9/26/2021

Data Mining Report PGP -DSBA

Saloni Juwatkar
PGP – DATA SCIENCE AND BUSINESS ANALYTICS

Table of Contents

List of Figures

List of Tables

1 Problem Statement: 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.

Data Description

1. spending: Amount spent by the customer per month (in 1000s)
2. advance_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability_of_full_payment: Probability of payment done in full by the customer to the bank
4. current_balance: Balance amount left in the account to make purchases (in 1000s)
5. credit_limit: Limit of the amount in credit card (10000s)
6. min_payment_amt: minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s)

Sample Dataset

	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

Figure 1: PS1 - Sample Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

Figure 2: PS 1 - Data Info

The dataset is for credit card users and has a total of 210 records representing 7 variables. All variables are continuous in nature.

```
spending                        0
advance_payments                0
probability_of_full_payment     0
current_balance                 0
credit_limit                    0
min_payment_amt                 0
max_spent_in_single_shopping    0
dtype: int64
```

Figure 3: PS 1 : Null Data

There are no missing values in the dataset.

```
Number of duplicate rows = 0
```

Figure 4: PS 1: Duplicate Data

There is no Duplicate data in the dataset.

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

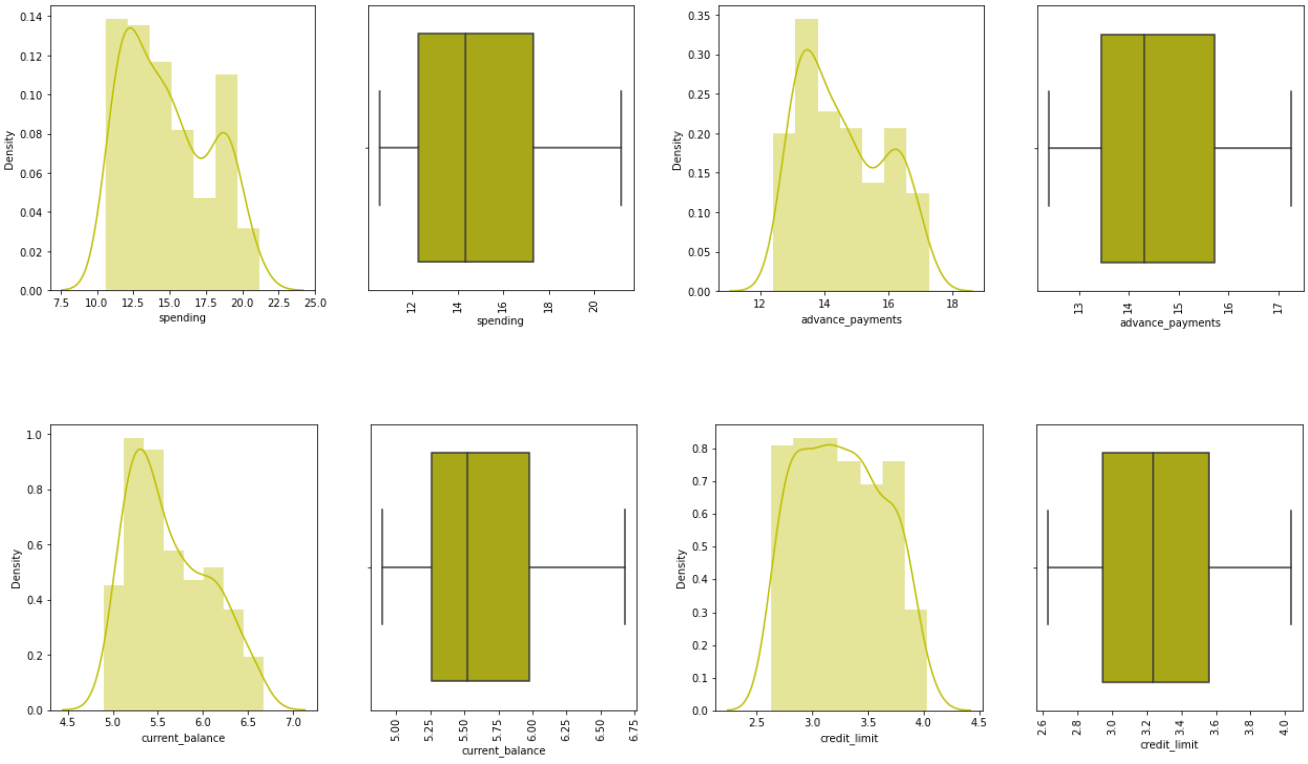
Univariate Analysis:

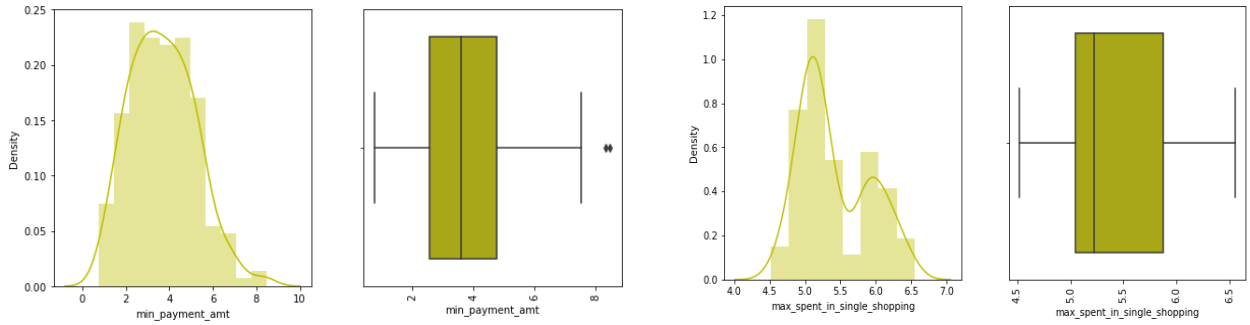
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

Figure 5: PS1 - Data Description

The number of rows of the data frame is 210.
The number of columns of the data frame is 7.

Distribution:





In the Dataset, only minimum payable amount has some outliers and we can ignore these.
The Data is a little right skewed.

MultiVariate Analysis:

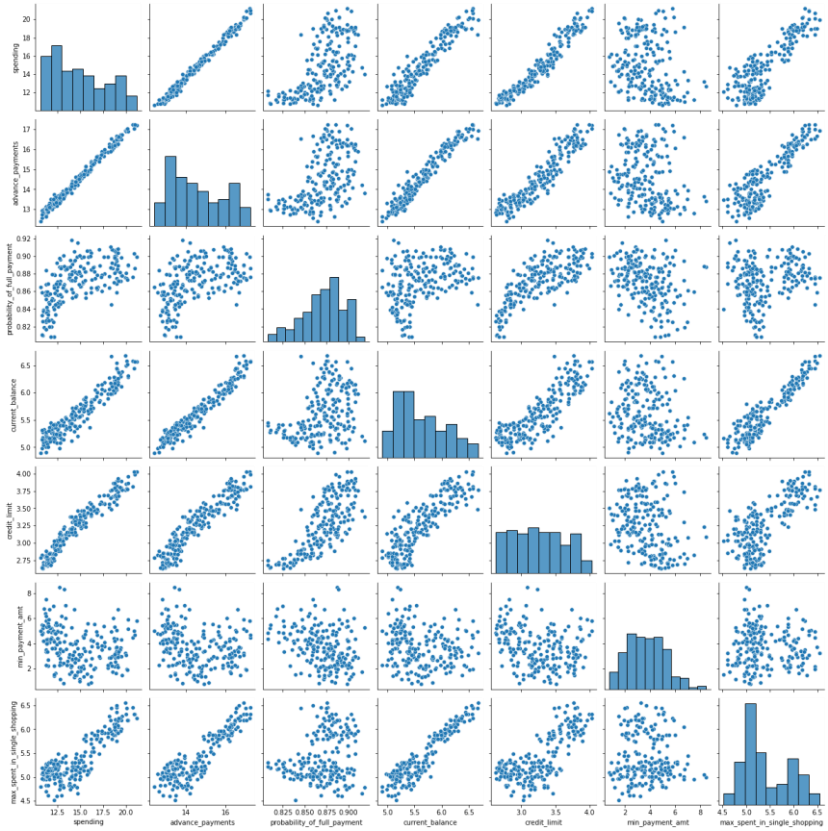


Figure 6: PS1 – Pair plot

Pair plot for the given data shows medium to strong correlation between the variables.

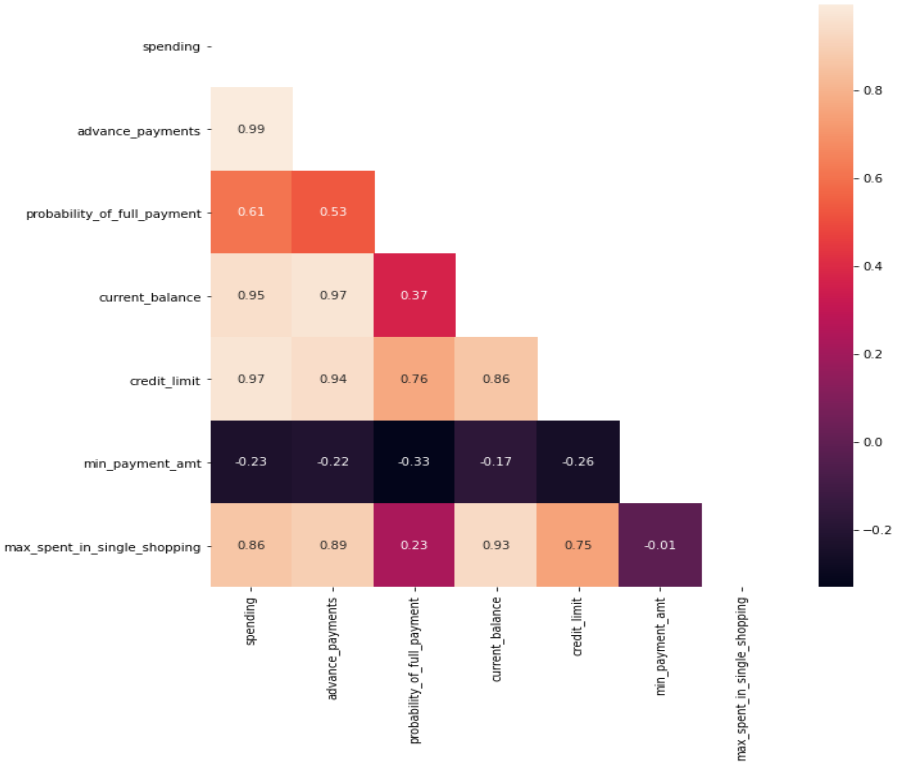


Figure 7: Heatmap

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

Yes, Scaling is necessary for clustering in this case as the variables:" spending" "advance_payments" are in the range of 10 – 20 while "probability_of_full_payment "is a probability number which will always be less than 1. This is just one example of how the data variables ranges are different. Due to this, variables with higher range may gain more weightage. if variables in the dataset have large difference in the variances it will disproportionately influence more on the construction of clusters. Hence, in order to bring the entire data to a common base line, Scaling is absolutely required.

Data Description before scaling:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

Figure 8: Data before scaling

Data Description after scaling:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02
mean	9.148766e-16	1.097006e-16	1.243978e-15	-1.089076e-16	-2.994298e-16	5.302637e-16	-1.935489e-15
std	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00
min	-1.466714e+00	-1.649686e+00	-2.668236e+00	-1.650501e+00	-1.668209e+00	-1.956769e+00	-1.813288e+00
25%	-8.879552e-01	-8.514330e-01	-5.980791e-01	-8.286816e-01	-8.349072e-01	-7.591477e-01	-7.404953e-01
50%	-1.696741e-01	-1.836639e-01	1.039927e-01	-2.376280e-01	-5.733534e-02	-6.746852e-02	-3.774588e-01
75%	8.465989e-01	8.870693e-01	7.116771e-01	7.945947e-01	8.044956e-01	7.123789e-01	9.563941e-01
max	2.181534e+00	2.065260e+00	2.006586e+00	2.367533e+00	2.055112e+00	3.170590e+00	2.328998e+00

Figure 9: PS1-Data after scaling.

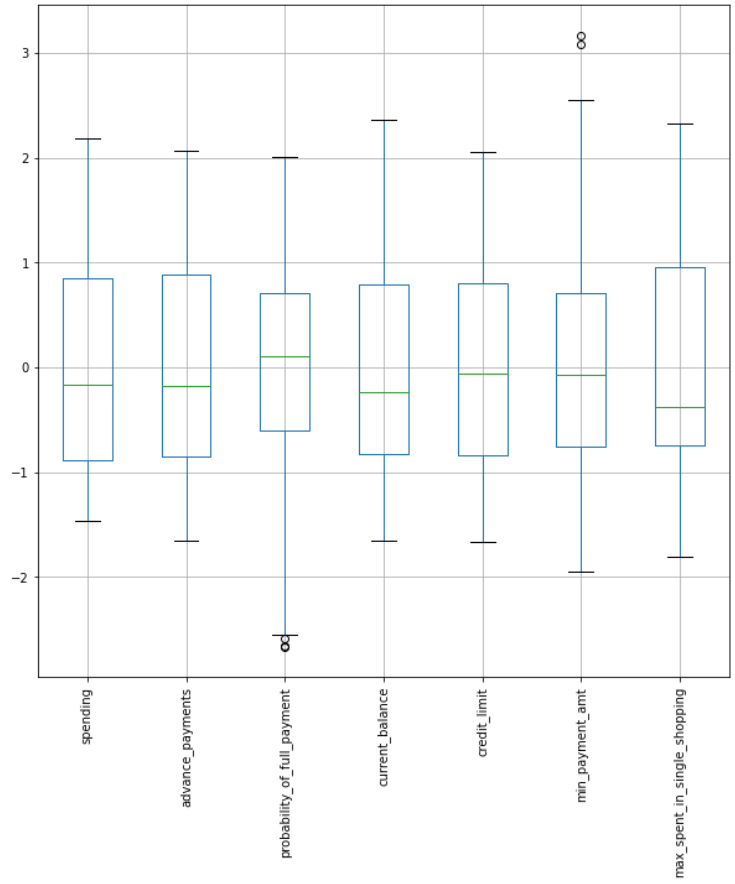


Figure 10:Scaled Data

From the above graphs, it is observed that the data is now standardised.

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

The Dendrogram obtained by applying hierarchical clustering to scaled data is as below:

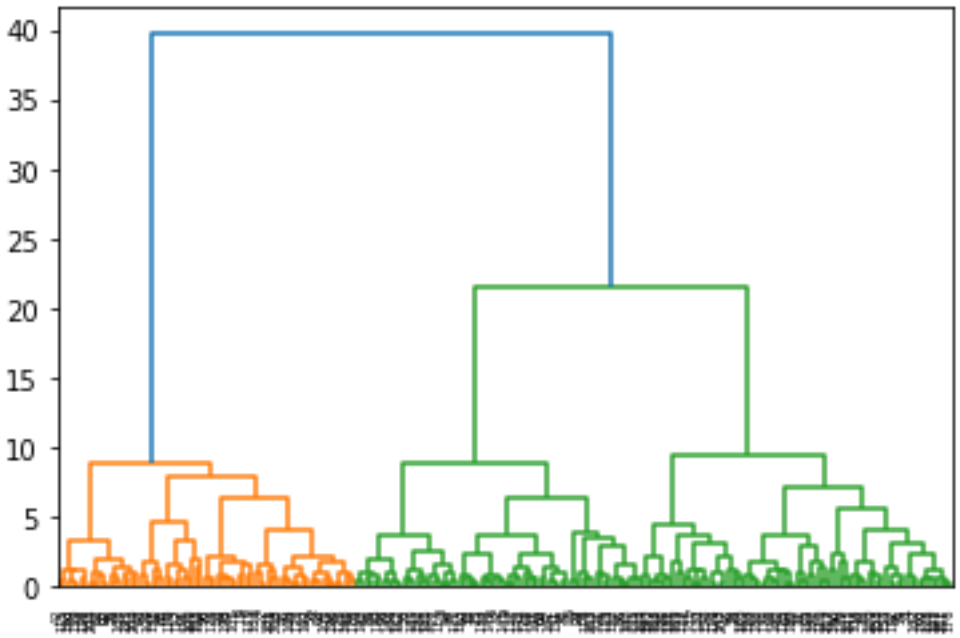


Figure 11: Hierarchical Clustering

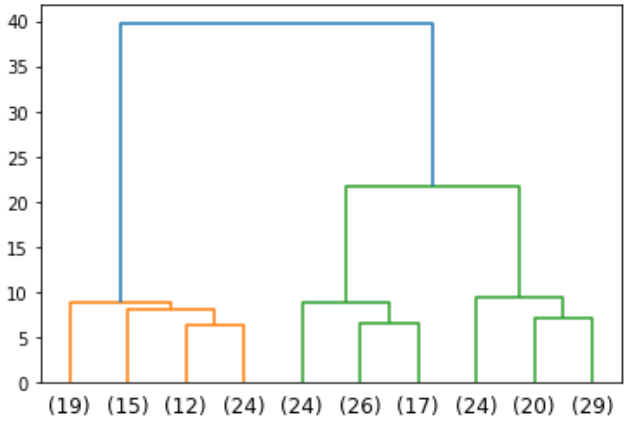


Figure 12: Truncated Dendrogram

From the dendrogram, we can derive the data into 3 clusters.

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.

From the Figure 11: Hierarchical Clustering, we can see that 3 clusters would be optimum but to support this analysis elbow curve of K-means inertia and silhouette score is used.

WSS:

```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.301721277542,
 326.36254154106956,
 289.22019649887096,
 262.44580353616084,
 243.82103621456784,
 222.1860907842431,
 206.51575924541532]
```

Clearly the optimum number of clusters as statistically derived is 3 as there as the values change is very less after 3rd cluster. However, if we look at the WSS Elbow curve plot, we find that there is a significant elbow forming for K=3.

Once we take a look at the data, we can see that instead of just High Spending and Low spending, it is more convenient to segment the customers into High spending, Medium Spending and Low spending customers with the limited amount of Data available.

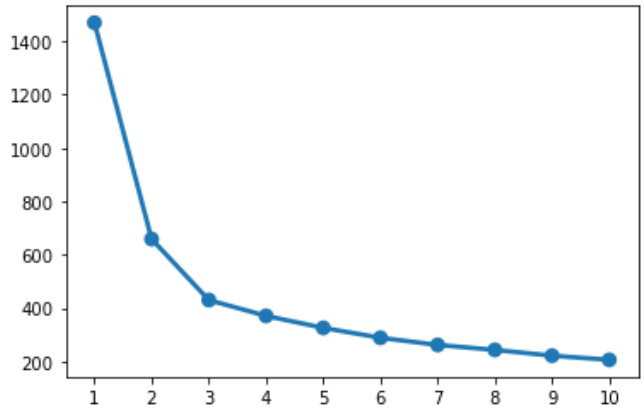


Figure 13: K-Means WSS score Elbow Graph

The minimum silhouette score is 0.00271 for 3 clusters which suggest that that the decided no of clusters are optimum.

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

The Three Segments into which we have divided the customer profiles are:

1. Low Spending

2. Medium Spending

3. High Spending

Reason: Apart from the spending amount, there is no data point which could be neatly segregated into different groups.

Table 1: Cluster Profile

	spending	advance_ payments	probability_ of_full _payment	current_ balance	credit_ limit	min_ payment_ amt	max_spent _in_single _shopping	clusters	sil_ width	Frequency
Cluster_ Means										
0	14.438	14.338	0.8816	5.5146	3.2592	2.7073	5.1208	2.8732	0.3398	71
1	11.857	13.248	0.8483	5.2318	2.8495	4.7424	5.1017	2.0833	0.3975	72
2	18.495	16.203	0.8842	6.1757	3.6975	3.6324	6.0417	1.0299	0.4688	67

Cluster 0:

Medium spending group, however the minimum payment is less compared to less spending group. We should give this group some promotional offers. This group also has a probability of 88.16% for full payments.

Cluster 1:

Least spending group. Minimum amount spent is more than others. Average of most parameters is more or less same. We can increase the credit limit of this cluster customers. This cluster to be encouraged to spend more through promotional offers.

Cluster 2:

Group spending more money. All parameters high compared to other clusters. Minimum amount spent is less than least spending group. Bank can increase minimum amount of this group. This cluster can be targeted for credit dependent promotional offers.

2 Problem Statement: 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 provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

Data Description

The dataset includes the following variables:

1. Target: Claim Status (Claimed)
2. Code of tour firm (Agency_Code)
3. Type of tour insurance firms (Type)
4. Distribution channel of tour insurance agencies (Channel)
5. Name of the tour insurance products (Product)
6. Duration of the tour (Duration)
7. Destination of the tour (Destination)
8. Amount of sales of tour insurance policies (Sales)
9. The commission received for tour insurance firm (Commission)
10. Age of insured (Age)

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

	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

Figure 14: P2 - Data Description

Out of the 10 columns there are 6 object type variables, 2 float types and 2 integer types 'Claimed' is the Target variable.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Age           3000 non-null   int64  
 1   Agency_Code   3000 non-null   object 
 2   Type          3000 non-null   object 
 3   Claimed       3000 non-null   object 
 4   Commision     3000 non-null   float64
 5   Channel       3000 non-null   object 
 6   Duration      3000 non-null   int64  
 7   Sales         3000 non-null   float64
 8   Product Name  3000 non-null   object 
 9   Destination   3000 non-null   object 
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

Figure 15: P2 - Data Info

There are no missing values in the dataset:

```
Age             0
Agency_Code     0
Type            0
Claimed         0
Commision       0
Channel         0
Duration        0
Sales           0
Product Name    0
Destination     0
dtype: int64
```

Figure 16: P2 - Null Data

	Age	Commision	Duration	Sales
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	38.091000	14.529203	70.001333	60.249913
std	10.463518	25.481455	134.053313	70.733954
min	8.000000	0.000000	-1.000000	0.000000
25%	32.000000	0.000000	11.000000	20.000000
50%	36.000000	4.630000	26.500000	33.000000
75%	42.000000	17.235000	63.000000	69.000000
max	84.000000	210.210000	4580.000000	539.000000

Figure 17: P2 - Data Description

There are 139 duplicate records in the dataset, which need to be removed. After removal of these records, the final shape of the dataset is 2861 records.

```
Number of duplicate rows = 139
```

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
63	30	C2B	Airlines	Yes	15.0	Online	27	60.0	Bronze Plan	ASIA
329	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
407	36	EPX	Travel Agency	No	0.0	Online	11	19.0	Cancellation Plan	ASIA
411	35	EPX	Travel Agency	No	0.0	Online	2	20.0	Customised Plan	ASIA
422	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
...	...	...	...	...	...	...	...	...	...	...
2940	36	EPX	Travel Agency	No	0.0	Online	8	10.0	Cancellation Plan	ASIA
2947	36	EPX	Travel Agency	No	0.0	Online	10	28.0	Customised Plan	ASIA
2952	36	EPX	Travel Agency	No	0.0	Online	2	10.0	Cancellation Plan	ASIA
2962	36	EPX	Travel Agency	No	0.0	Online	4	20.0	Customised Plan	ASIA
2984	36	EPX	Travel Agency	No	0.0	Online	1	20.0	Customised Plan	ASIA

139 rows × 10 columns

Figure 18: Duplicate Data – Before

Univariate Analysis:

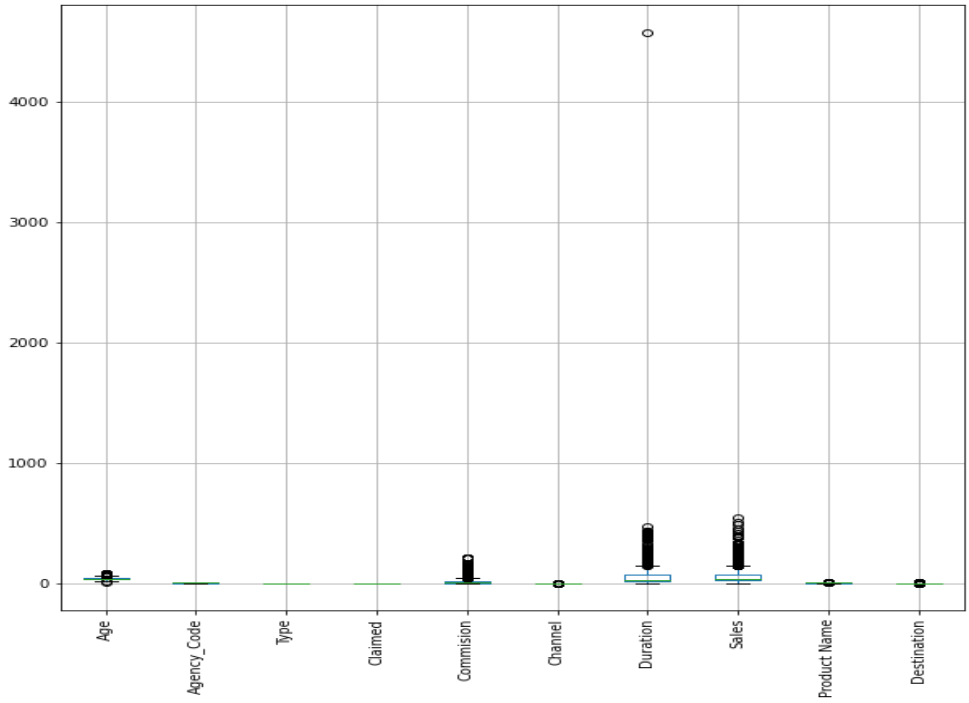


Figure 19: P2 - Boxplot

Multivariate Analysis:

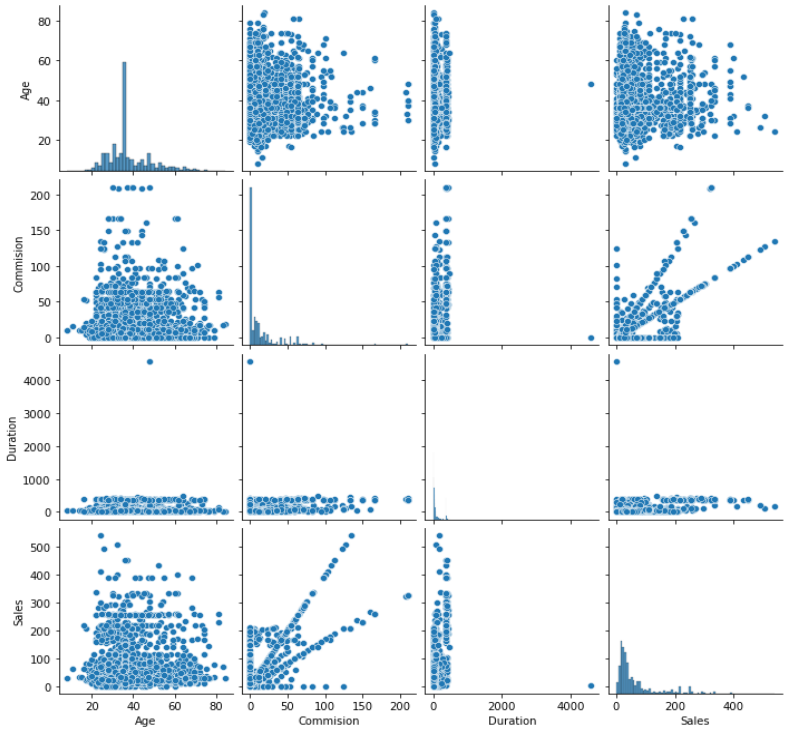


Figure 20: P2 - Multivariate Analysis (Pair Plot)

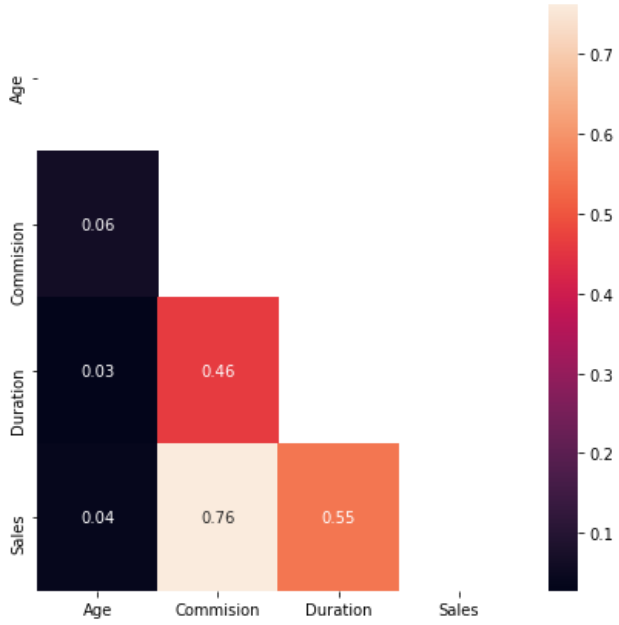


Figure 21: P2 - Heatmap

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

The criteria chosen to split the data into test and train sets is Claimed as to decide which all parameters/ variables affect the claim status.

The test data comprises 30% of the overall records in the dataset.

The Test and train sets to be used in the classification model are as follows:

```
X_train (2002, 9)
X_test (859, 9)
train_labels (2002,)
test_labels (859,)
```

Figure 22: Train and Test Set Size

X_train:

Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
69	0	0	6.00	1	7	15.0	0	0
36	2	1	0.00	1	29	35.0	2	0
60	1	1	41.58	1	8	69.3	2	1
36	0	0	9.75	1	70	39.0	4	0
36	2	1	0.00	1	39	51.0	1	2

Figure 23: Train Set

X_test

Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
31	1	1	0.00	0	402	97.0	2	0
68	2	1	0.00	1	60	29.0	1	0
42	0	0	21.00	1	11	84.0	4	0
44	1	1	23.76	1	51	39.6	2	0
50	1	1	35.64	1	111	59.4	2	0

Figure 24: Test Set

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.

A. CART:

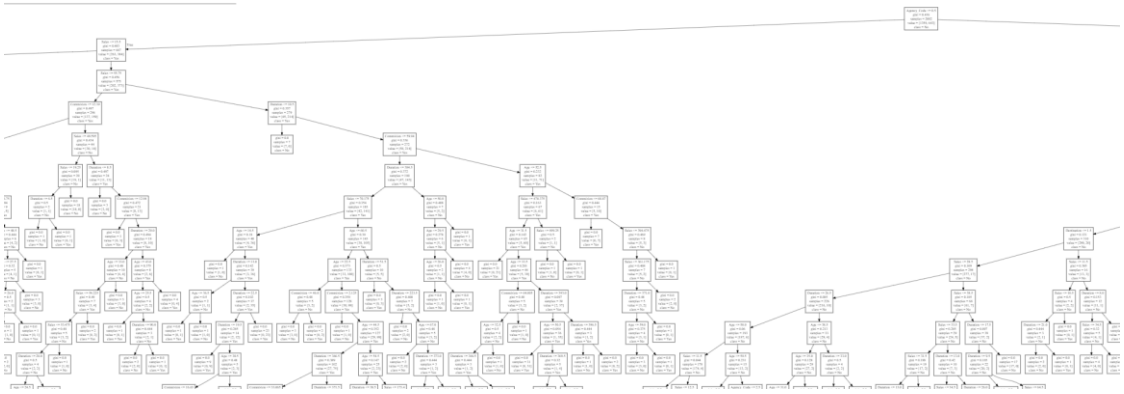


Figure 25: Decision Tree Before Pruning

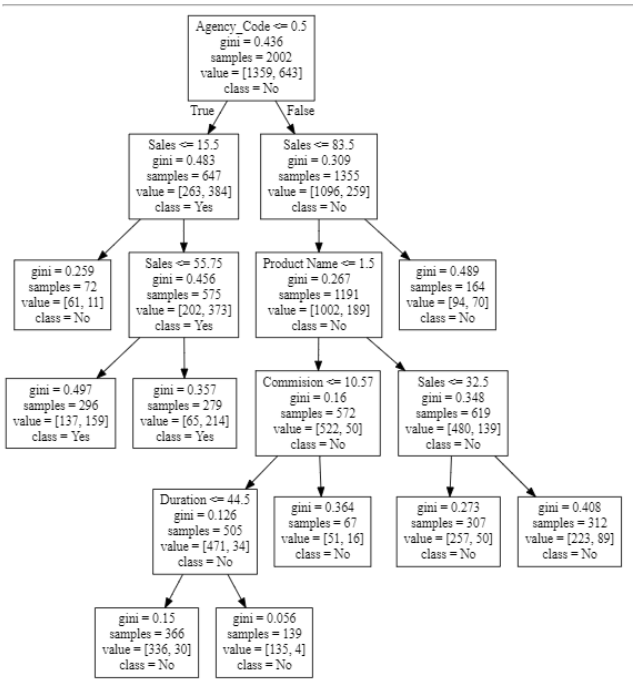


Figure 26: Decision Tree - CART Pruned

Optimal parameters for building Decision Tree Classifiers were:

{'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 50, 'min_samples_split': 450}

Agency_Code happens to be the most important influencing variable on predicting the target variable behaviour.

	Imp
Age	0.000000
Agency_Code	0.619073
Type	0.000000
Commision	0.015189
Channel	0.000000
Duration	0.002489
Sales	0.314424
Product Name	0.048825
Destination	0.000000

1. Confusion Matrix

1. Test

```
array([[510,  78],
       [109, 162]], dtype=int64)
```

2. Train

```
array([[1157,  202],
       [ 270,  373]], dtype=int64)
```

2. ROC Curve with AUC score:

1. Test

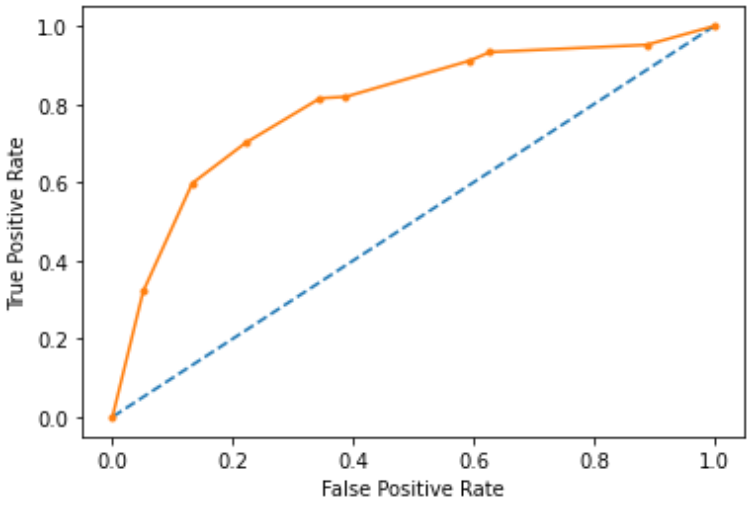


Figure 27 : CART - ROC(Test)

AUC: 0.796

2. Train

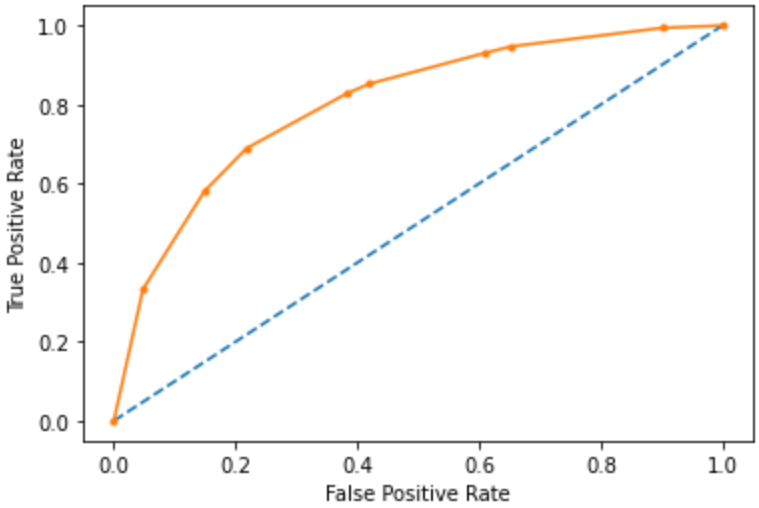


Figure 28: CART - ROC(Train)

AUC: 0.803

3. Classification Report

1. Test

	precision	recall	f1-score	support
0	0.82	0.87	0.85	588
1	0.68	0.60	0.63	271
accuracy			0.78	859
macro avg	0.75	0.73	0.74	859
weighted avg	0.78	0.78	0.78	859

Figure 29: CART - Test - Classification Report.

2. Train

	precision	recall	f1-score	support
0	0.81	0.85	0.83	1359
1	0.65	0.58	0.61	643
accuracy			0.76	2002
macro avg	0.73	0.72	0.72	2002
weighted avg	0.76	0.76	0.76	2002

Figure 30: CART - Train - Classification Report.

Conclusion of CART:

Training and Test sets are almost similar, but from the classification report Precision, recall and f1-score have improved for the Test data slightly and thus the model is a good model.

Agency_Code is the most important for predicting Claim status.

B. Random Forest:

Optimal parameters for building Decision Tree Classifiers were:

RandomForestClassifier(max_depth=10, max_features=4, min_samples_leaf=10, min_samples_split=30, n_estimators=150, random_state=1)

Agency_Code happens to be the most important influencing variable on predicting the target variable behaviour.

	Imp
Agency_Code	0.252259
Sales	0.207481
Product Name	0.179632
Commision	0.126912
Duration	0.112591
Age	0.076571
Type	0.027728
Destination	0.015568
Channel	0.001257

1. Confusion Matrix

a. Test

```
array([[525,  63],
       [119, 152]], dtype=int64)
```

b. Train

```
array([[1229,  130],
       [ 258,  385]], dtype=int64)
```

2. ROC Curve with AUC score:

a. Test

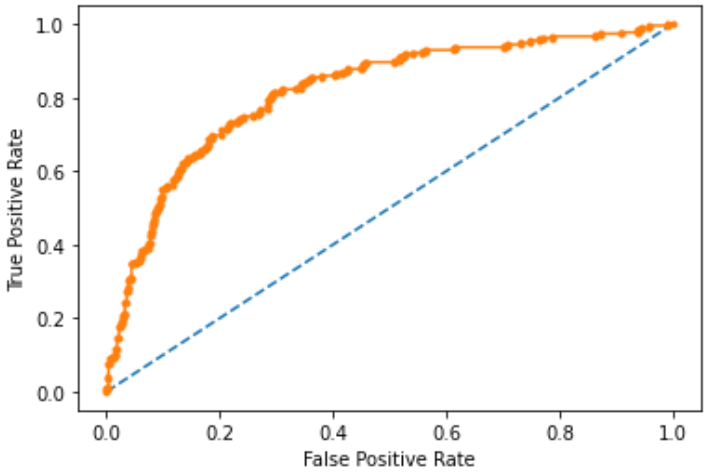


Figure 31: Random Forest - ROC(Test)

AUC : 0.818

b. Train

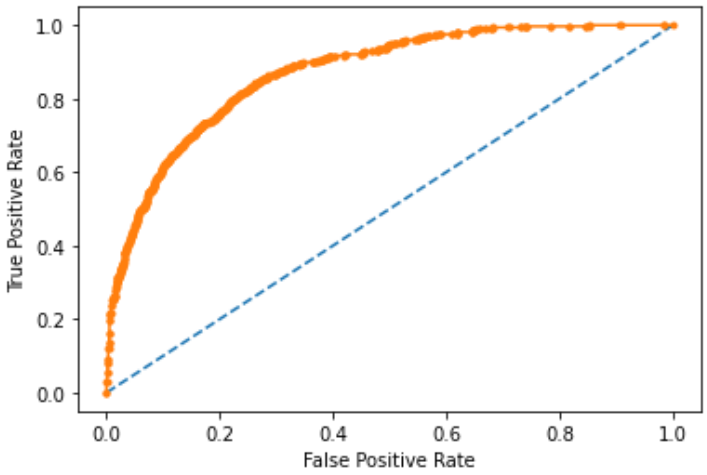


Figure 32: Random Forest - ROC(Train)

AUC : 0.871

3. Classification Report

a. Test

	precision	recall	f1-score	support
0	0.82	0.89	0.85	588
1	0.71	0.56	0.63	271
accuracy			0.79	859
macro avg	0.76	0.73	0.74	859
weighted avg	0.78	0.79	0.78	859

Figure 33: Random Forest - Classification Report (Test)

b. Train

	precision	recall	f1-score	support
0	0.83	0.90	0.86	1359
1	0.75	0.60	0.66	643
accuracy			0.81	2002
macro avg	0.79	0.75	0.76	2002
weighted avg	0.80	0.81	0.80	2002

Figure 34: Random Forest - Classification Report (Train)

Conclusion for Random Forest :

Training and Test sets are almost similar, but from the classification report Precision, recall and f1-score have degraded for the Test data slightly.

Agency_Code is the most important for predicting Claim status.

C. ANN:

1. Confusion Matrix

a. Test

```
array([[553,  35],
       [181,  90]], dtype=int64)
```

b. Train

```
array([[1279,   80],
       [ 445,  198]], dtype=int64)
```

2. ROC Curve with AUC score:

a. Test

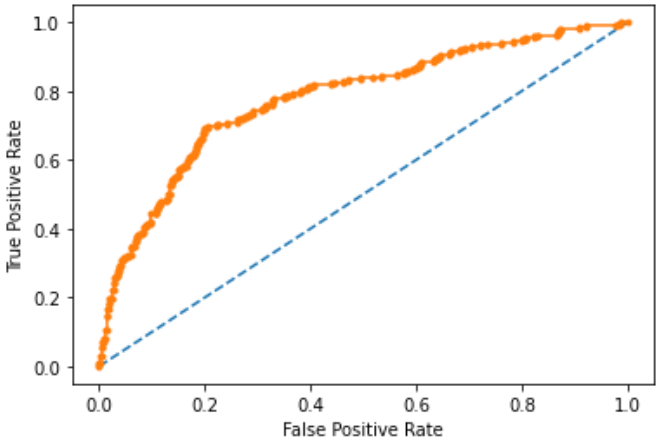


Figure 35: ANN - ROC(Test)

AUC: 0.78

b. Train

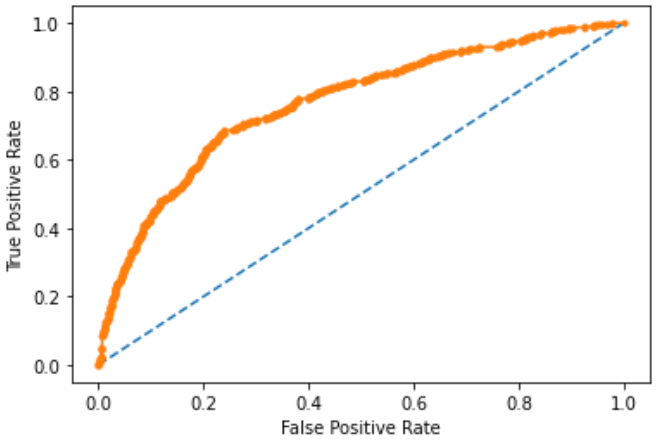


Figure 36:ANN - ROC(Train)

AUC: 0.768

3. Classification Report

a. Test

	precision	recall	f1-score	support
0	0.75	0.94	0.84	588
1	0.72	0.33	0.45	271
accuracy			0.75	859
macro avg	0.74	0.64	0.65	859
weighted avg	0.74	0.75	0.72	859

Figure 37: ANN - Classification Report (Test)

b. Train

	precision	recall	f1-score	support
0	0.74	0.94	0.83	1359
1	0.71	0.31	0.43	643
accuracy			0.74	2002
macro avg	0.73	0.62	0.63	2002
weighted avg	0.73	0.74	0.70	2002

Figure 38: ANN - Classification Report (Train)

Conclusion for ANN:

Training and Test set results are almost similar, but the sensitivity score is low for both the training and testing data.

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

Comparison of all models on basis on 2.3 is as follows

	CART		Random Forest		ANN	
	Test	**Train**	**Test**	**Train**	**Test**	**Train**
Accuracy	78	76	79	81	75	74
AUC	79.6	80.3	81.8	87.1	78	76.8
Recall	60	58	56	60	33	31
Precision	68	65	71	75	72	71
F1 Score	63	61	63	66	45	43

Table 2: Comparison of CART, Random Forest and ANN models

It can be concluded that Random Forest Model is better than CART or ANN model for the following reasons:

1. Accuracy of Random Forest is better than other models.
2. Precision, Recall and F1-score are also slightly better than the CART or ANN models.
3. AUC is better than all other models.

Therefore, to predict the claim status, Random Forest model shall provide a better understanding of the predictions. Next best model is CART and ANN is least preferable model.

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

On Basis of Random Forest model, the insurance company can predict the higher probability of the claim being filed depending on various attributes. The Agency code is on the most important factor on which prediction is done, thus emphasis must be given to the agency to train it to profile the customers in order.

The End!