Data Mining

25th July 2021

Vinay Santosh
PGP Data Science and Business Analytics
March_B 2021 (Online)

CONTENTS:

TABLES:

FIGURES:

Problem 1: Customer Segmentation

1.1 Exploratory Data Analysis:

The given data set consists of various activities of customers, at a leading bank, over the past few months. The goal is to segment customers based on their credit card activities in order to provide promotional offers to them. The following are the list of the customer's credit card activities:

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)

	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

Table 1.1 a

```
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)
```

Figure 1.1 a

The given dataset consists of 210 entries, all of which are of numeric data types, and 7 different features based on the customer's credit card activities. There are no missing values or duplicate entries found in the dataset. From the description of the given data, we can observe that there is a significant difference in values of mean, standard deviation, minimum and maximum between the various features.

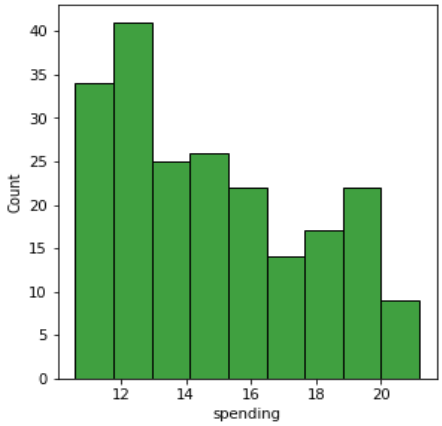
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	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

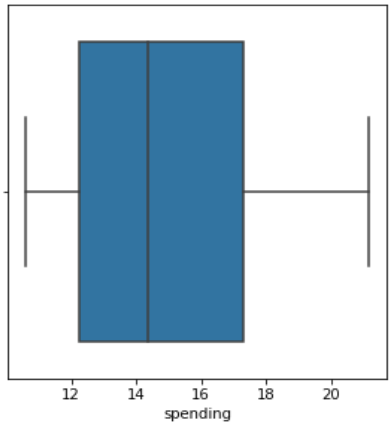
Table 1.1 b

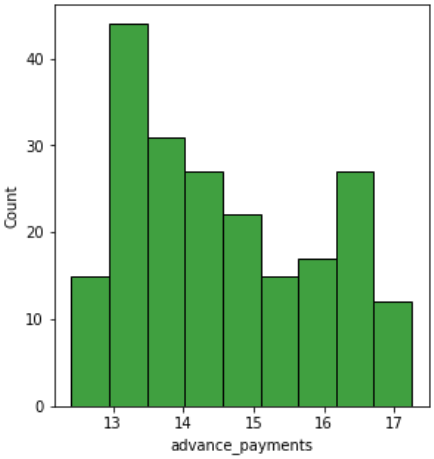
Univariate Analysis:

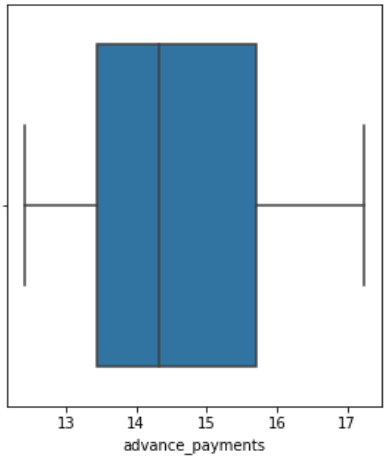
By performing analysis on separate features individually, we can observe how they are distributed across the dataset and check for any underlying outliers. The table (Table 1.1 c) given below displays the histogram and boxplot of the features individually. The distribution seems to be similar for most of the features with less to no outliers present in them.

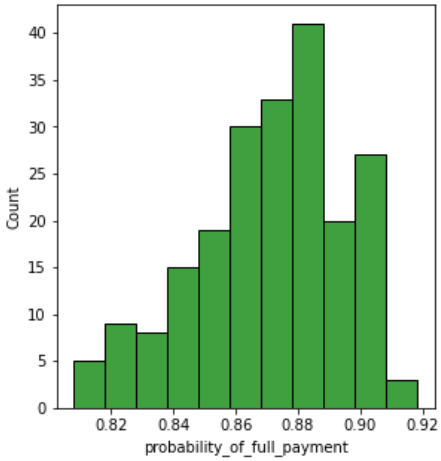
The features “probability_of_full_payment” and “min_payment_amt” are the only two features that have a few outliers and skewness in them. This shows that there is a larger group of customers that have a higher probability of paying the full amount and customers who have a lesser minimum amount to be paid tend to pay off the required amount.

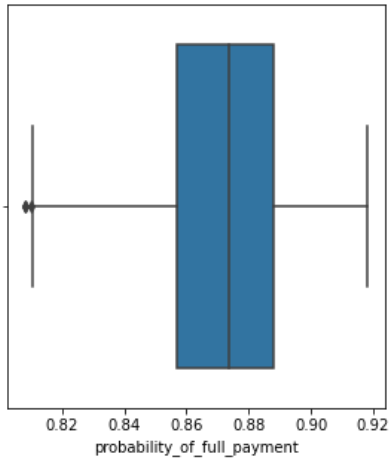


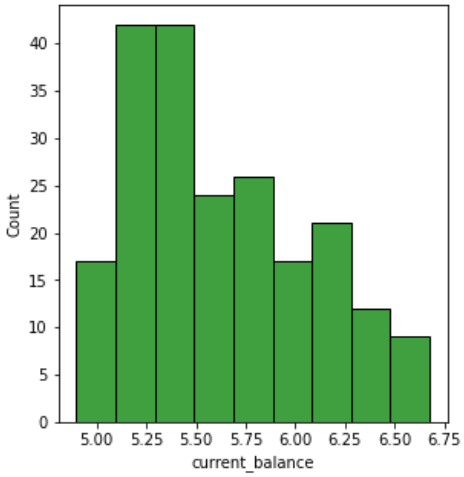


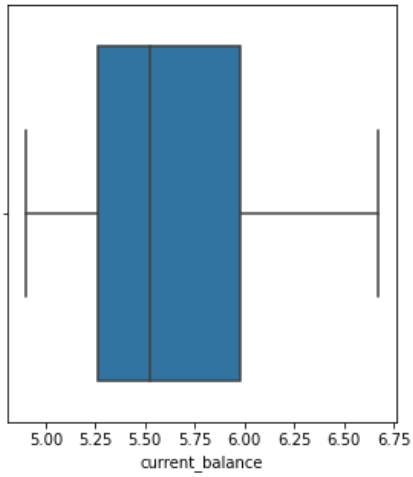


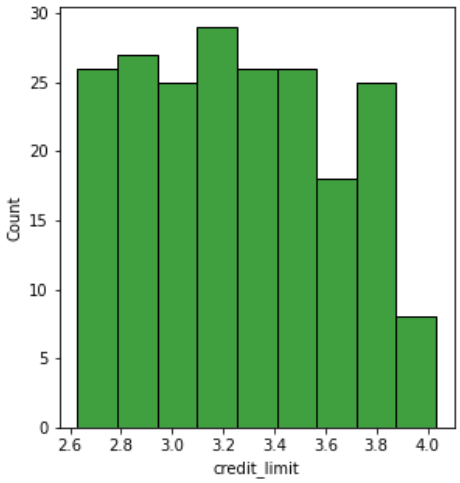


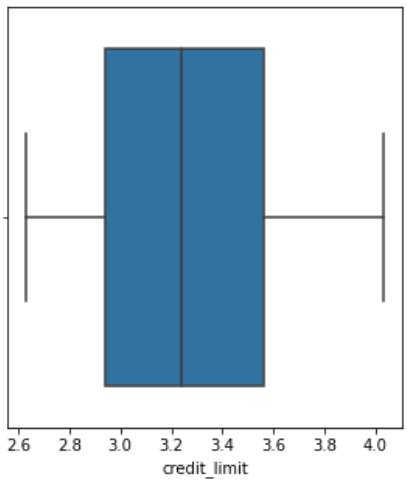


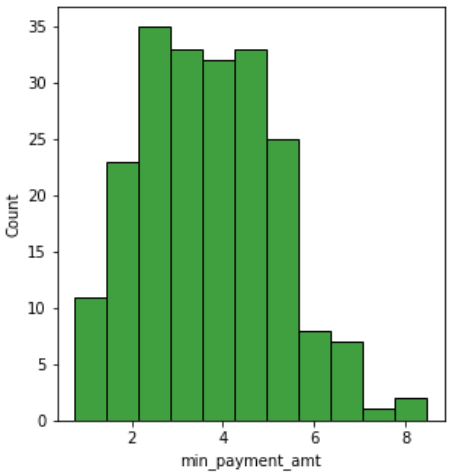


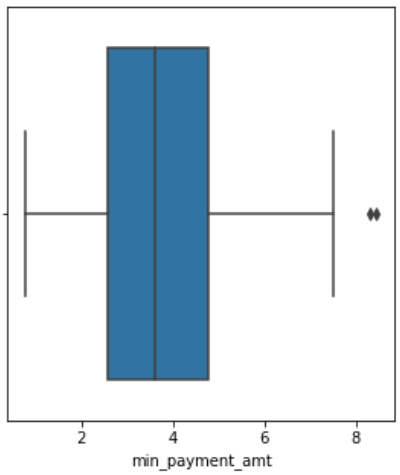

Count
current_balance


current_balance


Count
credit_limit


credit_limit


Count
min_payment_amt


min_payment_amt

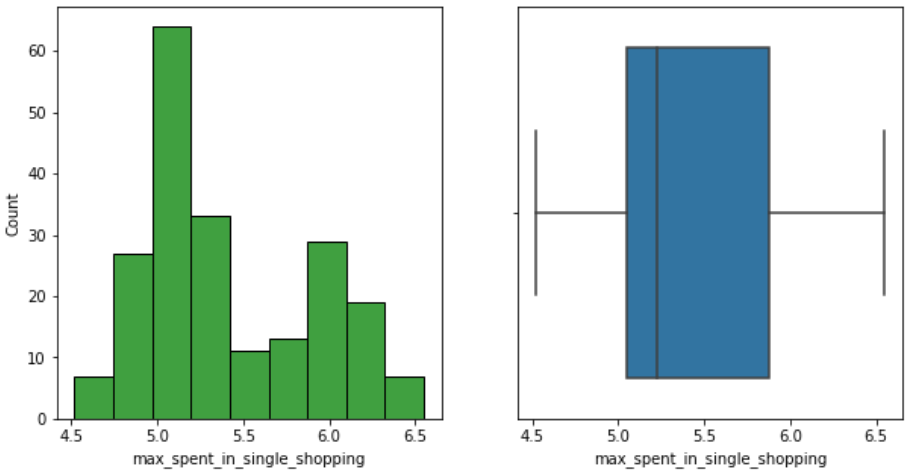


Table 1.1 c

Multivariate Analysis:

- By performing multivariate analysis, we can determine the relationship of features and whether or not they influence each other.
- In this analysis, we plot a heatmap (Figure 1.1 b) that displays the correlation between the various features. The values range from -1 to +1.
- Features with darker colors have a lower correlation(closer to -1) and lighter colors have a higher correlation(closer to +1).
- The features, “spending” and “advance_payments” show the high correlation with most of the other features while “min_payment_amt” shows the least correlation with most of the other features.

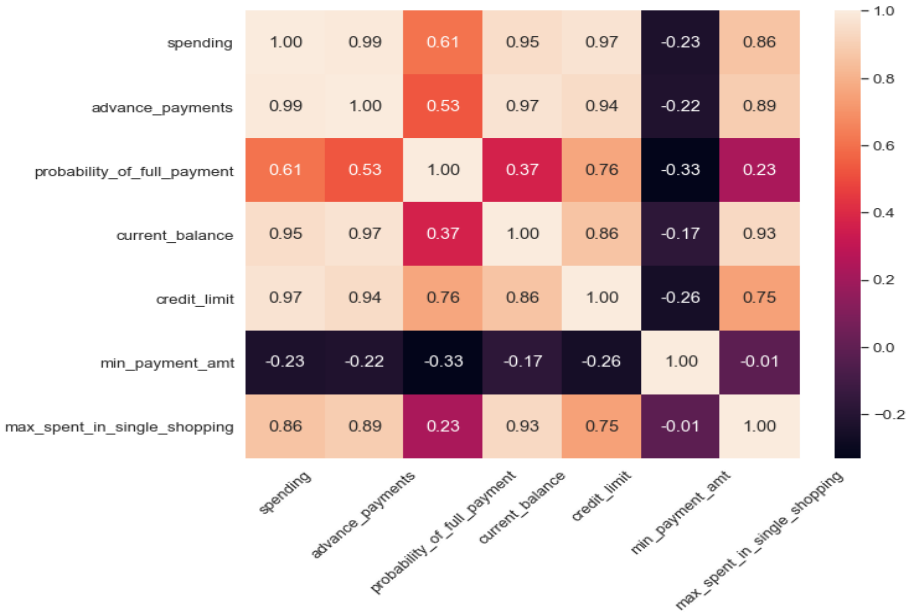


Figure 1.1 b

- The figure below (Figure 1.1 c) displays a scatterplot between each of the given variables in the off diagonals. We can observe a linear relationship with most of the features which indicate that they are highly correlated with each other.

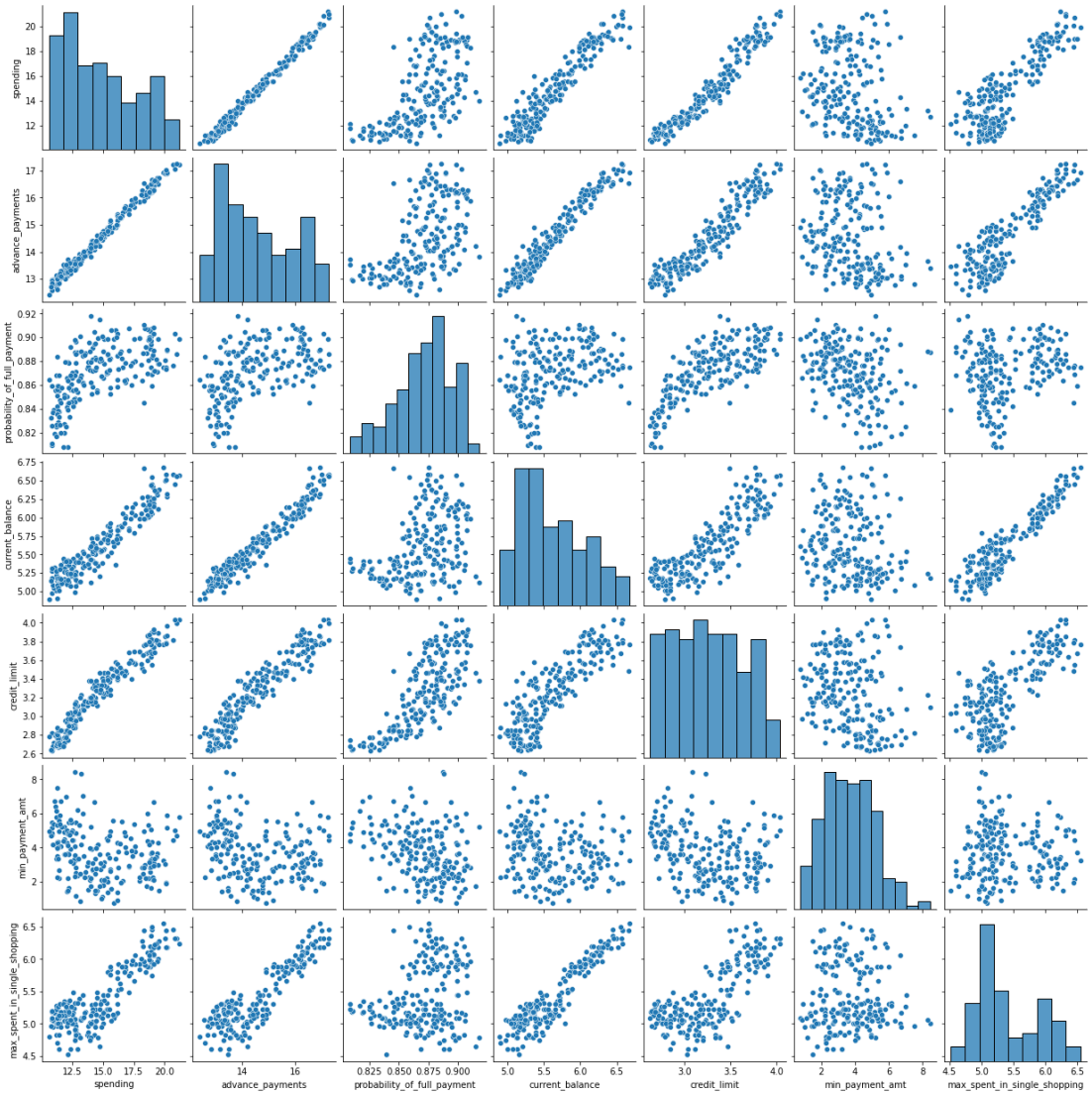


Figure 1.1 c

1.2 Data Scaling:

- Since clustering is based on distance metrics, scaling or normalizing of data is required. From the description of the dataset, we can observe differences in the minimum and

maximum values as well as irregularities in the means and standard deviation of some other features.

- The magnitude of some of the features vary and scaling will help in bringing their means closer to 0 and standard deviation closer to 1.

1.3 Hierarchical Clustering:

- Hierarchical Clustering is a distance-based algorithm that is used for segmentation problems.
- In this algorithm, each entry is initially treated and a cluster and its distances are measured. Further clusters are formed based on the distances until all of the entries are fitted into one cluster.
- This technique uses a process of linkage in order to measure the distances between the clusters.
- For our problem statement, we implement Ward's linkage method which creates compact and even-sized clusters, ideal for smaller datasets.
- Upon performing Ward's linkage, we arrive at a truncated dendrogram which is a tree-like graph that visualizes the clustered data. From the dendrogram, we observe that there are two main clusters (orange and green).
- The numbers below represent the number of rows or entries in that cluster.
- We can decide the number of clusters required, using the Fcluster method, and generate the output which can be fitted to our dataset.
- This can help us in further analysis by choosing the required cluster.

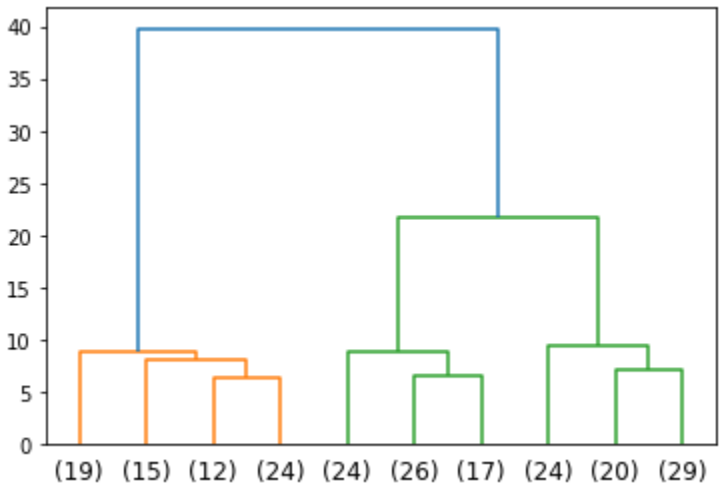


Figure 1.3

1.4 K-Means Clustering:

- K-means Clustering is a non-hierarchical approach to clustering or segmenting data.
- We assign a certain K value that pre-determines the number of clusters the dataset will be segmented into.
- Once the segmentation is done, the within-cluster sum of squares is calculated for each K value.
- These values are then plotted using an elbow plot to visualize the variation when the K values change.
- From the figure below(FIgure 1.4 a), we can notice a significant variation in the WSS when the K value changes from 1 to 2.
- The variation continues from 2 to 3 but is not as steep as the initial drop. This gives us an indication of which K value to choose for the given problem.

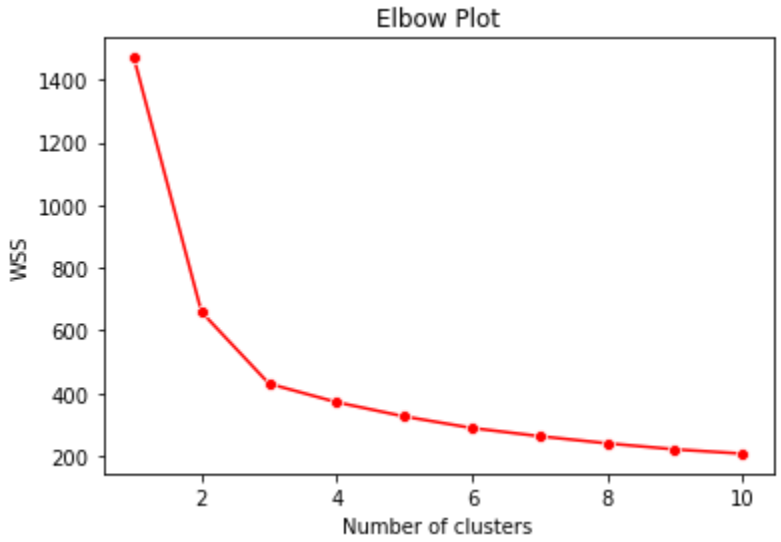


Figure 1.4

- To find out whether the chosen cluster produces the optimum results, we calculate the silhouette score. This measures how tightly the observations are clustered and the average distance between them.
- A silhouette score closer to +1 means that it is tightly bound. A score less than 0 suggests that we try another value for k.
- The silhouette width is calculate for each cluster and we check if there are any negative values.

- Upon calculating the various WSS for k values ranging from 1 - 10 as well as the silhouette score and width, the ideal value of k for this problem is 3.
- The cluster labels are extracted and attached to the original dataset. We can now segment the data and extract insights.
- The table below shows the head of the original dataset with the extracted cluster labels attached.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	2
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	0
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	2
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	1
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	2

Table 1.4

1.5 ClusterProfiles, Insights, and Recommendations:

- The table below (Table 1.5) shows the mean values of all the variables in different clusters.
- We can observe that Cluster1 consists of the group of customer who spend the least on their credit cards and have lower credit limits.
- On an average Cluster1 has the highest minimum amount paid monthly towards their credit cards. This shows that the customers in Cluster1 spend less and are consistent in their credit card payments.
- Customers in Cluster2 have the highest spending patterns among the three clusters. The also tend to have a higher credit limit and maximum amount of money spent in a single purchase. This indicates that Cluster2 are high spenders and use their credit cards more frequently.

Cluster0:	**Cluster1:**	**Cluster2:**
spending 14.437887 **advance_payments** 14.337746 **probability_of_full_payment** 0.881597 **current_balance** 5.514577 **credit_limit** 3.259225 **min_payment_amt** 2.707341 **max_spent_in_single_shopping** 5.120803	**spending** 11.856944 **advance_payments** 13.247778 **probability_of_full_payment** 0.848253 **current_balance** 5.231750 **credit_limit** 2.849542 **min_payment_amt** 4.742389 **max_spent_in_single_shopping** 5.101722	**spending** 18.495373 **advance_payments** 16.203433 **probability_of_full_payment** 0.884210 **current_balance** 6.175687 **credit_limit** 3.697537 **min_payment_amt** 3.632373 **max_spent_in_single_shopping** 6.041701

Table 1.5 a

- From the scatter plots given below (Table 1.5 b), we can observe the various spending patterns of different clusters.
- Customers in Cluster1 have lower credit limits and a lower probability of paying the amounts in full. However, they have a higher rate of paying the minimum amount.

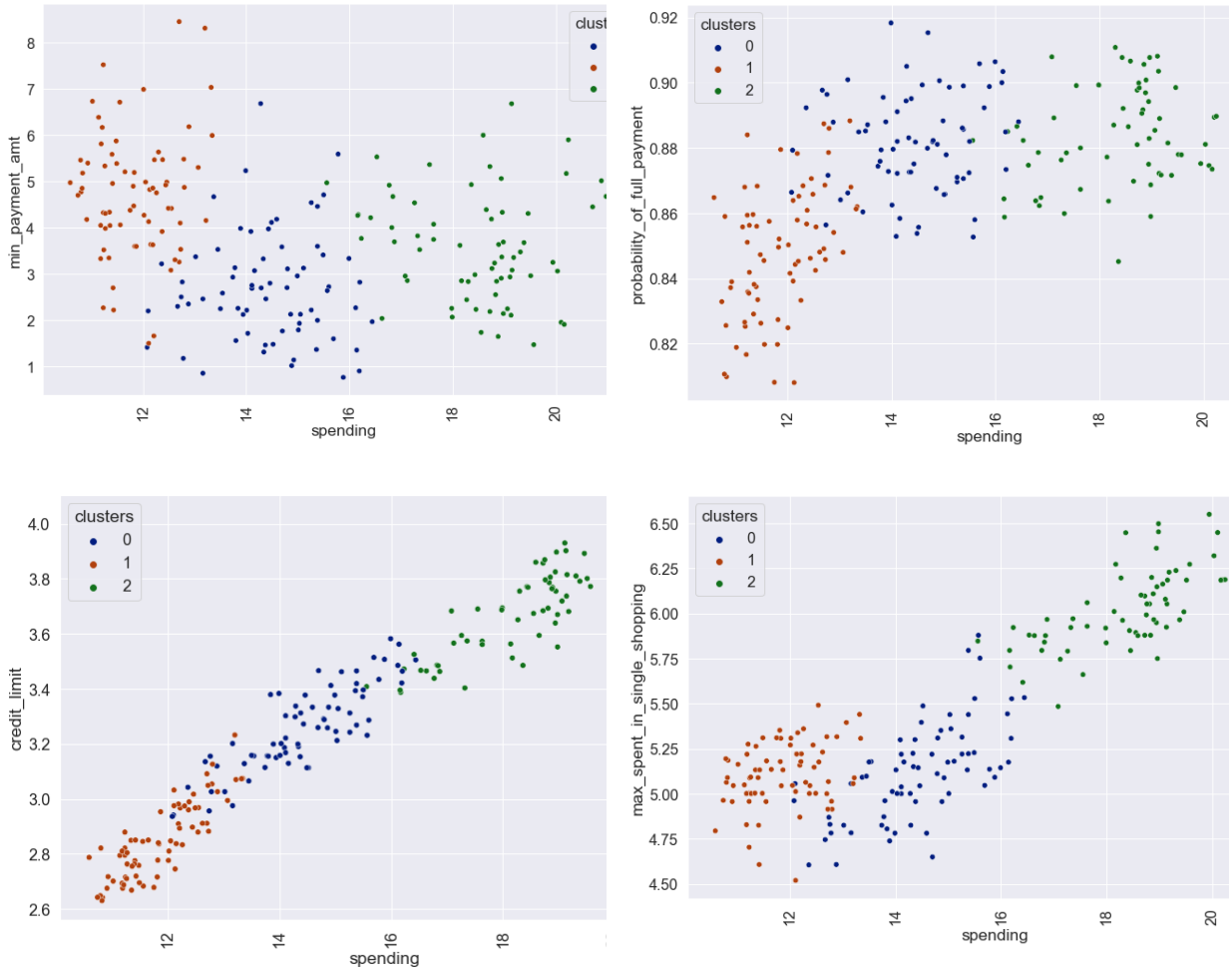


Table 1.5 b

- Based on the above insights and cluster profiles, we can promote special offers to specific targets.
- Customers in Cluster1 spend less and maintain a good balance with consistent bill payments. We can hence offer a special savings plan and special discounts for using their credit cards for shopping.
- Customers in Cluster2 are high spenders who use their credit cards frequently. The customers in this cluster tend to pay their bills on time and make higher advanced payments.

- We can offer an increase in credit limit or provide an upgrade to new credit cards with competitive interest rates and features.
- An offer with a rewards scheme for making purchases using credit cards is recommended to customers in Cluster0.

Problem 2: Predicting Insurance Claim Status using CART, RF and ANN

2.1 Exploratory Data Analysis:

- The given dataset, for this problem, consists of data from tour insurance firms over the past few years. The goal is to build a model for an insurance firm that wants to predict the future claims based on the given data.
- There are 10 different features present in the data that provides information on the insurance plan, travel destination, age of the insured etc. The table below (Table 2.1 a) shows the top five rows of the dataset with all the features.

	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

Table 2.1 a

- The description of the data provides us information on the average values for numeric data and the most common categorical type.
- We can observe from the table below that most customers prefer a customized plan for insurance which they choose online through a travel agency. The most preferred destination in Asia.
- The highest sale generated from tour insurance is 539 with an average of 60. The maximum commission received from the sale of insurance is 210.2 with an average of 14.5 per sale.
- The majority of the customers have not claimed their insurance. The duration of the tours differ significantly with each sale.

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000.000000	3000	3000	3000	3000.000000	3000	3000.000000	3000.000000	3000	3000
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2076	NaN	2954	NaN	NaN	1136	2465
mean	38.091000	NaN	NaN	NaN	14.529203	NaN	70.001333	60.249913	NaN	NaN
std	10.463518	NaN	NaN	NaN	25.481455	NaN	134.053313	70.733954	NaN	NaN
min	8.000000	NaN	NaN	NaN	0.000000	NaN	-1.000000	0.000000	NaN	NaN
25%	32.000000	NaN	NaN	NaN	0.000000	NaN	11.000000	20.000000	NaN	NaN
50%	36.000000	NaN	NaN	NaN	4.630000	NaN	26.500000	33.000000	NaN	NaN
75%	42.000000	NaN	NaN	NaN	17.235000	NaN	63.000000	69.000000	NaN	NaN
max	84.000000	NaN	NaN	NaN	210.210000	NaN	4580.000000	539.000000	NaN	NaN

Table 2.1 b

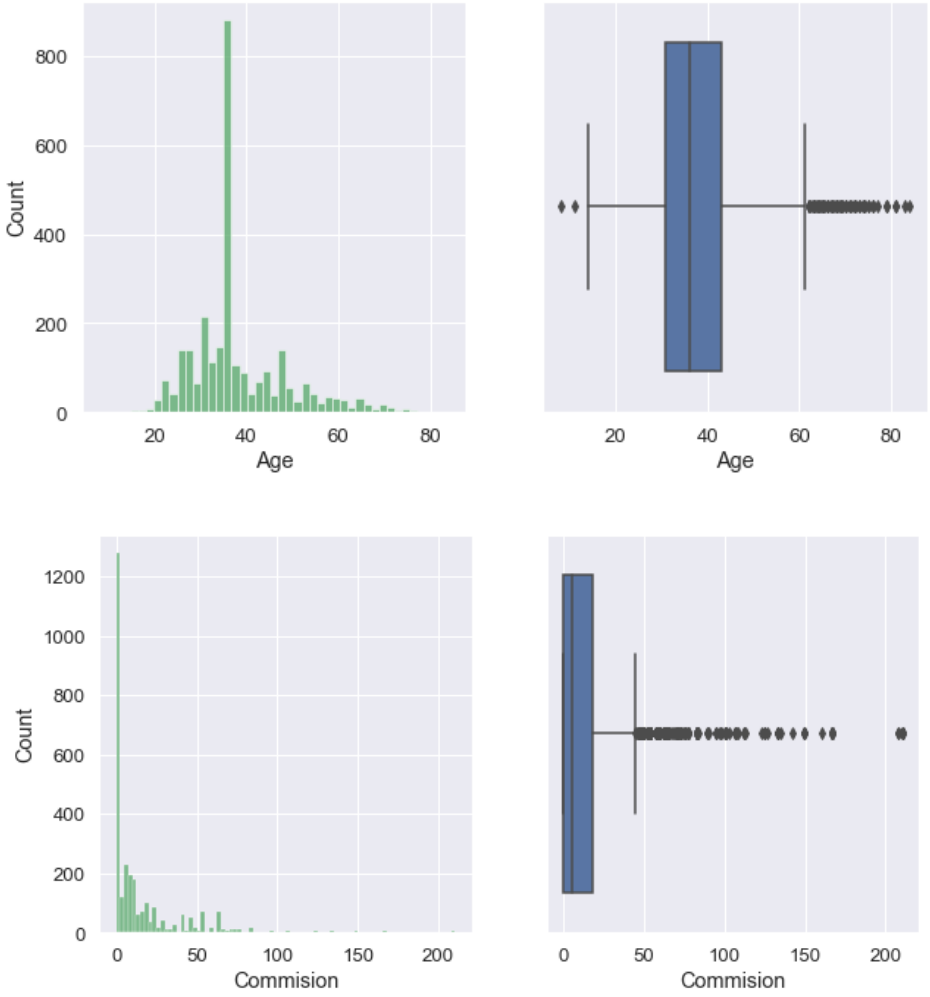
- From the information below (Figure 2.1 a), we can see that there are 3000 entries with 10 features as mentioned earlier.
- There are 6 features of “object” data-type and 4 of numeric data-types.
- The dataset also contain 139 duplicate entries which are dropped for ease of analysis. There are no null values present in the dataset.
- The shape of the dataset after dropping the duplicates will be (2861, 10)

```
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)
```

Figure 2.1 a

Univariate Analysis:

- By performing univariate analysis on the numeric data, we can observe it's distribution and underlying outliers if any.
- From the table given below (Table 2.1 c), we can notice that there are outliers present in all of the numeric variables; most of them are present above the upper quartile range (on the higher side).
- We notice that Sales, Duration and Commission are heavily skewed to the left.



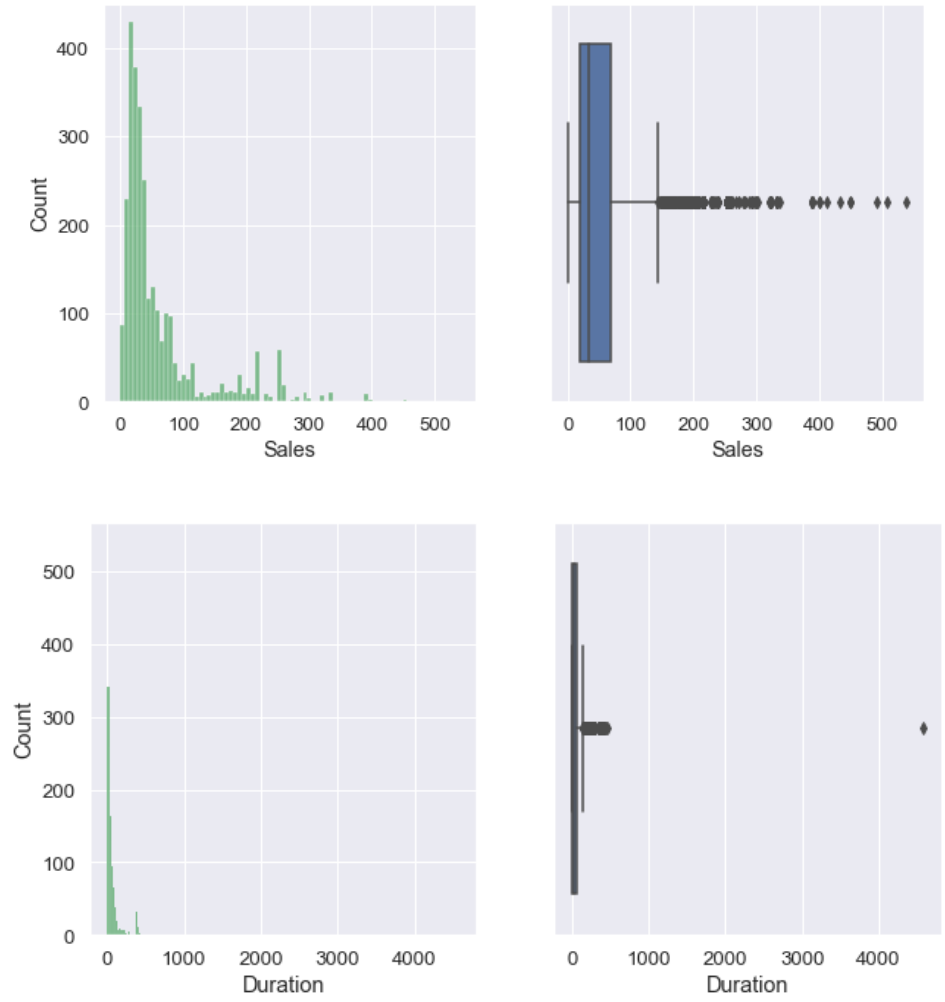


Table 2.1 c

The table (Table 2.1 d) below shows the categorical variables present in the dataset. We can see the distribution of different values in each category. We can draw the following insights from the plots:

- We can observe that the majority of the customers prefer to purchase insurance from a travel agency as compared to the insurance provided by the airlines.
- Almost all the purchases are made online. As mentioned earlier, Asia seems to be the most commonly visited destination.
- Customers prefer a customised plan the most followed by cancellation plan. 69% of the customers have not claimed their insurance.

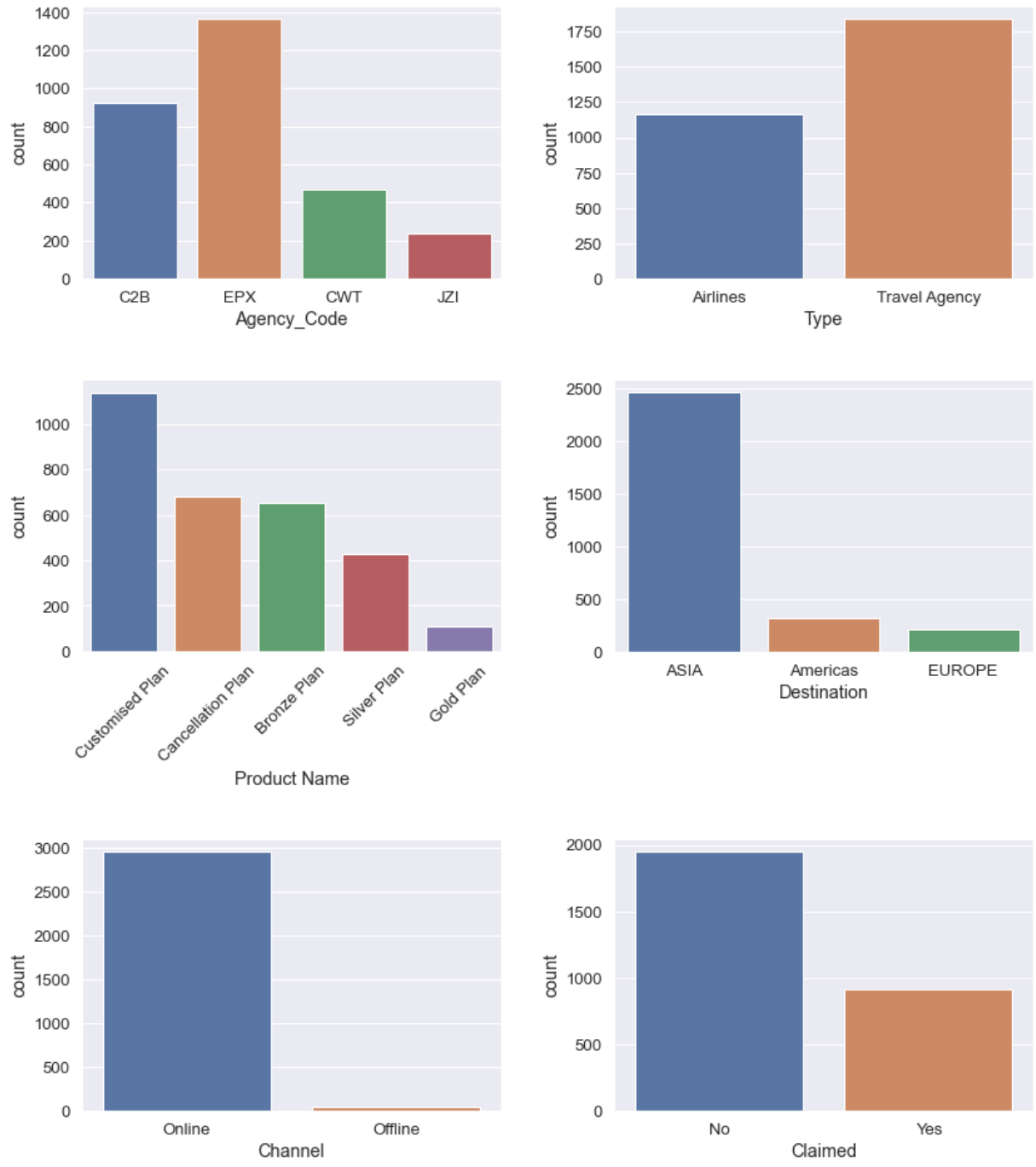


Table 2.1 d

Multivariate Analysis:

- We can observe the correlation between the features with the help of a heatmap.
- Most of the features do not display a high correlation with each other except for the feature “Sales” which has a high correlation with “Commission” indicating that higher sales leads to higher commission for the insurance firm.

- The Duration feature does not seem to vary much and has an extreme outlier. There is no negative correlation between any of the features.

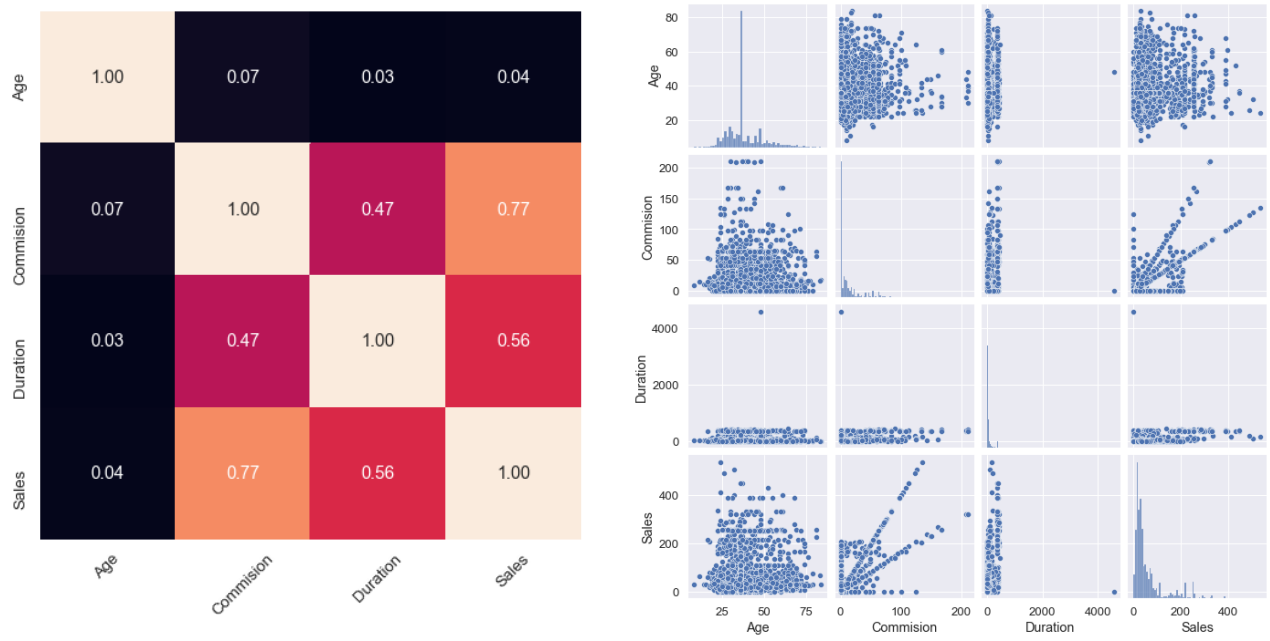


Table 2.1 e

2.2 Model Building and Feature Importance:

Splitting the Dataset:

- To build prediction models, we must split the given dataset into training and testing sets.
- We define the target variable separately in order to make predictions on them later.
- For this given problem, we split the data by the ratio 7:3. That is to say 70% of the data is used for training and 30% for testing.
- We use a random_state to ensure that the values generated remain the same at different instances of running the program.
- Upon splitting the data, the train set has a total of 2002 entries while the test set consists of 859 entries.
- We fit the split data in order to train and test the data.

Grid Search and building Models:

- To build the various predictive models, we pass in certain parameters for each model we build.

- These parameters are measured based on the dataset at hand for different models. We can pass in a group of values for each parameter that may fall within the range to predict the best results.
- Once the chosen parameters are set for the models we implement a grid search for each of the algorithms.
- Cross-validation is an import parameter that is passed to the model inorder to test various combinations of parameters. This helps in identifying overfitting of data.
- The best grid parameters extracted for the different models are given in the table below:

Decision Tree Classifier:	**Random Tree Classifier:**	**ANN:**
criterion: " gini"	**max_depth: 10**	**hidden_layer_sizes: 300**
max_depth: 10	**max_features: 6**	**max_iter: 5000**
min_samples_leaf: 50	**min_samples_leaf: 5**	**solver: "adam"**
min_samples_split: 300	**min_samples_split: 50**	**tol: 0.01**
	n_estimators: 300	

Table 2.2

- On running the models, we can determine probabilities of the target class and predict the outcomes of the test data.
- For the decision tree classifier and random forest classifier, we can extract the importance of features in the model.
- The figures (Figure 2.2 a and b)below are graphs of feature importance in the decision tree classifier and random forest respectively.
- These graphs show the percentage of importance that each feature plays in the model.
- We can observe that the decision tree classifier does not use all of the features to predict the results.
- "Agency_Code" and "Sales" contribute the highest in both of the classifiers.

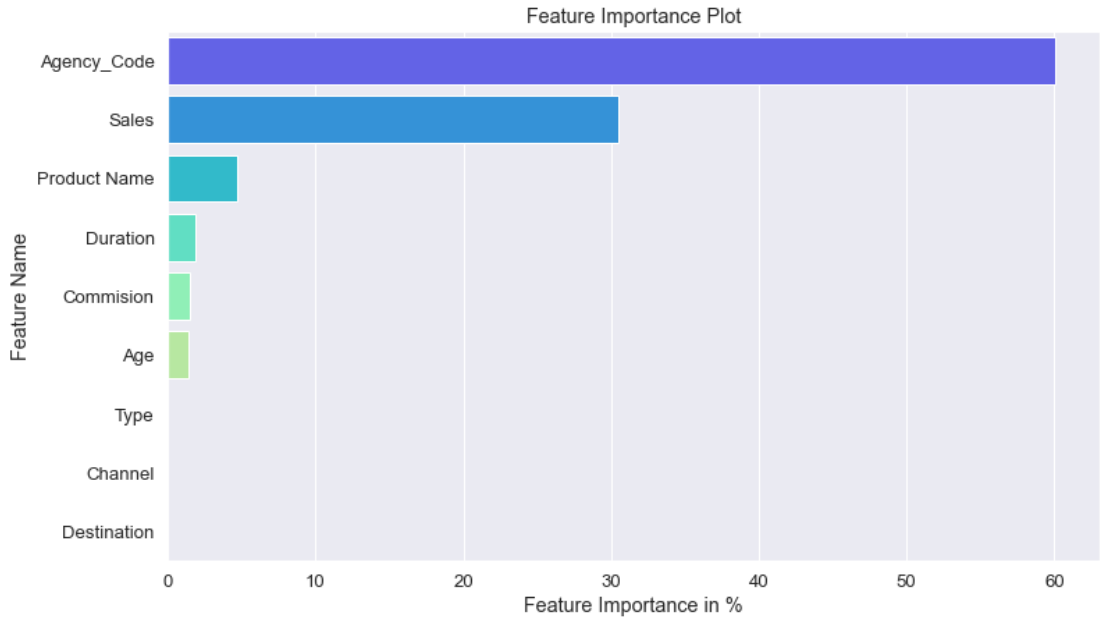


Figure 2.2 a

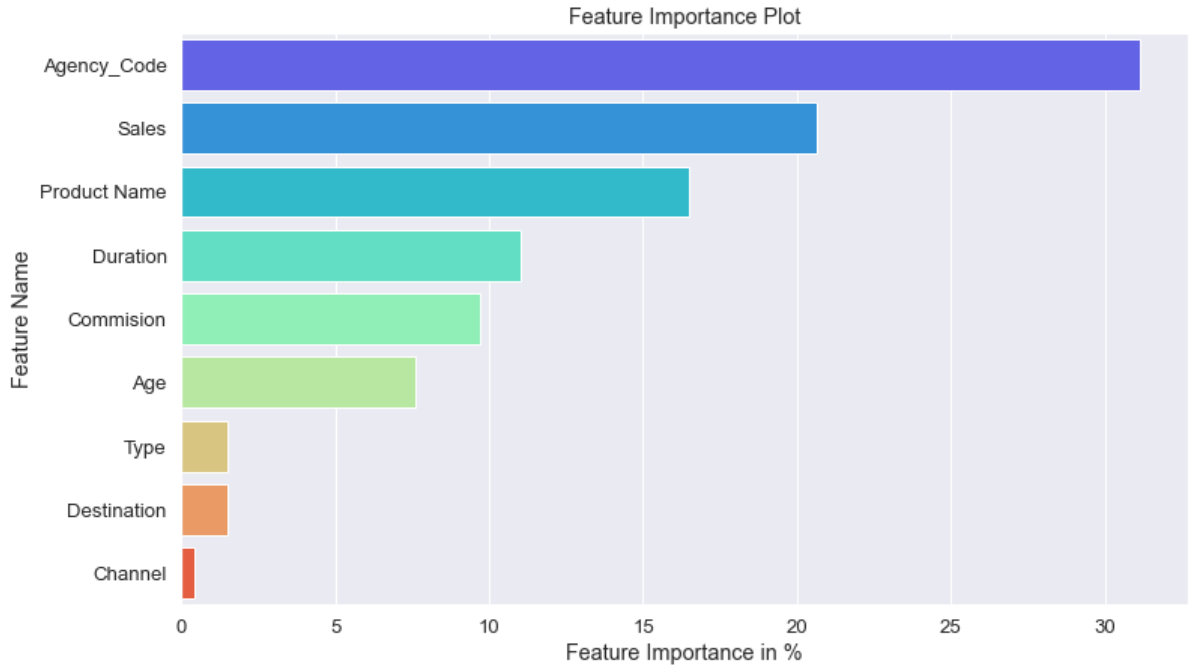


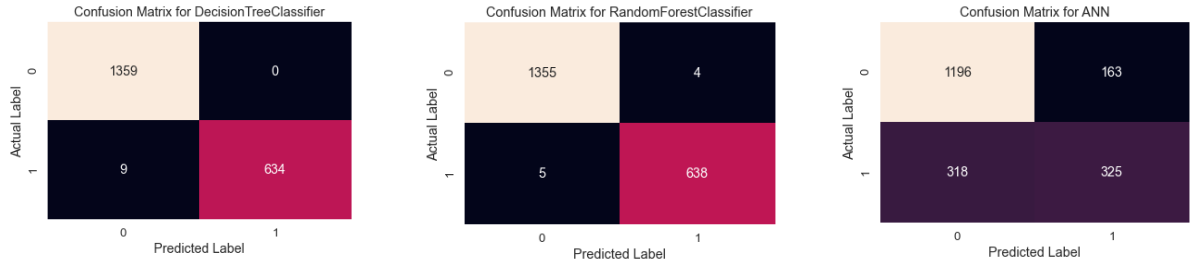
Figure 2.2 b

2.3 Model Evaluation:

- We can see how well the built model predicts based on certain performance metrics like accuracy, precision, classification reports and confusion matrices.

- The confusion matrix is a matrix of right and wrong predictions. The predictions are categorised as True Positives, False Positives, False Negatives and True Negatives.
- The True Positives and True Negatives are the number of predictions the model guessed correct. False Negatives are predictions made by the model as true but in reality were false. In our case, did the customer claim for insurance or not.
- We can observe from the confusion matrices that in the train set, CART and RF models predicted almost all of the samples correctly whereas ANN has a lot of incorrect predictions (higher number of False Negatives) of claim status.

Train Set:



Test Set:

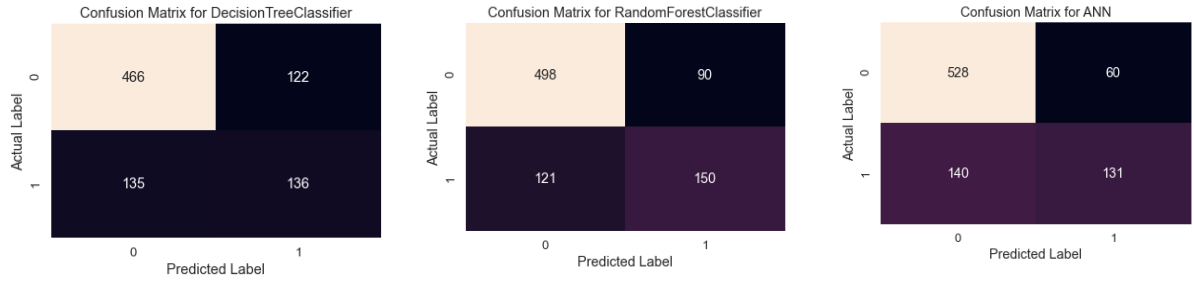


Table 2.3

- The confusion matrix of the test set, however, shows that the ANN does a better job at predicting customers who did not claim their insurance as compared to the other two models.
- The accuracy score of the models indicate how accurate the overall predictions in each model are. It is the ratio of the sum of right predictions to all of the predictions in the model.

- The train set of both CART and RF do an extremely good job in predicting the target class and have exceptional accuracy, precision and recall.
- The neural network model predicts 76% of the train set accurately.
- From the score and confusion matrices of the test set, we can observe that there is a drop in accuracy in the CART and RF models.
- The neural network performs slights better than it did with the training data.
- This indicates that CART and RF models have overfitted the training data and performs poorly on the training set.
- The ROC (Receiver Operating Characteristic) Curve plots the true positive rates versus the false positive rate.
- The area under the curve (AUC) is an aggregate measure of performance of the models across all thresholds.
- If the ROC_AUC score is closer to 1, we can say that the model is highly accurate in making predictions and if it's closer to 0, the model does not make accurate predictions.
- The figures given below show the ROC curves for both the train and test sets. We can observe that for the train set, both CART and RF have similar curves and overlap each other.

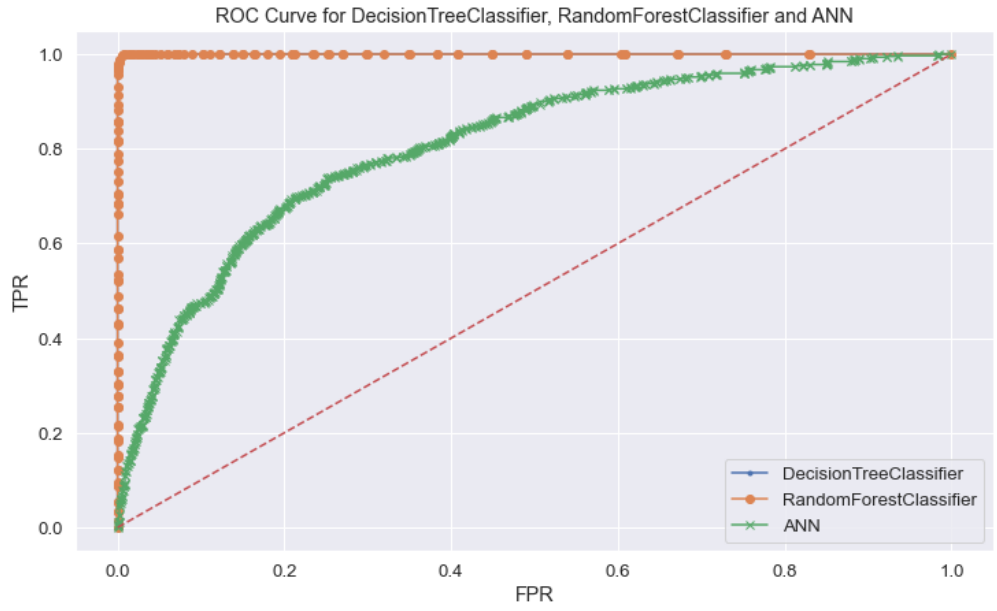


Figure 2.3 a

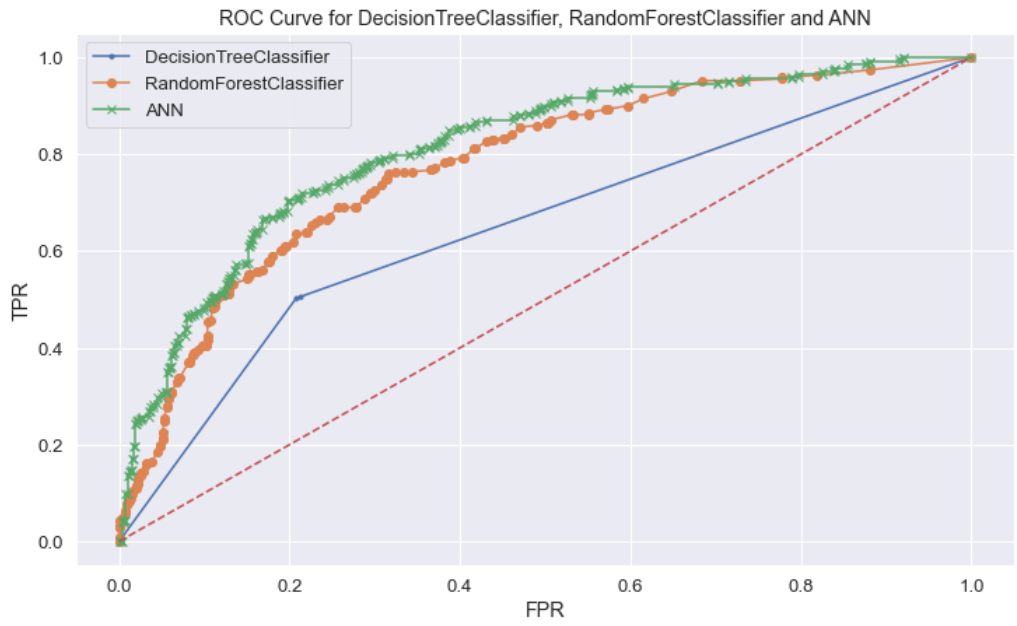


Figure 2.3 b

2.4 Final Model:

Comparing Model Performance Evaluations:

	Accuracy	Precision	Recall	F1 Score	ROC_AUC Score
CART	Train: 1.00 Test: 0.70	Train: 1.00 Test: 0.53	Train: 0.99 Test: 0.50	Train: 0.99 Test: 0.51	Train: 1.00 Test: 0.65
RF	Train: 1.00 Test: 0.75	Train: 0.99 Test: 0.62	Train: 0.99 Test: 0.55	Train: 0.99 Test: 0.59	Train: 1.00 Test: 0.78
ANN	Train: 0.76 Test: 0.77	Train: 0.67 Test: 0.69	Train: 0.51 Test: 0.48	Train: 0.57 Test: 0.57	Train: 0.81 Test: 0.81

Table 2.4

- By comparing the different performance metrics of training and testing sets, we can conclude that the CART and RF models overfits the data and does a poor job in predicting the customers who tend to claim tour insurance.

- Since the target proportions are imbalanced, the accuracy score does not play an important role in choosing the best model for the problem.
- When comparing the different AUC scores, we can observe that there is less to no difference in performance of the neural network model. Random forest has the next best closest AUC score for the test data.
- The recall is comparatively higher for random forest, which is essential as we need to reduce the number of false negatives in order to accurately predict whether future customers claim their insurance.
- Hence, based on the recall score, we can choose the **random forest classifier** to predict future insurance claims.

2.5 Inference and Recommendations:

- As the business requirement is to be able to predict the claim status of the customer.
- The feature importance of the random forest classifier shows the important variables that may help in predicting the claim status.
- Customers who travel for a longer duration of time tend to make an insurance claim.
- It is recommended that the insurance firm provides an altered policy for these travellers.
- They can also give the customer options to choose their insurance based on travel durations.
- It is evident that most customers choose customised plans. The firm can build a plan for the traveller's based on their itinerary
- The agency chosen by the travellers play the most significant role in identifying the future claims.
- Most claims come from certain agencies that can be contacted by the management and provide a revised policy for future customers.
- By improving sales at the various agencies providing the insurance can help recover from the increased claim in frequency.