Project

Karthiheswar

Table of Contents

1 Project Objective

The objective of the report is to explore all the projects data set in Python and generate insights about the data set. This exploration report will consist of the following:

- Importing the dataset in Python
- Understanding the structure of dataset
- Graphical exploration
- Applying different clustering techniques
- Clustering profiles
- Checking null values and performing descriptive statistics
- Scaling the variables
- Splitting the data into train and test and building models
- Insights from the dataset

2 Clustering the bank and marketing dataset

2.1 Reading the data and exploratory data analysis

Reading the dataset (head)

	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

Exploratory data analysis

Describing the data

	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

S.no	Description	IQR values for all attributes	Difference between highest and lowest values for all attributes
1	Spending	5.035000	10.5900
2	Advance_payments	2.265000	4.8400
3	Probability_of_full_payment	0.030875	0.1102
4	Current_balance	0.717500	1.7760
5	Credit_limit	0.617750	1.4030
6	Min_payment_amt	2.207250	7.6909
7	Max_spent_in_single_shopping	0.832000	2.0310

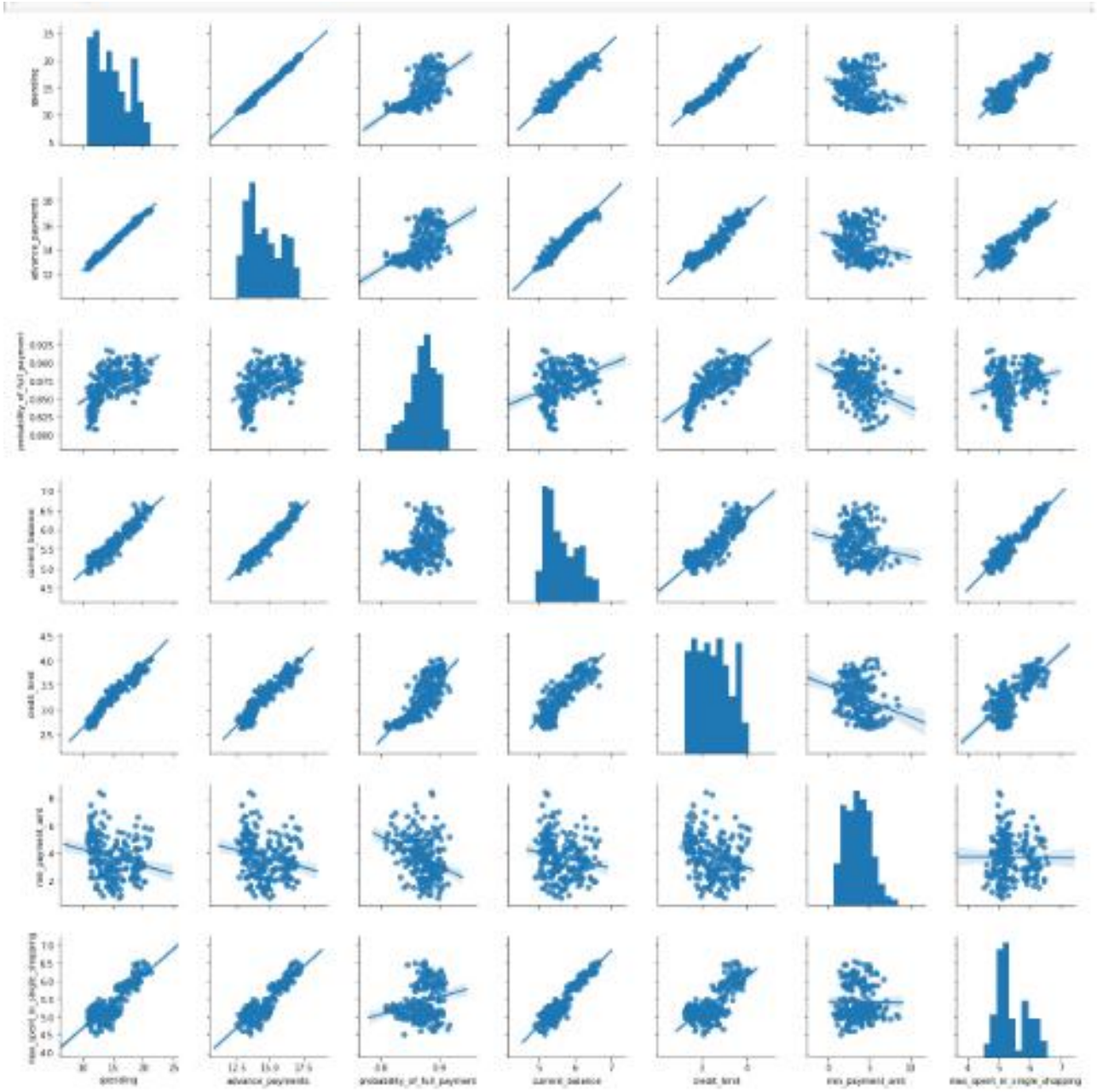
Covariance of each attribute against every other attribute

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopp
spending	8.466351	3.778443	0.041823	1.224704	1.066911	-1.004356	1.235
advance_payments	3.778443	1.705528	0.016332	0.562666	0.466065	-0.426766	0.571
probability_of_full_payment	0.041823	0.016332	0.000558	0.003852	0.006798	-0.011777	0.002
current_balance	1.224704	0.562666	0.003852	0.196305	0.143992	-0.114290	0.203
credit_limit	1.066911	0.466065	0.006798	0.143992	0.142668	-0.146543	0.139
min_payment_amt	-1.004356	-0.426766	-0.011777	-0.114290	-0.146543	2.260684	-0.008
_spent_in_single_shopping	1.235133	0.571753	0.002634	0.203125	0.139068	-0.008187	0.241

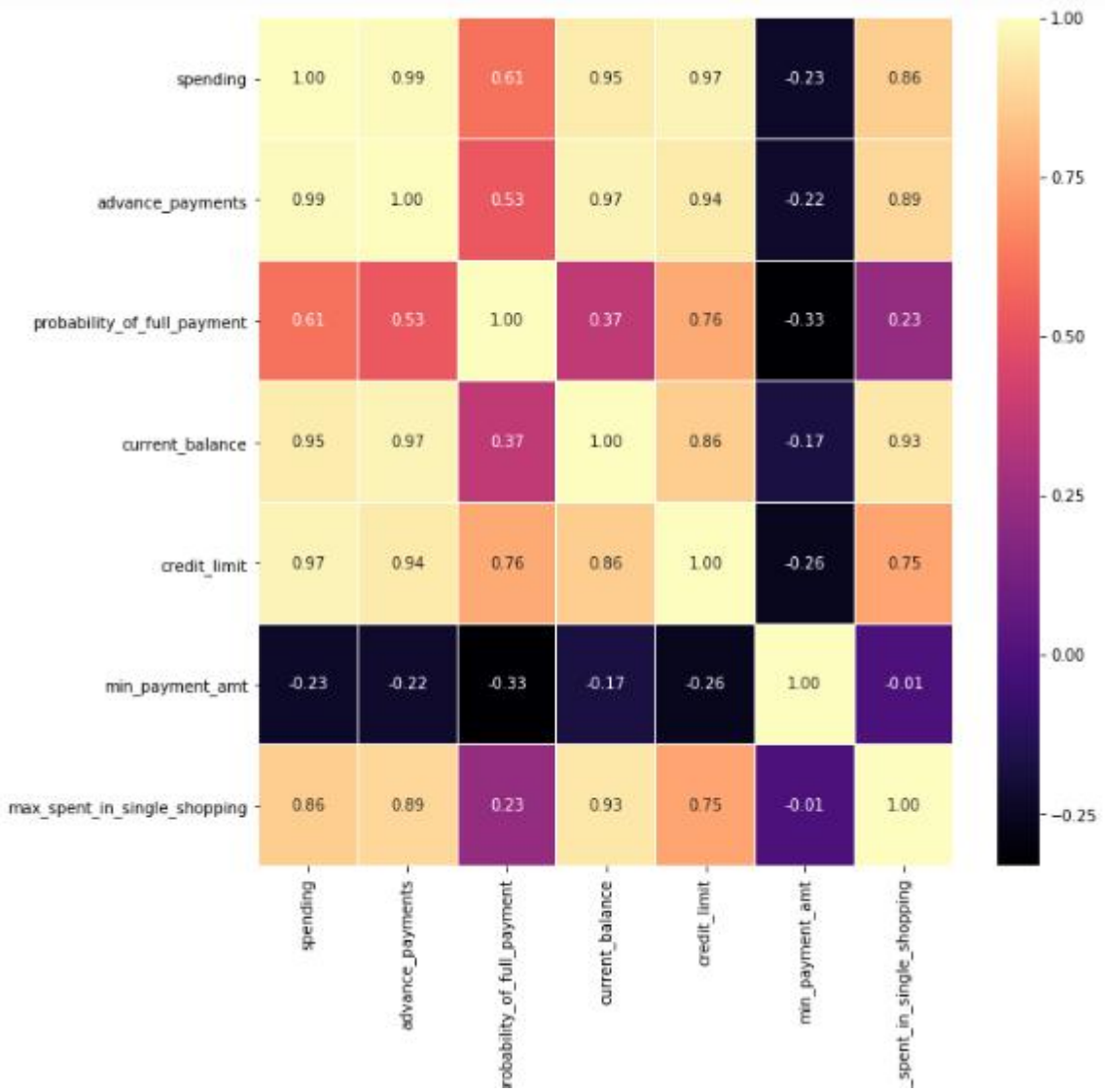
Correlation coefficient between every pair of attributes

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopp
spending	1.000000	0.994341	0.608288	0.949985	0.970771	-0.229572	0.863
advance_payments	0.994341	1.000000	0.529244	0.972422	0.944829	-0.217340	0.890
probability_of_full_payment	0.608288	0.529244	1.000000	0.367915	0.761635	-0.331471	0.226
current_balance	0.949985	0.972422	0.367915	1.000000	0.860415	-0.171562	0.932
credit_limit	0.970771	0.944829	0.761635	0.860415	1.000000	-0.258037	0.749
min_payment_amt	-0.229572	-0.217340	-0.331471	-0.171562	-0.258037	1.000000	-0.011
_spent_in_single_shopping	0.863693	0.890784	0.226825	0.932806	0.749131	-0.011079	1.000

Scatter plot



Heatmap



Skeweness

S.no	Description	Skeweness of every attribute
1	Spending	0.399889
2	Advance_payments	0.386573
3	Probability_of_full_payment	-0.537954
4	Current_balance	0.525482
5	Credit_limit	0.134378
6	Min_payment_amt	0.401667
7	Max_spent_in_single_shopping	0.561897

2.2 Scaling for clustering

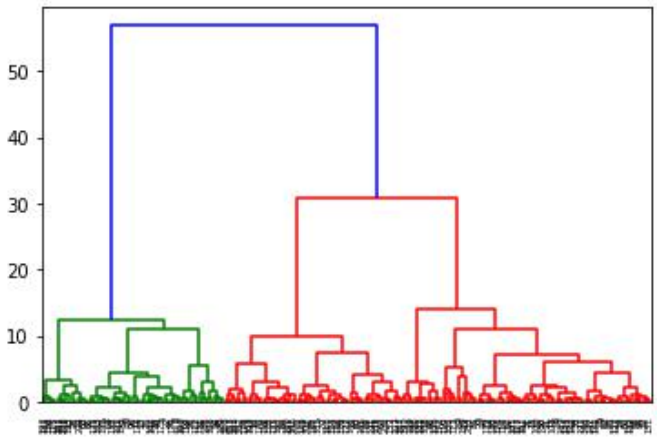
Scaled data

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.178230	2.367533	1.338579	-0.298806	2.328998
1	0.393582	0.253840	1.501773	-0.600744	0.858236	-0.242805	-0.538582
2	1.413300	1.428192	0.504874	1.401485	1.317348	-0.221471	1.509107
3	-1.384034	-1.227533	-2.591878	-0.793049	-1.639017	0.987884	-0.454961
4	1.082581	0.998364	1.196340	0.591544	1.155464	-1.088154	0.874813

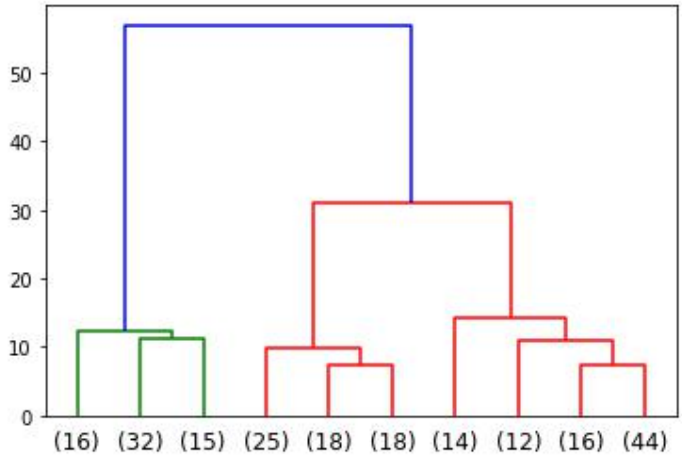
Scaling is definitely necessary as each columns contains different range of variables and different parameters, it should be scaled.

2.3 Hierarchical clustering to scaled data using Dendrogram

Hierarchical clustering



The optimum number of clusters are 10



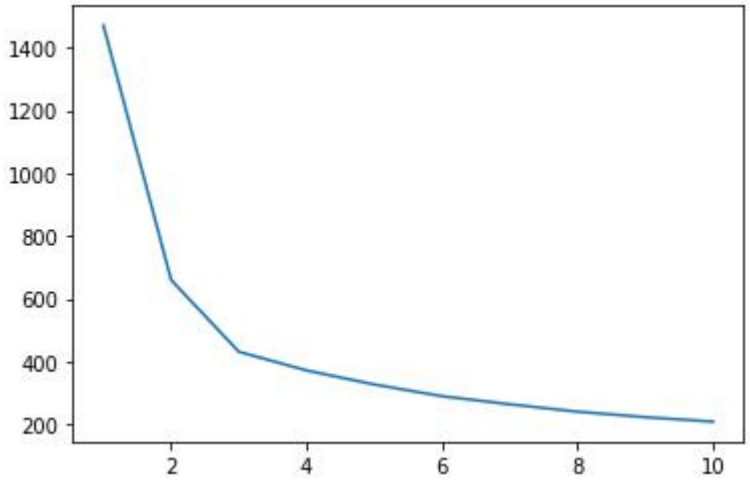
The optimum number of clusters are 10 as these 10 cluster shows the proper hierarchical relationship between objects.

2.4 K-Means clustering on scaled data

K-means clustering values from 1 to 10

[1470.0,
659.1717544870407,
430.65897315130053,
371.301721277542,
326.53057813155976,
289.2201964988712,
263.5084204019588,
239.91744118551287,
222.51271082015415,
208.10735185286126]

Elbow curve



Silhouette_score is 0.4007270552751299 and Min Silhouette_samples is 0.002713089347678533

2.5 Cluster profiles

Clus_kmeans	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	freq
0	1.256682	1.261966	0.560464	1.237883	1.164852	-0.045219	1.292308	67
1	-1.030253	-1.006649	-0.964905	-0.897685	-1.085583	0.694804	-0.624809	72
2	-0.141119	-0.170043	0.449606	-0.257814	0.001647	-0.661919	-0.585893	71

The cluster 0 has very less frequency and cluster 1 has highest variables. Where cluster 1 has very less spending and cluster 0 has very high spending. So in order to increase the spending, cluster 1 has to be concentrated more.

3 CART-RF-ANN on insurance dataset

3.1 Data Ingestion: Reading the dataset, performing descriptive statistics and null value condition check

Reading the dataset (head)

	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

Descriptive statistics

	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00

S.no	Description	Null value condition
1	Age	0
2	Agency_Code	0
3	Type	0
4	Claimed	0
5	Commision	0
6	Channel	0
7	Duration	0
8	Sales	0
9	Product_Name	0
10	Destination	0

Descriptive statistics after converting objects into integers

	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Agency_Code	3000.0	1.306333	0.994060	0.0	0.0	2.00	2.000	3.00
Type	3000.0	0.612333	0.487299	0.0	0.0	1.00	1.000	1.00
Claimed	3000.0	0.308000	0.461744	0.0	0.0	0.00	1.000	1.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Channel	3000.0	0.984667	0.122895	0.0	1.0	1.00	1.000	1.00
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00
Product_Name	3000.0	1.661667	1.258726	0.0	1.0	2.00	2.000	4.00
Destination	3000.0	0.250000	0.575277	0.0	0.0	0.00	0.000	2.00

The given dataset is imported and there are no null values present. Every columns containing discrete variables are converted into continuous variables for purpose of model building. Descriptive statistics are obtained for both original data and converted data.

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

RANDOM FOREST

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=7, max_features=3, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=50, min_samples_split=150,
                       min_weight_fraction_leaf=0.0, n_estimators=301,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Classification report of train data

	precision	recall	f1-score	support
0	0.81	0.91	0.86	1471
1	0.70	0.51	0.59	629
accuracy			0.79	2100
macro avg	0.76	0.71	0.72	2100
weighted avg	0.78	0.79	0.78	2100

Classification report of test data

	precision	recall	f1-score	support
0	0.76	0.92	0.83	605
1	0.72	0.40	0.52	295
accuracy			0.75	900
macro avg	0.74	0.66	0.68	900
weighted avg	0.75	0.75	0.73	900

CART DECISION TREE

DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=13,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=100, min_samples_split=200,
min_weight_fraction_leaf=0.0, presort=False,
random_state=None, splitter='best')

	Imp
Duration	0.263249
Sales	0.199095
Agency_Code	0.194797
Age	0.176348
Commision	0.093175
Product_Name	0.041322
Destination	0.022423
Channel	0.007262
Type	0.002329

Classification report of train data

	precision	recall	f1-score	support
0	0.81	0.90	0.85	1471
1	0.69	0.50	0.58	629
accuracy			0.78	2100
macro avg	0.75	0.70	0.72	2100
weighted avg	0.77	0.78	0.77	2100

Classification report of test data

	precision	recall	f1-score	support
0	0.77	0.92	0.84	605
1	0.72	0.42	0.53	295
accuracy			0.76	900
macro avg	0.74	0.67	0.68	900
weighted avg	0.75	0.76	0.74	900

ARTIFICIAL NEURAL NETWORK

Best grid search using ANN
{'activation': 'relu',
 'hidden_layer_sizes': (100, 100, 100),
 'max_iter': 10000,
 'solver': 'adam',
 'tol': 0.01}

Classification report of train data

	precision	recall	f1-score	support
0	0.85	0.86	0.85	1471
1	0.66	0.64	0.65	629
accuracy			0.79	2100
macro avg	0.75	0.75	0.75	2100
weighted avg	0.79	0.79	0.79	2100

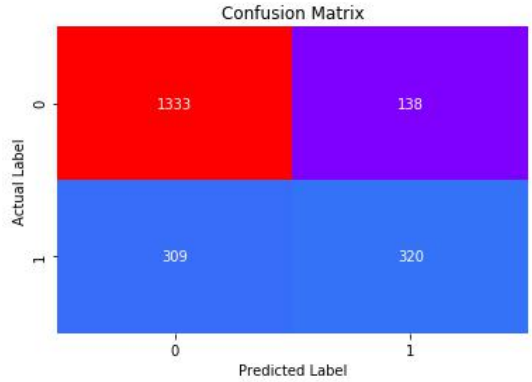
Classification report of test data

	precision	recall	f1-score	support
0	0.79	0.88	0.84	605
1	0.69	0.53	0.60	295
accuracy			0.77	900
macro avg	0.74	0.71	0.72	900
weighted avg	0.76	0.77	0.76	900

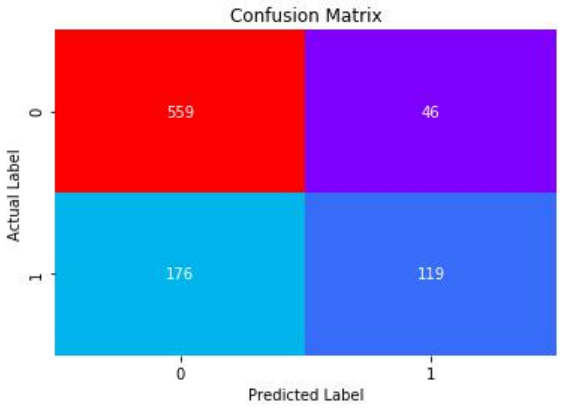
3.3 Performance Metrics: Performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model

RANDOM FOREST

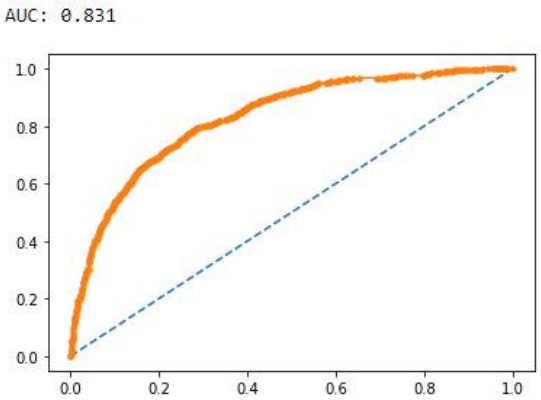
Confusion matrix of train data



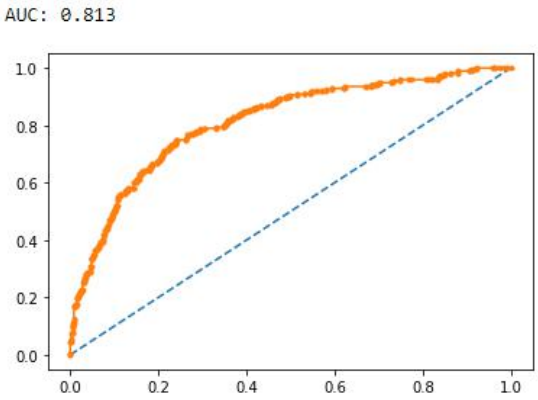
Confusion matrix of test data



AUC and ROC for the training data

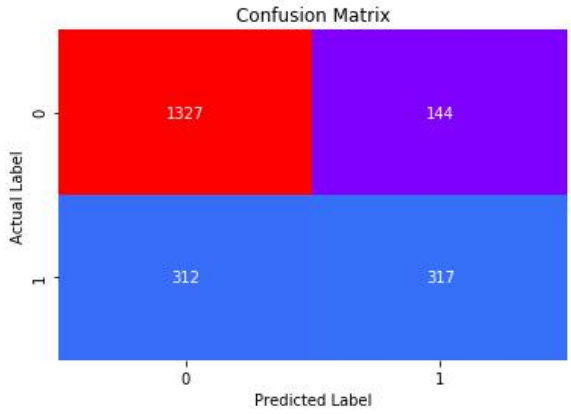


AUC and ROC for the testing data

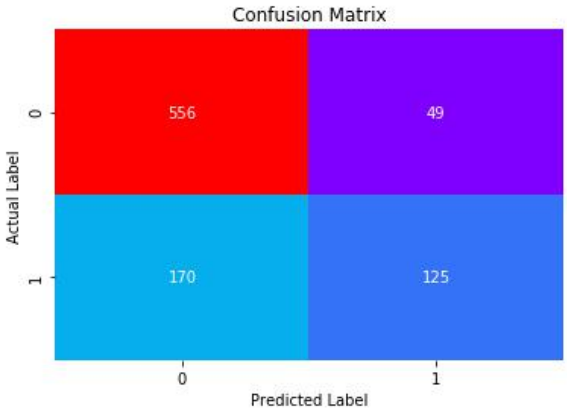


CART DECISION TREE

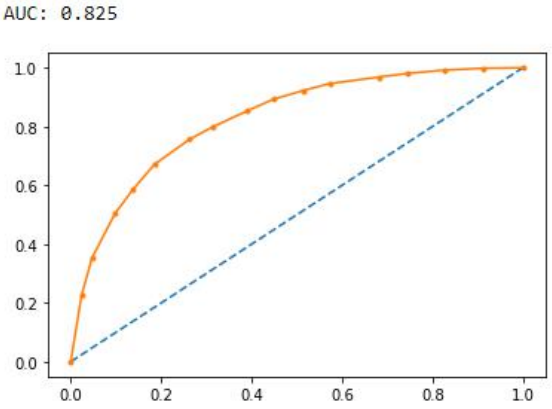
Confusion matrix of train data



Confusion matrix of test data

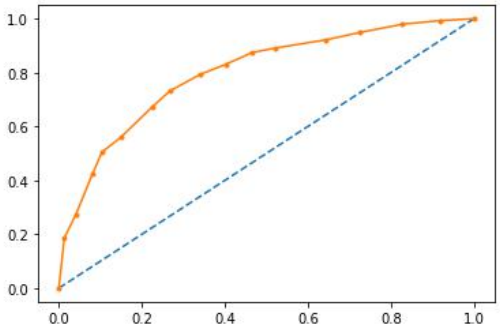


AUC and ROC for the training data



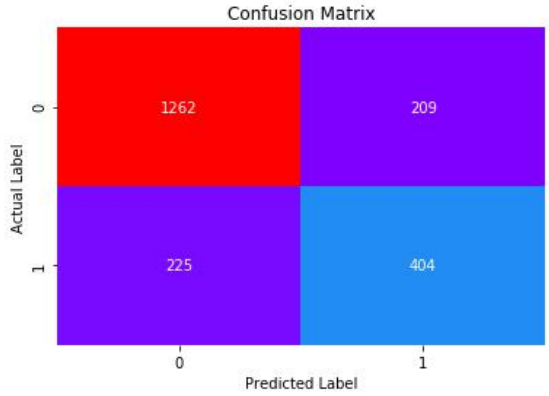
AUC and ROC for the testing data

AUC: 0.799

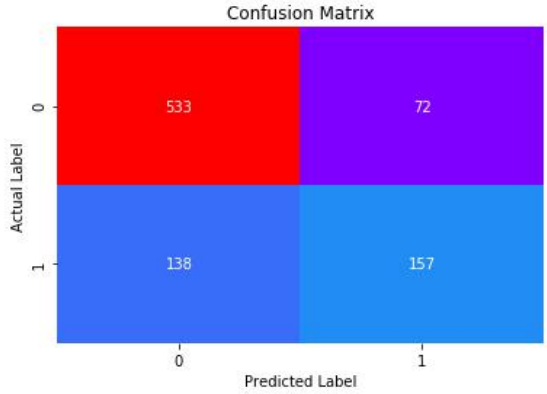


ARTIFICIAL NEURAL NETWORK

Confusion matrix of train data

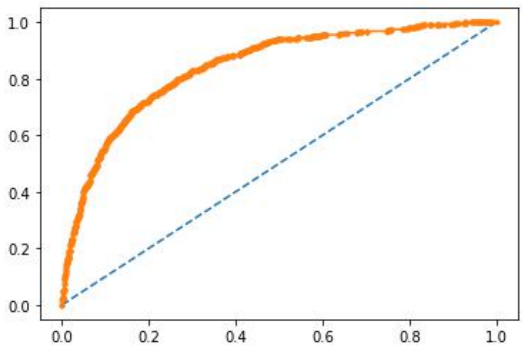


Confusion matrix of test data



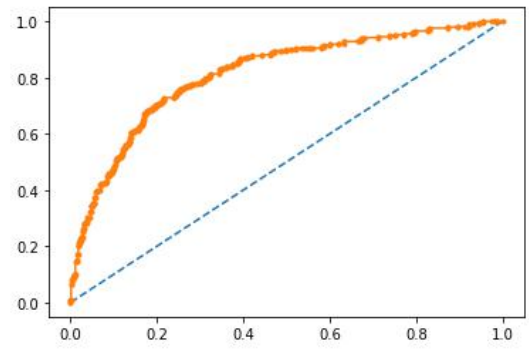
AUC and ROC for the training data

AUC: 0.843



AUC and ROC for the testing data

AUC: 0.813



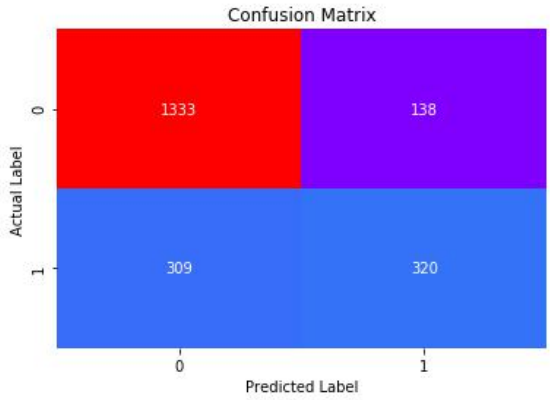
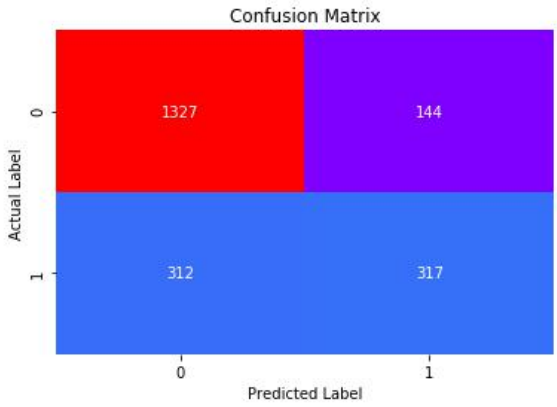
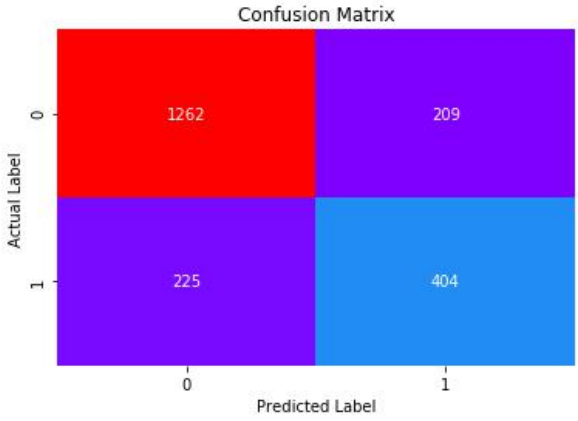
3.4 Final Model: Comparison of all models

Comparison of classification report for train data					
Random forest model		precision	recall	f1-score	support
	0	0.81	0.91	0.86	1471
	1	0.70	0.51	0.59	629
	accuracy			0.79	2100
	macro avg	0.76	0.71	0.72	2100
	weighted avg	0.78	0.79	0.78	2100
Decision tree model		precision	recall	f1-score	support
	0	0.81	0.90	0.85	1471
	1	0.69	0.50	0.58	629
	accuracy			0.78	2100
	macro avg	0.75	0.70	0.72	2100
	weighted avg	0.77	0.78	0.77	2100
Artificial neural network model		precision	recall	f1-score	support
	0	0.85	0.86	0.85	1471
	1	0.66	0.64	0.65	629
	accuracy			0.79	2100
	macro avg	0.75	0.75	0.75	2100
	weighted avg	0.79	0.79	0.79	2100

1. The artificial neural network model predicts highest precision of 0.85 for not claiming but least precision of 0.66 for claiming.
2. Random forest and Decision tree model predicts same precision of 0.81 for not claiming but Random forest predicts highest precision of 0.7 for claiming.
3. But however accuracy for all model merely same.

3.5 Inference: Business insights and recommendations

The following table describes the confusion matrix of train data for all 3 models

Random forest model	
Decision tree model	
Artificial neural network model	

Artificial neural network model delivers comparatively better confusion matrix as it as it shows higher true negative but however Random forest model shows higher true positive.

The majority of people traveling through travel agency doesn't claim insurance. Among people traveling through airlines, people who claim their insurance are nearly equal to those who doesn't claim insurance.

4 Appendix A – Source Code

Karthiheswar_DataMining.ipynb