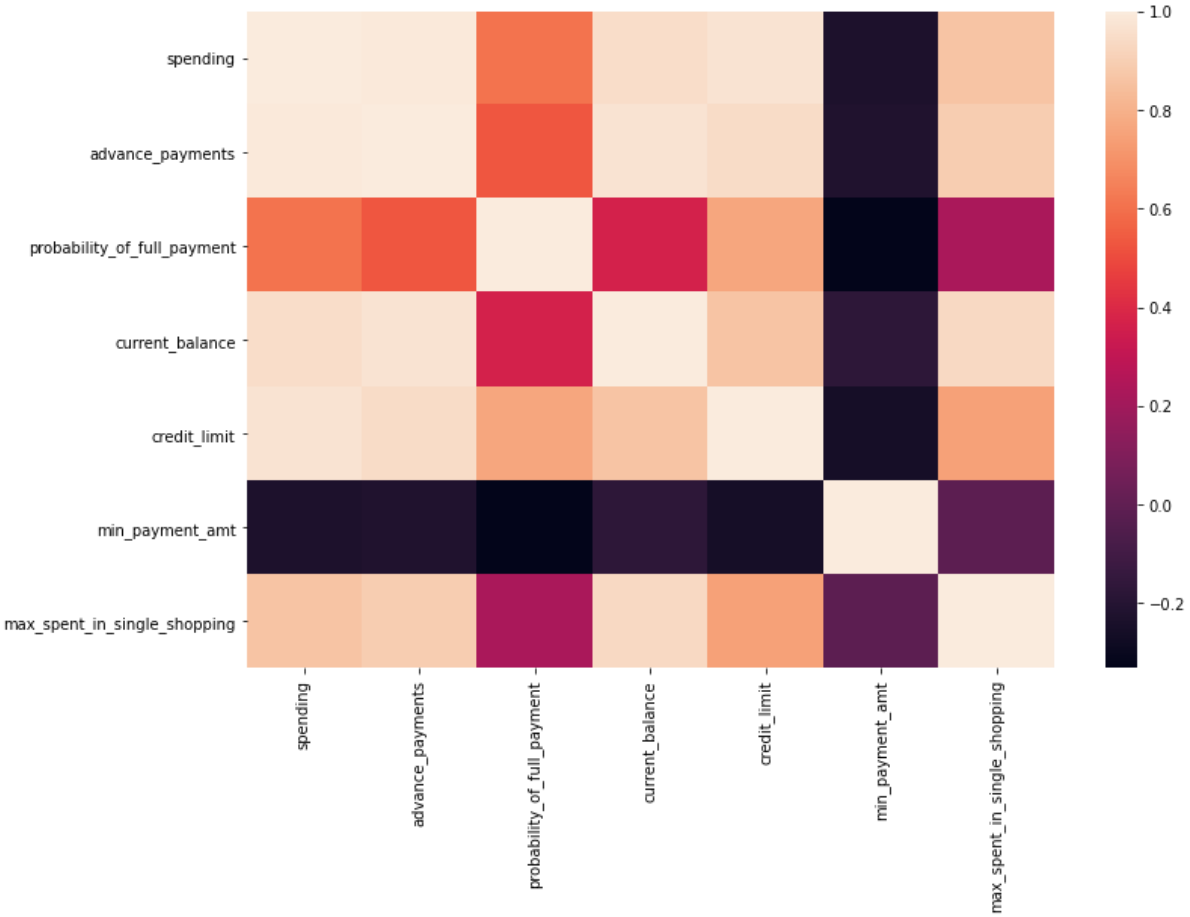
Problem 1: Clustering

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.

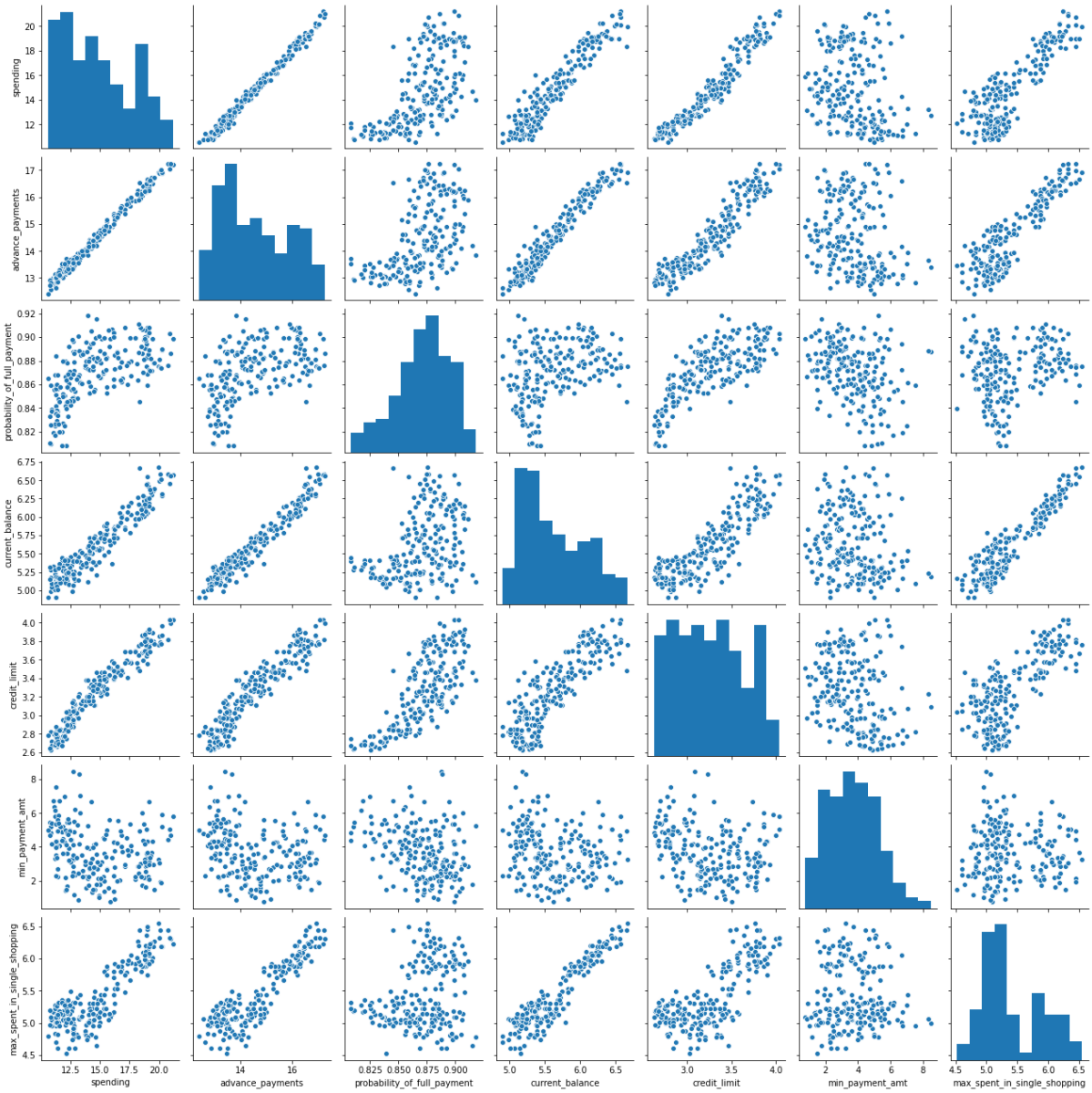
1.1 Read the data and do exploratory data analysis. Describe the data briefly.

Solution

1. The data doesn't contain any null values.
2. The variables show a strong positive and mild negative correlations with each other (correlation heatmap)
    a. Maximum spent in single shopping and minimum payment amount are only negatively correlated.



3. None of the variables are normally distributed as inferred by pairplot. Except:
    'probability of full payment' → slightly negatively skewed.
    'minimum payment amount' → slightly positively skewed.


spending
advance_payments
probability_of_full_payment
current_balance
credit_limit
min_payment_amt
max_spent_in_single_shopping

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

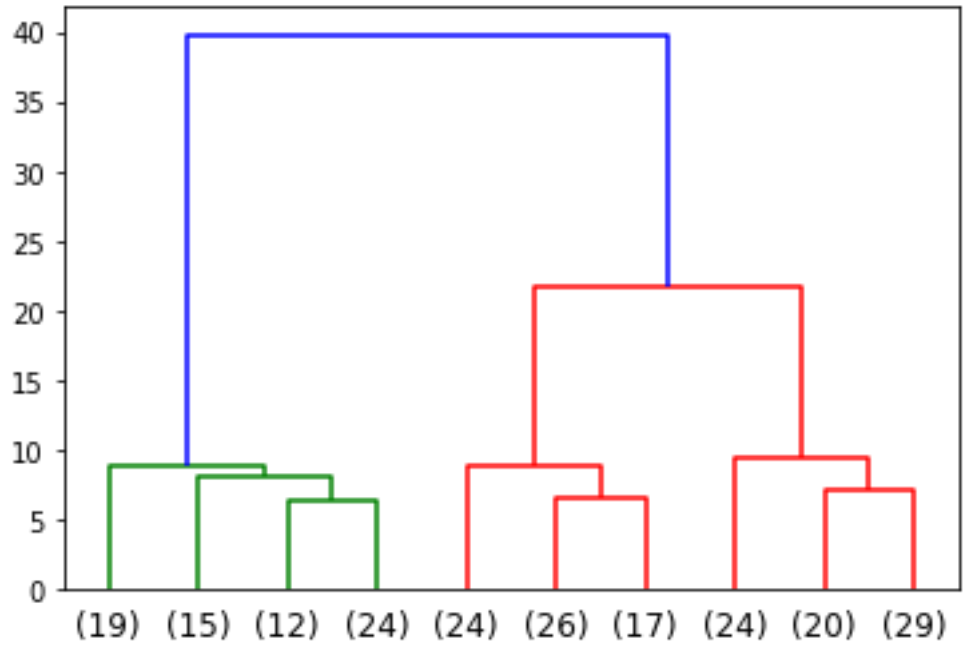
Solution

The Scaling is required for the following reasons:

	count	mean	std	min	25%	50%	75%	max
spending	210	14.848	2.910	10.590	12.270	14.355	17.305	21.180
advance_payments	210	14.559	1.306	12.410	13.450	14.320	15.715	17.250
probability_of_full_payment	210	0.871	0.024	0.808	0.857	0.873	0.888	0.918
current_balance	210	5.629	0.443	4.899	5.262	5.524	5.980	6.675
credit_limit	210	3.259	0.378	2.630	2.944	3.237	3.562	4.033
min_payment_amt	210	3.700	1.504	0.765	2.562	3.599	4.769	8.456
max_spent_in_single_shopping	210	5.408	0.491	4.519	5.045	5.223	5.877	6.550

1. All the variables are at different scales, for example **Current balance** is in multiples of 1000 and advance paid is in multiples of 100.
2. Probability of repayment is a ratio (% figure).

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



The Dendrogram divides the data in 2 distinct clusters.

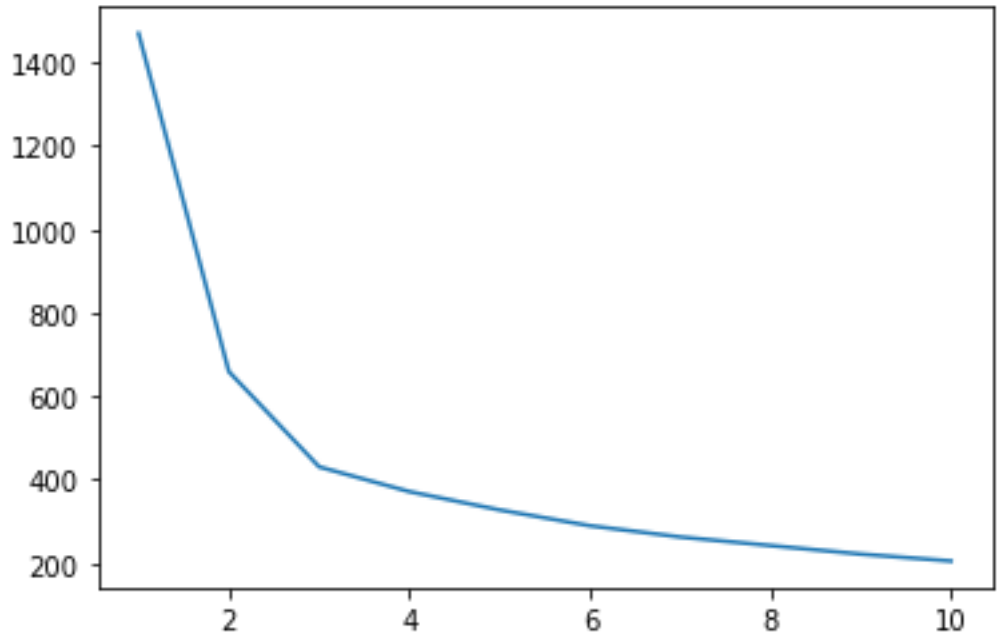
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
Agglo-Clusters							
0	13.09	13.77	0.86	5.36	3.05	3.73	5.1
1	18.37	16.15	0.88	6.16	3.68	3.64	6.02

1. Cluster 1 (labelled as 0):
    a. Characterised as consumer group with lower bank balance, with lower credit limit and lower repayment rate.
    b. This group is also low spenders in single shopping.
    c. Although, they make a higher proportion of advance cash payment as compared to other group/cluster.

2. Cluster 2 (labelled as 1):
    a. Characterised as consumers with relatively higher bank balance and more credit spending.
    b. This group is also entitled to higher credit limit and also spend more on a single shopping.

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 = 3

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Sil_width
KM Cluster								
0	18.5	16.2	0.88	6.18	3.7	3.63	6.04	0.47
1	11.86	13.25	0.85	5.23	2.85	4.74	5.1	0.4
2	14.44	14.34	0.88	5.51	3.26	2.71	5.12	0.34

The total Silhouette Score is 0.400.

The minimum Silhouette Sample is 0.0027. Hence, none of the records are incorrectly mapped to any cluster.

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

Hierarchical Clustering:

The Dendrogram divides the data in 2 distinct clusters.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
Agglo-Clusters							
0	13.09	13.77	0.86	5.36	3.05	3.73	5.1
1	18.37	16.15	0.88	6.16	3.68	3.64	6.02

1. Cluster 1 (labelled as 0):
    a. Characterised as consumer group with lower bank balance, with lower credit limit and lower repayment rate.
    b. This group is also low spenders in single shopping.
    c. Although, they make a higher proportion of advance cash payment as compared to other group/cluster.

2. Cluster 2 (labelled as 1):
    a. Characterised as consumers with relatively higher bank balance and more credit spending.
    b. This group is also entitled to higher credit limit and also spend more on a single shopping.

Recommendation:

Cluster 2 consumers have displayed more credit-based spending, coupled with higher spending on an absolute basis.

Higher credit limit and lesser cash spending coupled with a higher likelihood of repayment makes the consumers of Cluster 2 a better target to be targeted for making more spends.

The bank should target the Cluster 2 consumers for any promotional schemes.

K-Means Clustering

The WSS plot divides the data in 3 distinct clusters. (as discussed in Q 1.4)

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Sil_width
KM Cluster								
0	18.5	16.2	0.88	6.18	3.7	3.63	6.04	0.47
1	11.86	13.25	0.85	5.23	2.85	4.74	5.1	0.4
2	14.44	14.34	0.88	5.51	3.26	2.71	5.12	0.34

1. Cluster 1 (labelled as 0):
    a. Highest spenders, high repayment rate, higher balance and highest credit limits with lowest cash spend (as a proportion to total spend).

2. Cluster 2 (labelled as 1):
    a. Conservative spenders, higher cash spends along with low current balance.
    b. This cluster has the lowest credit limit too.
3. Cluster 3 (labelled as 2):
    a. Moderate spenders, with better credit limit and balance as compared to Cluster 2.

Recommendations:

1. The Consumers of Cluster 1 are the lowest hanging fruits for the promotional schemes. They are thrifty spenders and have credibility too in favour of credit spends.
2. Cluster 3 consumers are the ones which could be converted to aggressive spenders through promotion and post sales service.

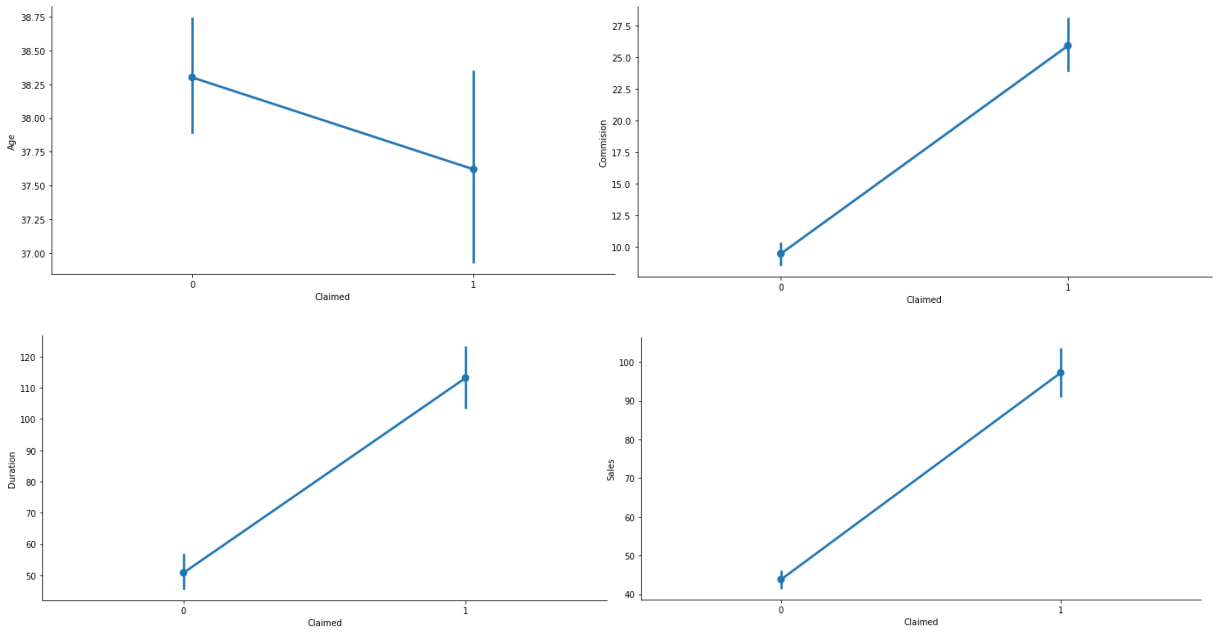
Problem 2: CART-RF-ANN

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.

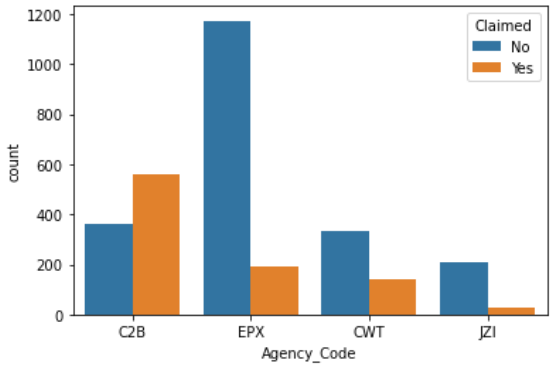
<u>For Continous variables</u>

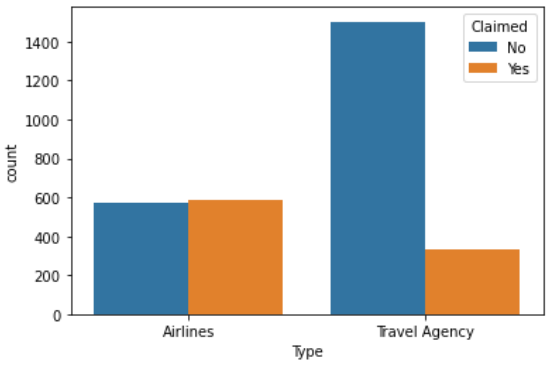
1. The Claim rate is inversely proportional to the Age.
2. Directly correlated to the Duration and Sales of the policy.
3. Surprisingly, there is also a positive correlation between Commissions and Claims. i.e. The more the policies sold through indirect channels, the higher the claims experienced.

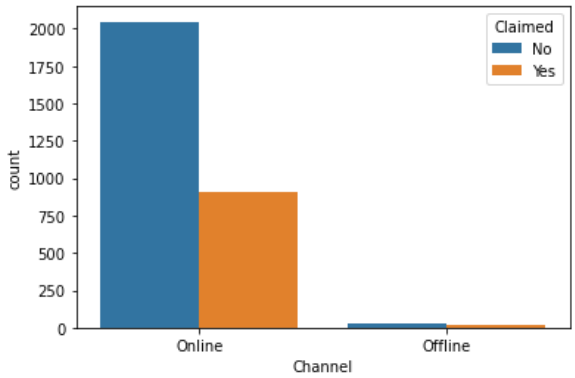


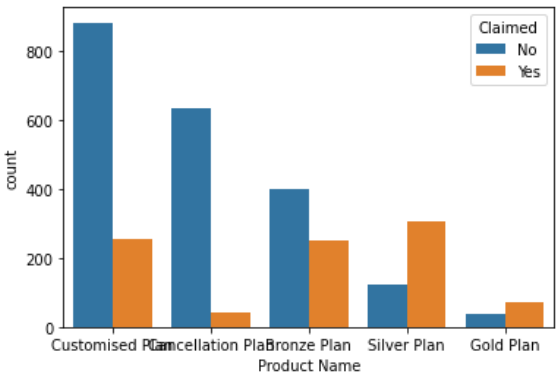
<u>For Categorical Variables:</u>

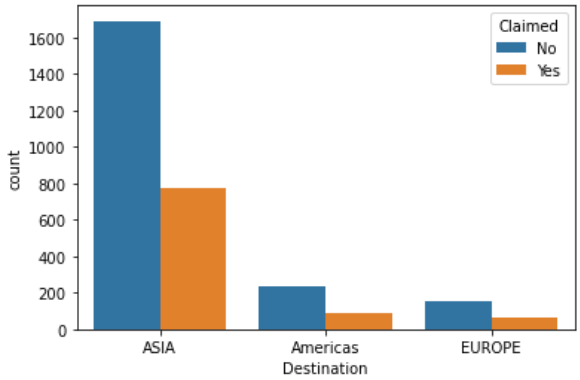
1. **Agency Code:** Tour company (C2B) claims most of the insurance claims.
2. **Airlines** Insurance have almost equal number of Claims as non claims.
3. Almost all the Insurances are processed over **online mode**.
4. **Gold Plan and Silver Plan** has the higher claims in the **Type** mix.
5. Travels insurance to **Asia** claims more than other destinations



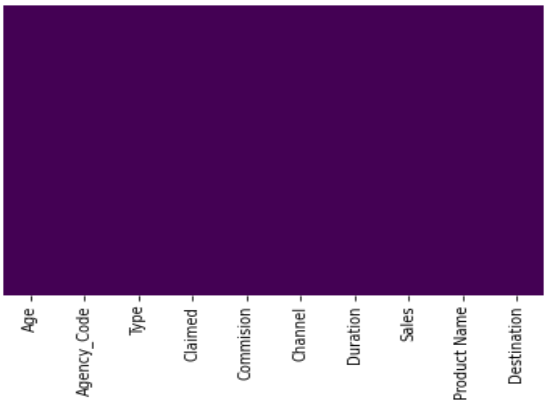








The data contains 0 (Zero) null values: As inferred by heatmap of null values.



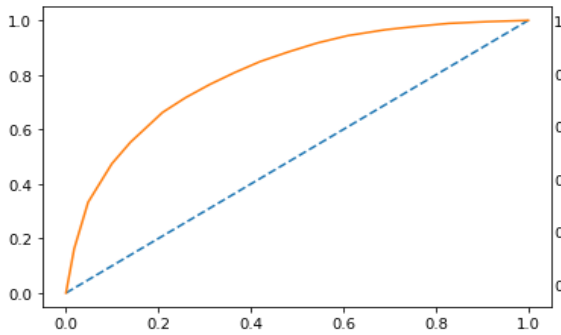
2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network

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.

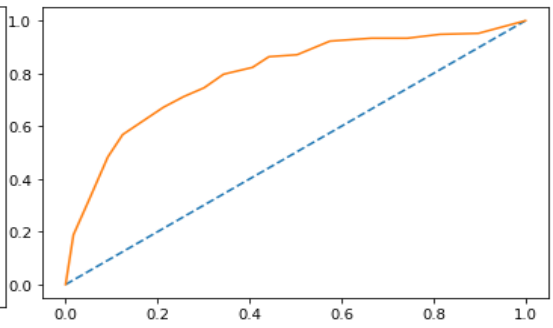
	CART TRAIN	CART TEST	RF_TRAIN	RF_TEST	NeuralNetwork_ TRAIN	NeuralNetwork _TEST
AUC	80.72%	79.50%	80.72%	70.29%	78.78%	80.58%
Recall	47.00%	48.00%	54.00%	51.00%	42.00%	41.00%
Accuracy	76.37%	77.42%	77.77%	77.30%	75.42%	76.14%
Precision	69.00%	71.00%	70.00%	69.00%	71.00%	71.00%
F1-Score	56.00%	57.00%	61.00%	59.00%	52.00%	52.00%

The ROC data with AUC value is as follows

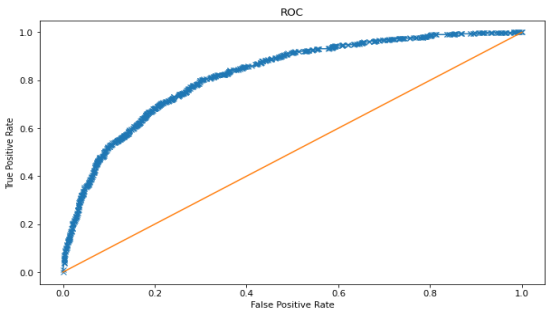
CART Train:AUC - 80.7%



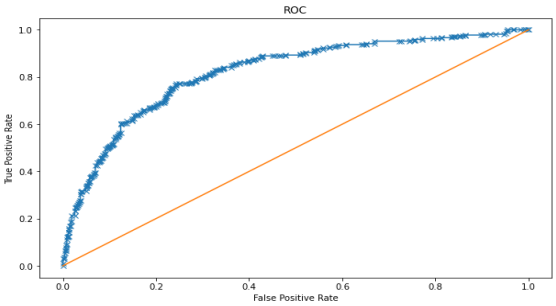
CART Test data: AUC= 79.5%



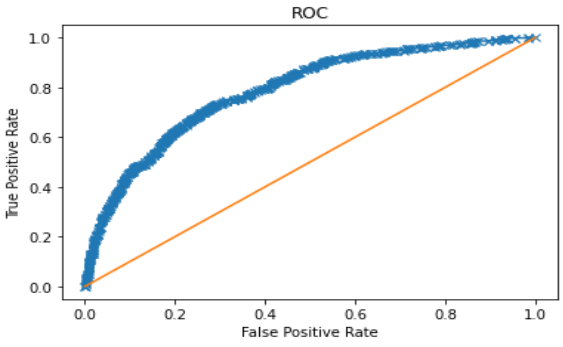
Rainforest: Train AUC = 71.41%



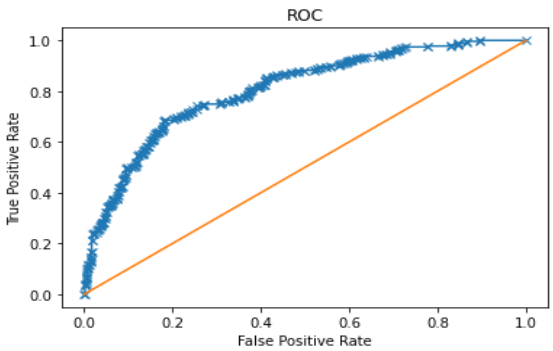
Rainforest: Test AUC = 70.28%



ANN Train data AUC = 78.78%



ANN Test Data AUC = 80.58%



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

Comparison of performance metrics from the 3 models

	CART TRAIN	CART TEST	RF_TRAIN	RF_TEST	NeuralNetwork_ TRAIN	NeuralNetwork _TEST
Recall	47.00%	48.00%	54.00%	51.00%	42.00%	41.00%
Precision	69.00%	71.00%	70.00%	69.00%	71.00%	71.00%
AUC	80.72%	79.50%	80.72%	70.29%	78.78%	80.58%
Accuracy	76.37%	77.42%	77.77%	77.30%	75.42%	76.14%
F1-Score	56.00%	57.00%	61.00%	59.00%	52.00%	52.00%

Best Optimised model: Random Forest Ensembler

Preferred model: RainForest Ensemble, because the sensitivity is the highest across the various models for the Insurance company. Rest other performance parameters are moderately above average.

As per the business case, the performance indicator we are keen to watch for are as listed in the rank of their preference:

1. Recall/ Sensitivity
2. Precision
3. AUC
4. Accuracy

Recall:. Based on the business case: where the Insurance business is experiencing higher frequency claims, it is important to use the model with higher Recall/Sensitivity ratio. It will be helpful to understand the ratio of Actual True predictions. Higher Sensitivity ratio means lower ‘False negatives’ and higher ‘True Positives’.

Higher the Recall value, better is the sensitivity of the model.

Precision: Given the problem of frequent claims, Ratio of Predicted Positives are also important for the prediction ability of the model. Lower the no of False predictions, lower the chances of fake claims in the business.

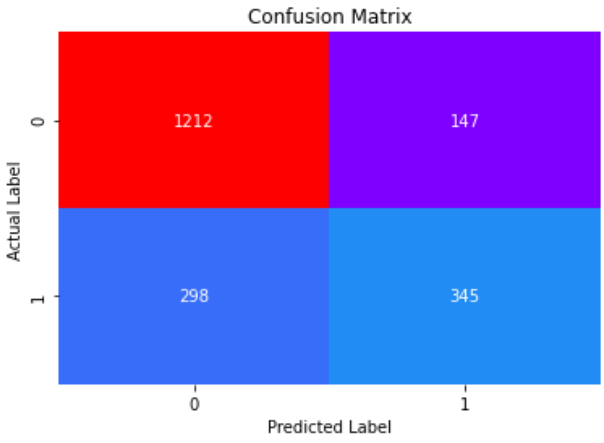
AUC: The area under the ROC curve is simple explanator of how True positive rates move against False positive rates.

Accuracy: The accuracy score of the model believes ‘lesser the false, better the predictions’ and hence, another important variable to look at.

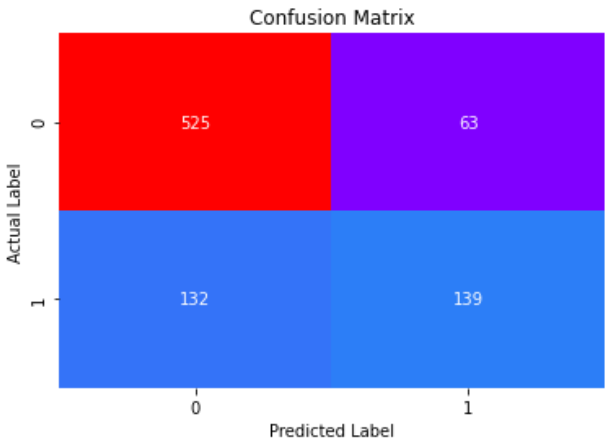
2.5 Inference: Basis on these predictions, what are the business insights and recommendations.

Preferred model: RainForest Ensemble, because the sensitivity is the highest across the various models, which is the most critical performance parameter for the Insurance company.

TRAIN Data (Rainforest Model)



Test data (Rainforest model)



Insights:

1. Agency codes are the biggest contributors to the Claimes filed.
2. The Claims are higher at a relatively younger age. Elderly customers, surprisingly do not record more claims.
3. The Claims filed through Agency channel C2B are maximum.
4. The Claims via Airlines travel insurance and travelling to Asia are biggest contributors to Claims filed.
5. Customers of Gold and Silver plans are the largest Claim filers across the type of Insurance plans.

Recommendations:

1. The company must re-verify the claims filed by customers:
    a. travelling to Asia,
    b. Insured online
    c. Via Commission Agent
    d. younger in age
    e. And travelling via Air

    To check for (False Positive) quandrant as they are likely to be fake claims as suggested by Precision measure. Failing to do so might impact the profitability of the business in the short term.

2. The company must assert and check on the claims predicted under False Negative (suggested by Recall measure) before rejecting, as they are likely to be actual claims which the model might have predicted wrongly. They could be:

a. Older/ Senior citizen customers.
b. Customers Insured via travel agency
c. Customers travelling to America and Europe

Failing to do so, might impact the reputation of the business badly and even might end up with customers filing for litigations in the long run.

The overall model performance is above moderate to start predicting if the customer will apply for insurance claim or not.