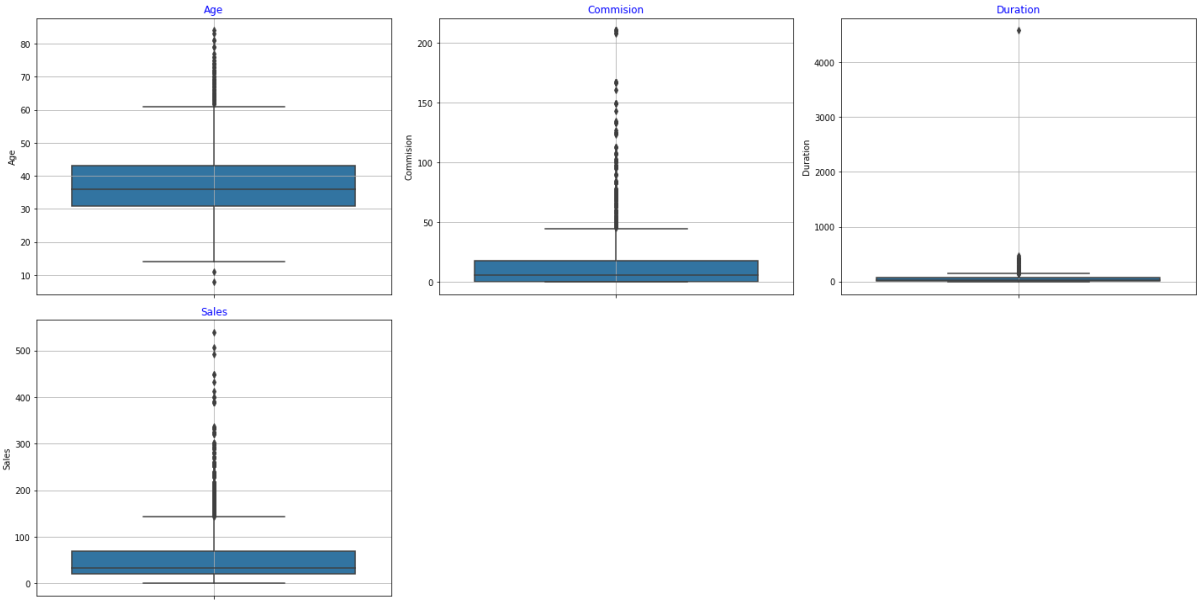
'Claimed' is the dependent classifier that we will predict once our model is trained.

The columns have data type composition as follows: float64(2), int64(2), object(6)

The dataset does not contain any null values, but it did have 139 duplicate rows. For our models to work better, we decided to drop the duplicates keeping the first instance.

The resulting data set had the shape (2861, 10).

All continuous features in our dataset have outliers as indicated by these boxplots.



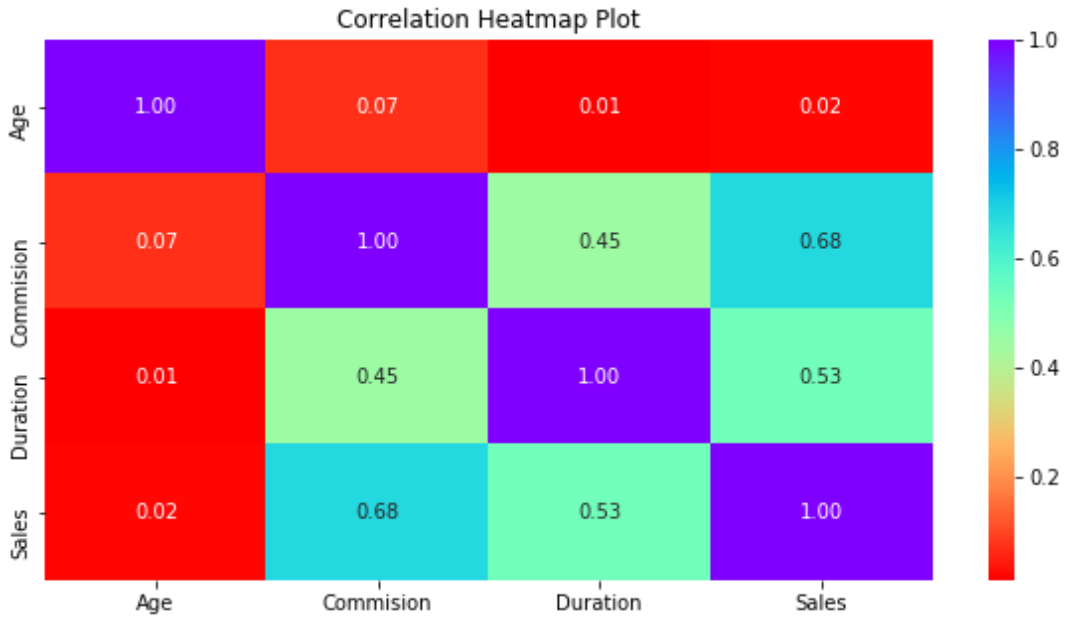
We treated these outliers using the IQR method and imputed whisker values wherever required.

We did a 70:30 split between the Train and Test data sets.

We also did feature scaling on the data. Although CART and Random Forest Classifier models do not require scaling, we used the scaled dataset for all three models as it does not impact the result of tree-based classifiers.

For ANN, scaling was required. We have used StandardScaler function for this purpose.

Let us also look at the correlation heatmap for the dataset



Most of the features do not show strong correlation, which is a good thing for predictive modelling. Sales and Commission show the strongest correlation amongst the features, which is understandable.

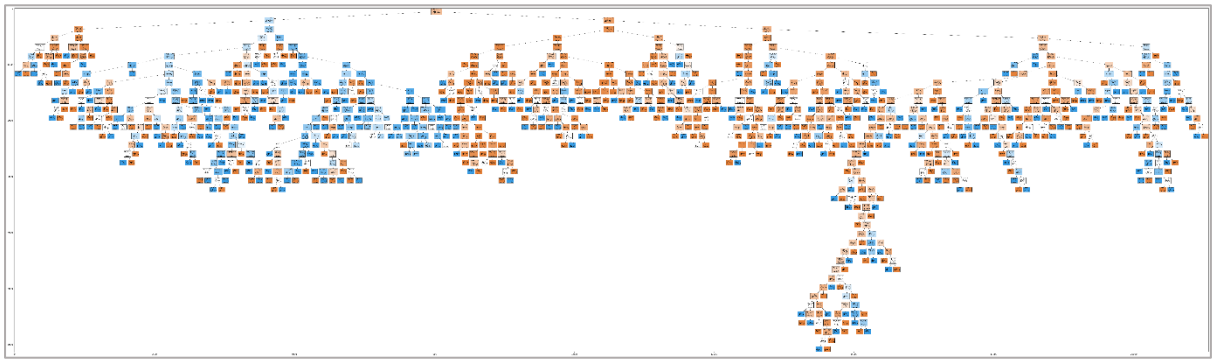
2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network

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

Building the models

Decision Tree Classifier

This is the Tree visualization using Gini criterion and all default parameters



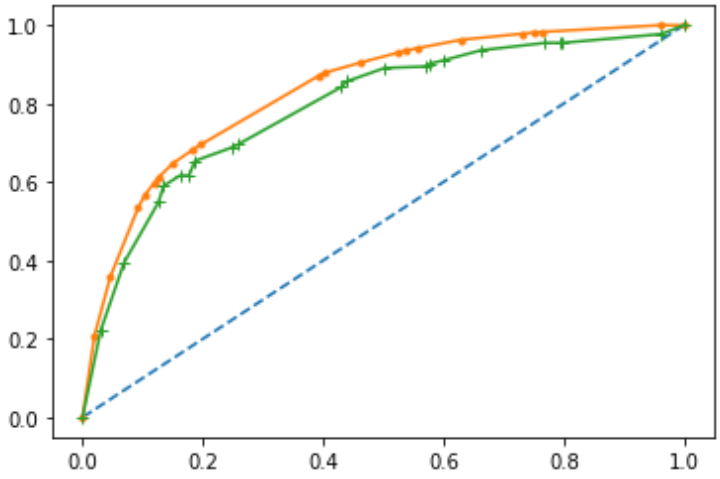
We used Grid Search Cross-Validation for obtaining the best parameter for the CART model.

After a few iterations to fine tune the model, we got the following result for best parameters.

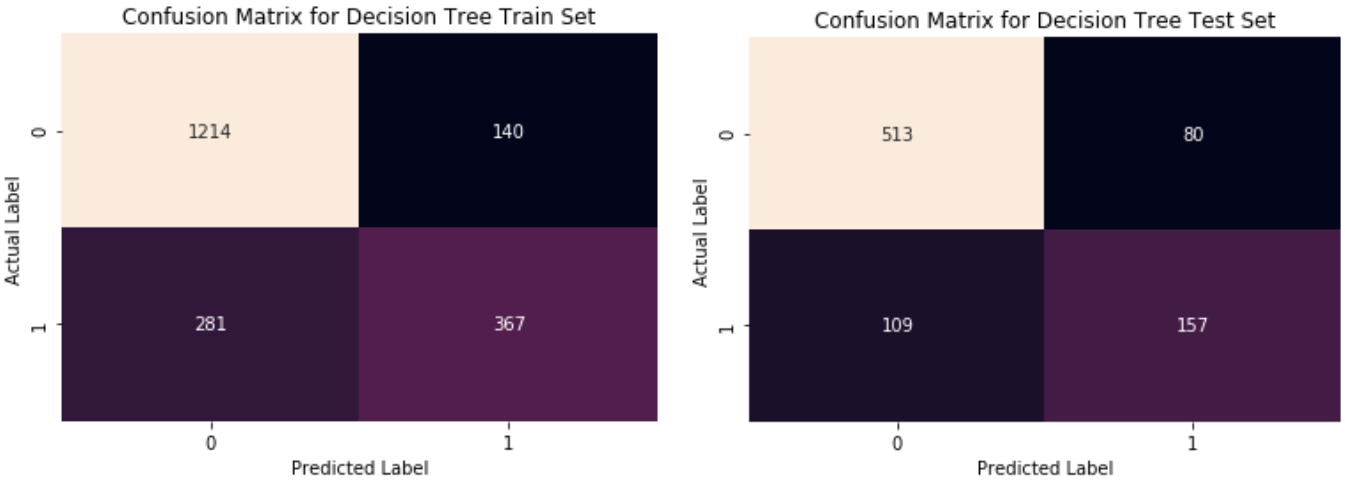
{'max_depth': 5, 'min_samples_leaf': 17, 'min_samples_split': 40}

The model gave AUC values as
AUC Train: 0.833
AUC Test: 0.792

The plot below visualises the ROC Curve. Orange line represents the Train data and Green represents the Test data.



Confusion matrix for Training and Test Data for CART



Classification Report for Training Data

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1354
1	0.72	0.57	0.64	648
accuracy			0.79	2002
macro avg	0.77	0.73	0.74	2002
weighted avg	0.78	0.79	0.78	2002

Classification Report for Test Data

	precision	recall	f1-score	support
0	0.82	0.87	0.84	593
1	0.66	0.59	0.62	266
accuracy			0.78	859
macro avg	0.74	0.73	0.73	859
weighted avg	0.77	0.78	0.78	859

Random Forest Classifier

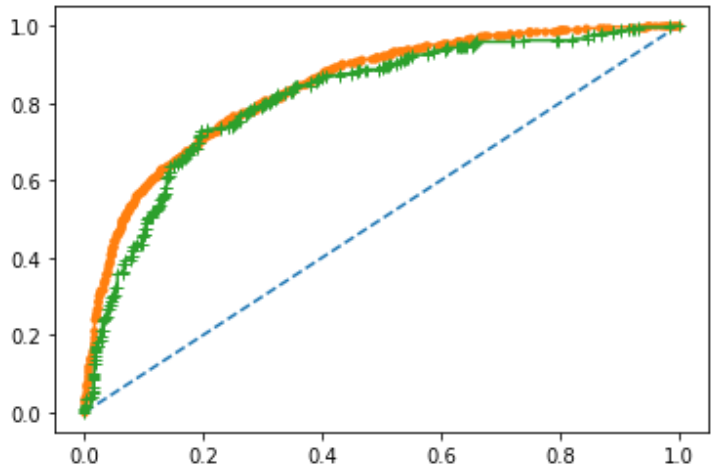
We used Grid Search Cross-Validation for obtaining the best parameter for the Random Forest Classifier.

After a few iterations to fine tune the model, we got the following result for best parameters.

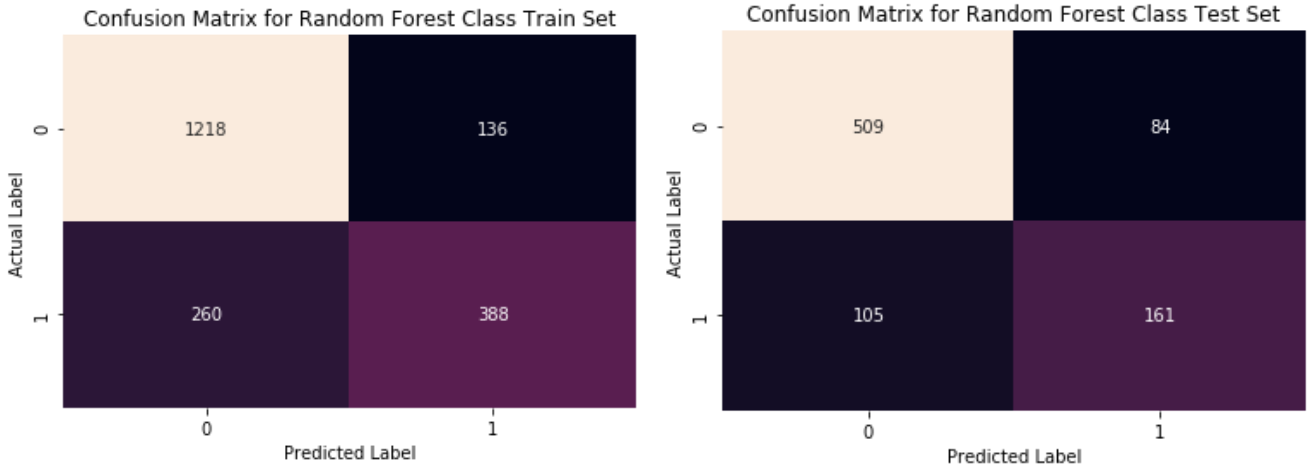
```
{'max_depth': 5,
 'max_features': 7,
 'min_samples_leaf': 7,
 'min_samples_split': 40,
 'n_estimators': 250}
```

The model gave AUC values as
AUC Train: 0.843
AUC Test: 0.815

The plot below visualises the ROC Curve. Orange line represents the Train data and Green represents the Test data.



Confusion matrix for Training and Test Data for Random Forest



Classification Report for Training Data

	precision	recall	f1-score	support
0	0.82	0.89	0.86	1354
1	0.73	0.59	0.65	648
accuracy			0.80	2002
macro avg	0.77	0.74	0.75	2002
weighted avg	0.79	0.80	0.79	2002

Classification Report for Test Data

	precision	recall	f1-score	support
0	0.83	0.86	0.84	593
1	0.66	0.61	0.64	266
accuracy			0.78	859
macro avg	0.75	0.74	0.74	859
weighted avg	0.78	0.78	0.78	859

Artificial Neural Network

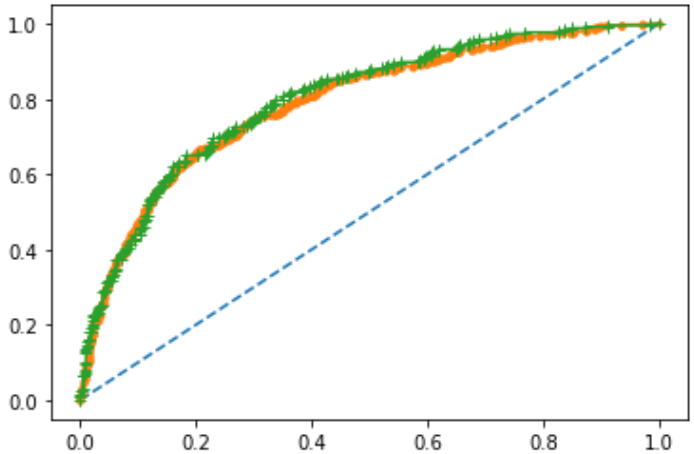
We used Grid Search Cross-Validation for obtaining the best parameter for the ANN Classifier.

After a few iterations to fine tune the model, we got the following result for best parameters.

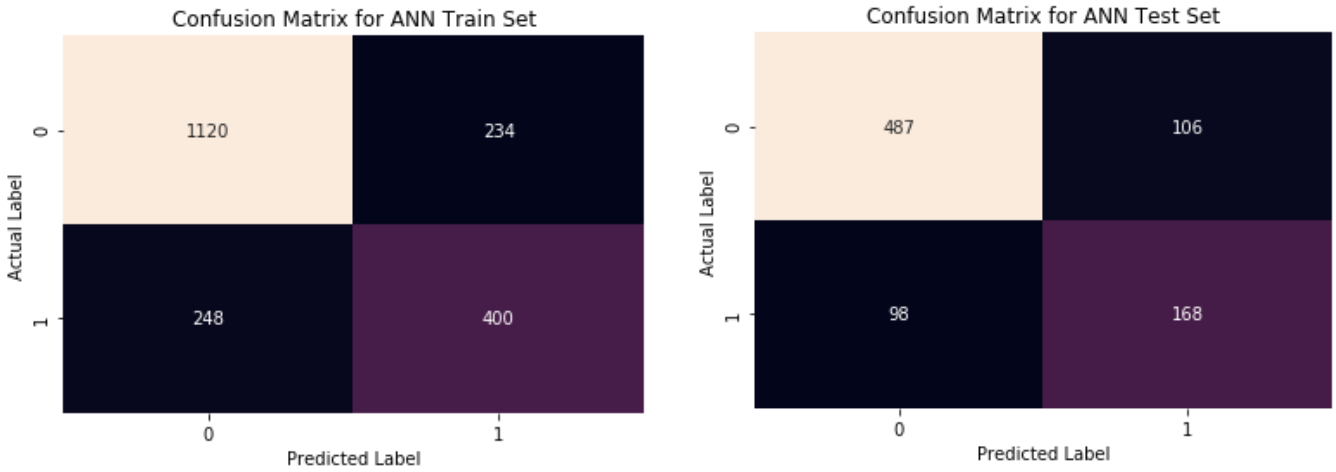
```
{'activation': 'relu',
 'hidden_layer_sizes': 100,
 'learning_rate': 'constant',
 'max_iter': 10000,
 'solver': 'adam',
 'tol': 0.01}
```

The model gave AUC values as
AUC Train: 0.792
AUC Test: 0.801

The plot below visualises the ROC Curve. Orange line represents the Train data and Green represents the Test data.



Confusion matrix for Training and Test Data for ANN



Classification Report for Training Data

	precision	recall	f1-score	support
0	0.82	0.83	0.82	1354
1	0.63	0.62	0.62	648
accuracy			0.76	2002
macro avg	0.72	0.72	0.72	2002
weighted avg	0.76	0.76	0.76	2002

Classification Report for Test Data

	precision	recall	f1-score	support
0	0.83	0.82	0.83	593
1	0.61	0.63	0.62	266
accuracy			0.76	859
macro avg	0.72	0.73	0.72	859
weighted avg	0.76	0.76	0.76	859

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

For model comparison, we will look at the Test performance that gives a better use case indication about real world application.

Since we are looking to predict Claims, we will look at recall score for 1 for all models.

Model	Accuracy	Precision	Recall	AUC
CART	78%	66%	59%	79.2%
Random Forest	78%	66%	61%	81.5%
ANN	76%	63%	62%	80.1%

While ANN shows the highest recall, the overall performance of the Random Forest model is best optimized.
However, we may be able to get better tuning for ANN if we have substantial compute resources available.

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

Using these models, we will be able to predict a potential claim around 60% of the time accurately. Due to the nature of business, a false positive prediction will not have negative impact on the business either.

This prediction can allow the agencies to either evaluate the risk associated or help decide the appropriate value for the premium to be charged from customers.