June 2021 | PG-DSBA-Online

PROJECT REPORT
DATA MINING

SUBMITTED BY
DEV TRIPATHI

Contents

List of Figures

Problem 1

Bank Marketing Dataset Analysis

Information about dataset

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 Dictionary for Market Segmentation

The dataset contains the following features and the information about these features is mentioned below:

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)

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

EDA

Sample of 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: Sample of Bank Marketing Dataset

Variable Information-

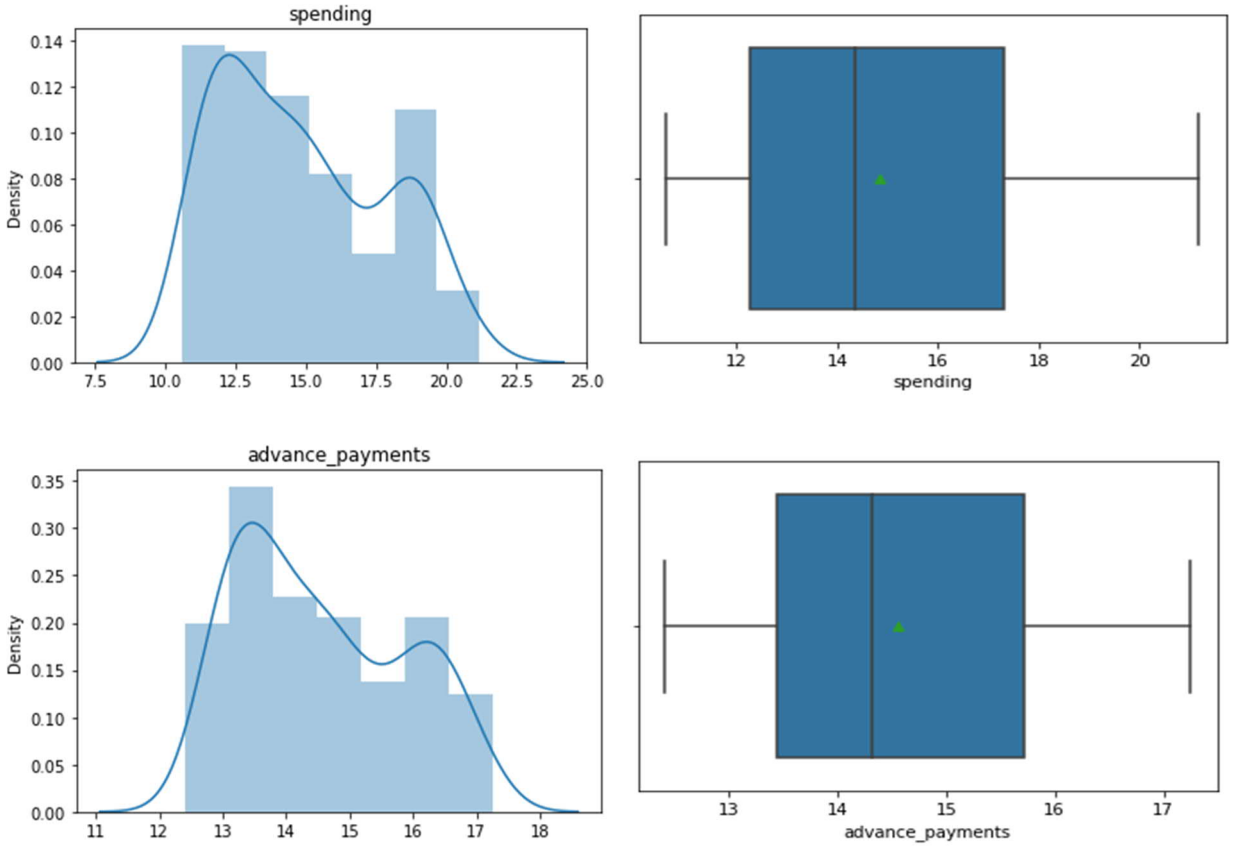
```
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 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
```

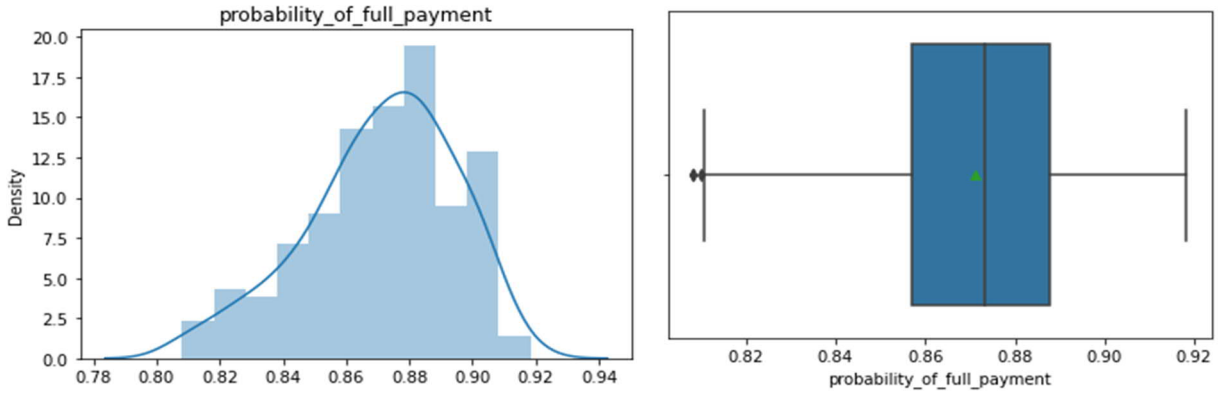
```
spending                      0
advance_payments              0
probability_of_full_payment   0
current_balance               0
credit_limit                  0
min_payment_amt               0
max_spent_in_single_shopping  0
```

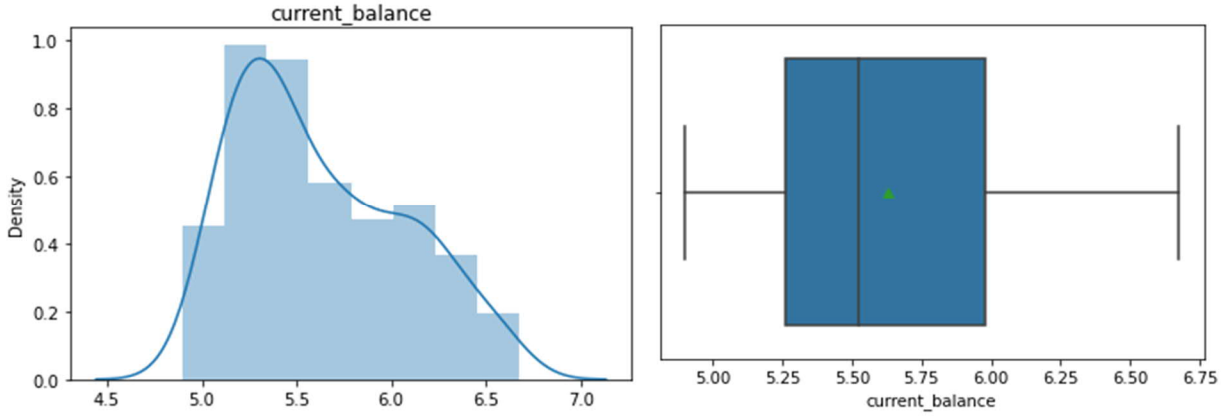
Figure 2: Feature Info and null values count of Bank Marketing Dataset

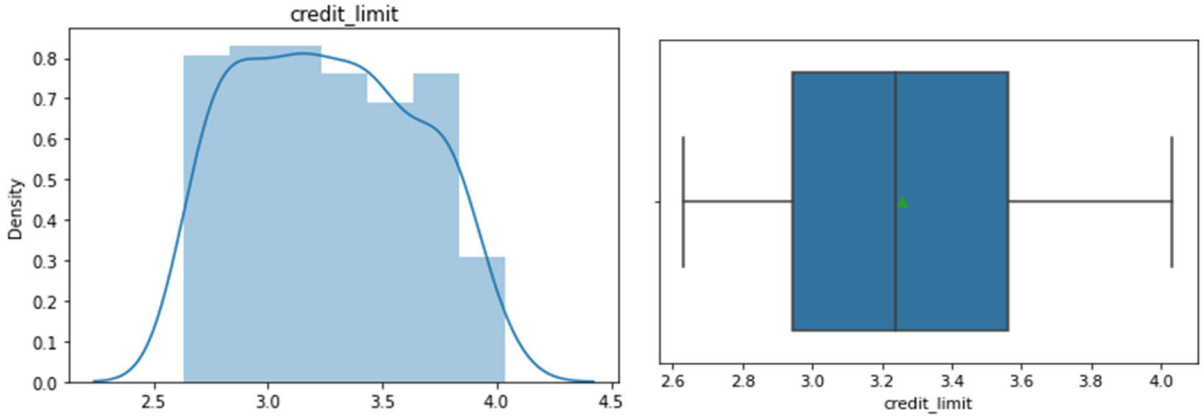
The dataset contains a total of 210 entries and 7 features with no missing values present. All the features are of 'float64' data type.

Univariate Analysis-




probability_of_full_payment
Density
probability_of_full_payment


current_balance
Density
current_balance


credit_limit
Density
credit_limit

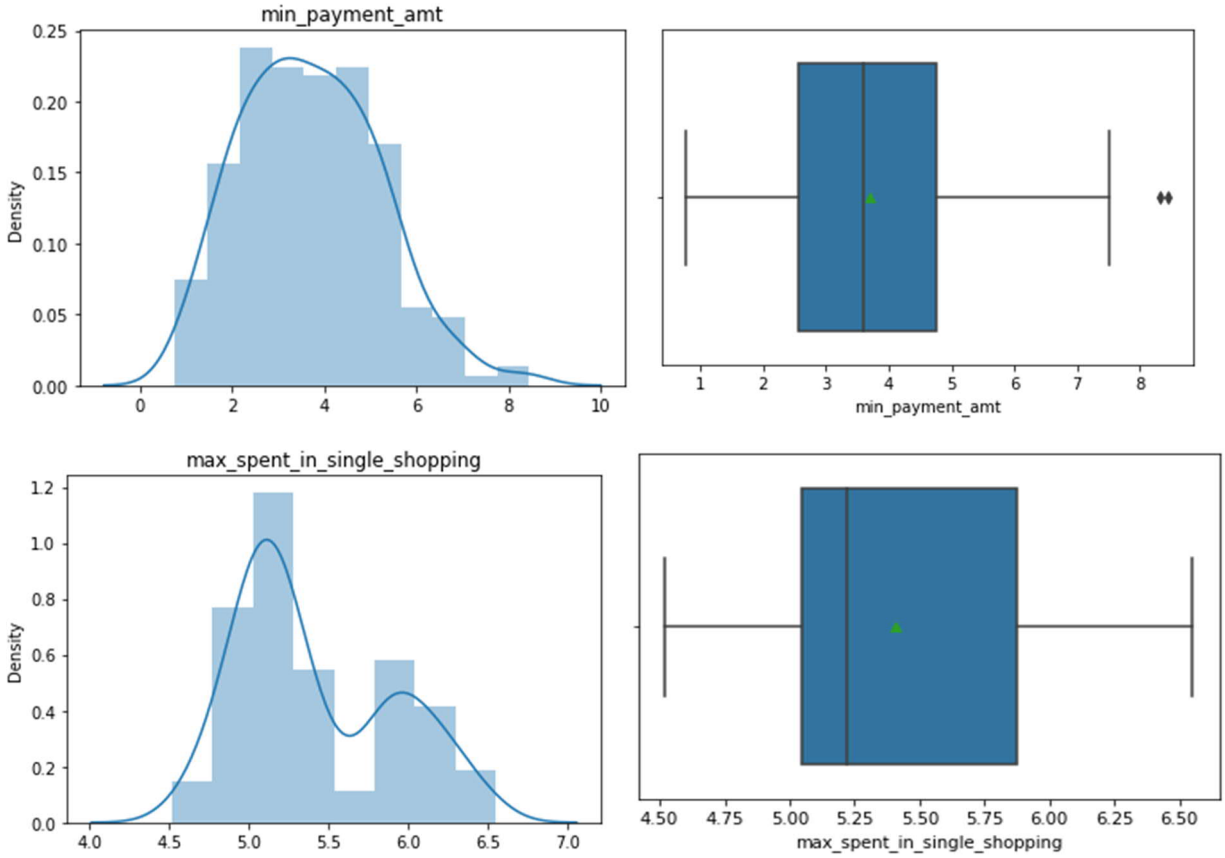


Figure 3: Univariate Curves for different features in the dataset

Now, conclusions drawn from this analysis can be briefly explained in the following points:

1. It is visible from the boxplots obtained, that the feature 'spending', 'advance_payment', 'current_balance' and 'max_spent_in_single_shopping' are right-skewed curves and 'probabitlity_of_full_payment' is a left-skewed curve. All the rest features are slightly skewed towards the right.
2. 'credit_limit' is almost the same for all the data points (with a few exceptions) in the dataset which is visible in the histogram.
3. The 'probability_of_full_payment' feature is left-skewed which means for most of the people probability of full payment is high and it is good from the bank's perspective.
4. The right skewness of most of the features such as spending, advance payment, current balance, and maximum amount spent in single shopping shows that most of the people spend or pay in a similar pattern but some exceptions are also present.

5. The min_payment_amt feature has got some outliers which can be explained as some of the people are using their credit cards for paying huge bills only and not for small amounts.

Bi-Variate Analysis-

Now for bivariate analysis, we have created a pair plot that simply shows a scatterplot for all the pairs of features present in the dataset.

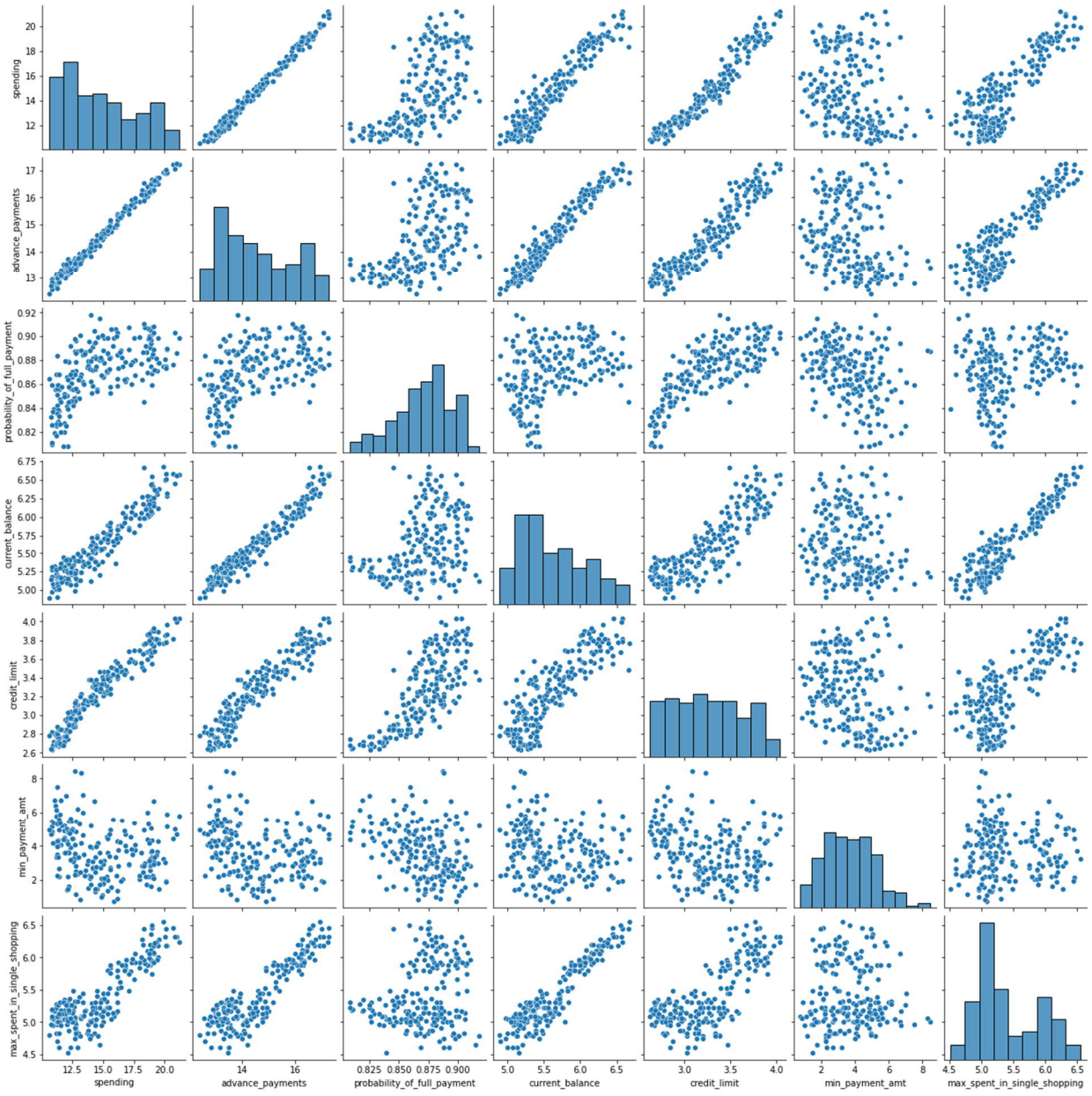


Figure 4: Pairplot for Bank Marketing Dataset

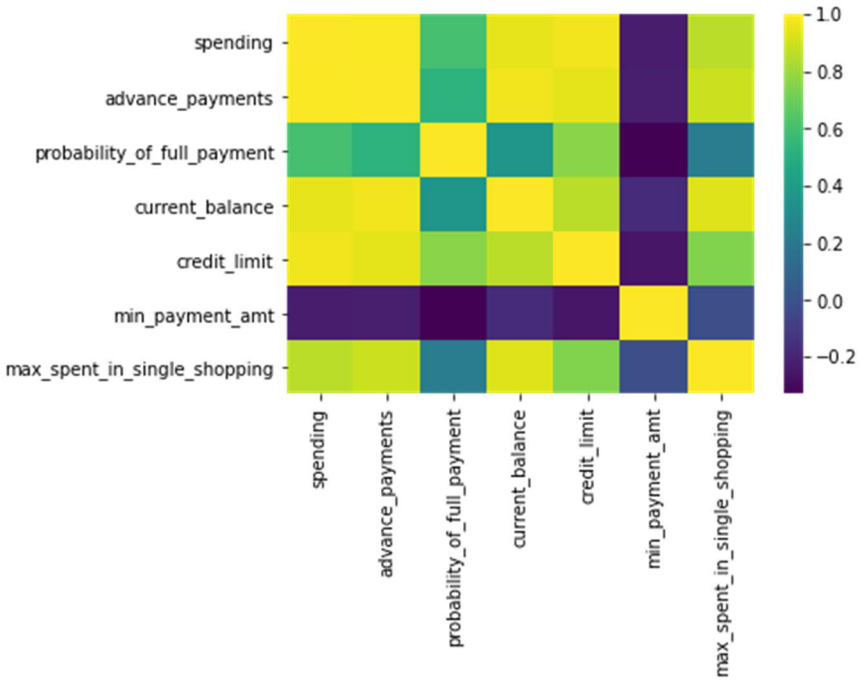


Figure 5: Correlation plot for different features

Now one can observe that the features have a high correlation between them except the 'max_spent_in_single_shopping' which doesn't have any significant correlation with any of the other features (which can also be observed from the pairplot). Also, 'probability_of_full_payment' has a somewhat lesser correlation with other features.

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

Sol:

Scaling.

Since in clustering, we are using distances in hyperspace like if we talk about Hierarchical clustering, in Ward linkage we used Euclidean distance and in K-Means also we are using distances from the centroid of the cluster for training our model. Hence, we must scale these features before training our model because this will impact the performance of our model significantly. Like if we look at our dataset the values of the probability of full payment are lying within 0 to 1 on the other hand the values for the 'spending' and

'advance_payments' feature is ranging from 10 to 21 and 12 to 18 respectively. What it means is if we train our model using these values for our distance calculations we are simply biased to these features initially.

Now for scaling we can go for Min-Max scaling or Normal scaling. Here, in our analysis, we have used Normal scaling. What it does is, it makes the mean of the distribution to zero and standard deviation equal to 1 (or at least tries to do so).

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

Sol:

Hierarchical Clustering

Now, for hierarchical clustering purposes, first, we have to choose a linkage-type as different types of linkage will result in different clustering results. In this project, we have used the Ward Linkage method and Euclidean for affinity. After this, we can create the **Dendrogram** which is nothing but the visual representation of the hierarchical clustering process in which the vertical straight lines denote the height where two items or two clusters combine. The higher the level of combining, the distant the individual items or clusters are.

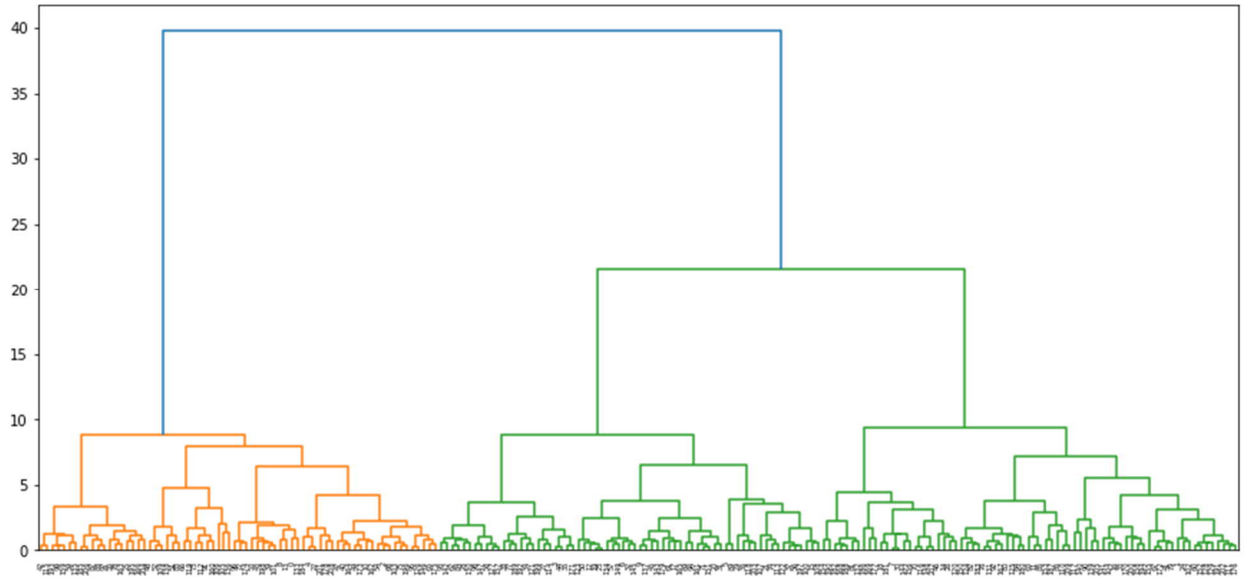


Figure 6: Ward's method of constructing clusters

As one can observe there are broadly three fairly heterogeneous clusters. These 3 clusters contain the following number of data points:

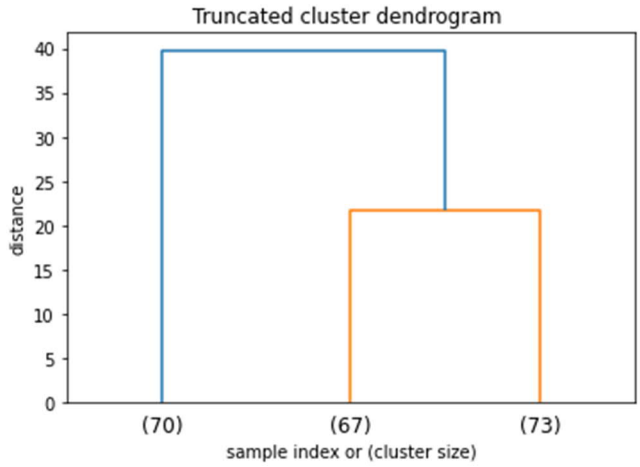


Figure 7: Truncated Dendrogram

Now for visualization of these clusters, we can again look at the pair plot but this time with the hue of cluster labels obtained from hierarchical clustering-

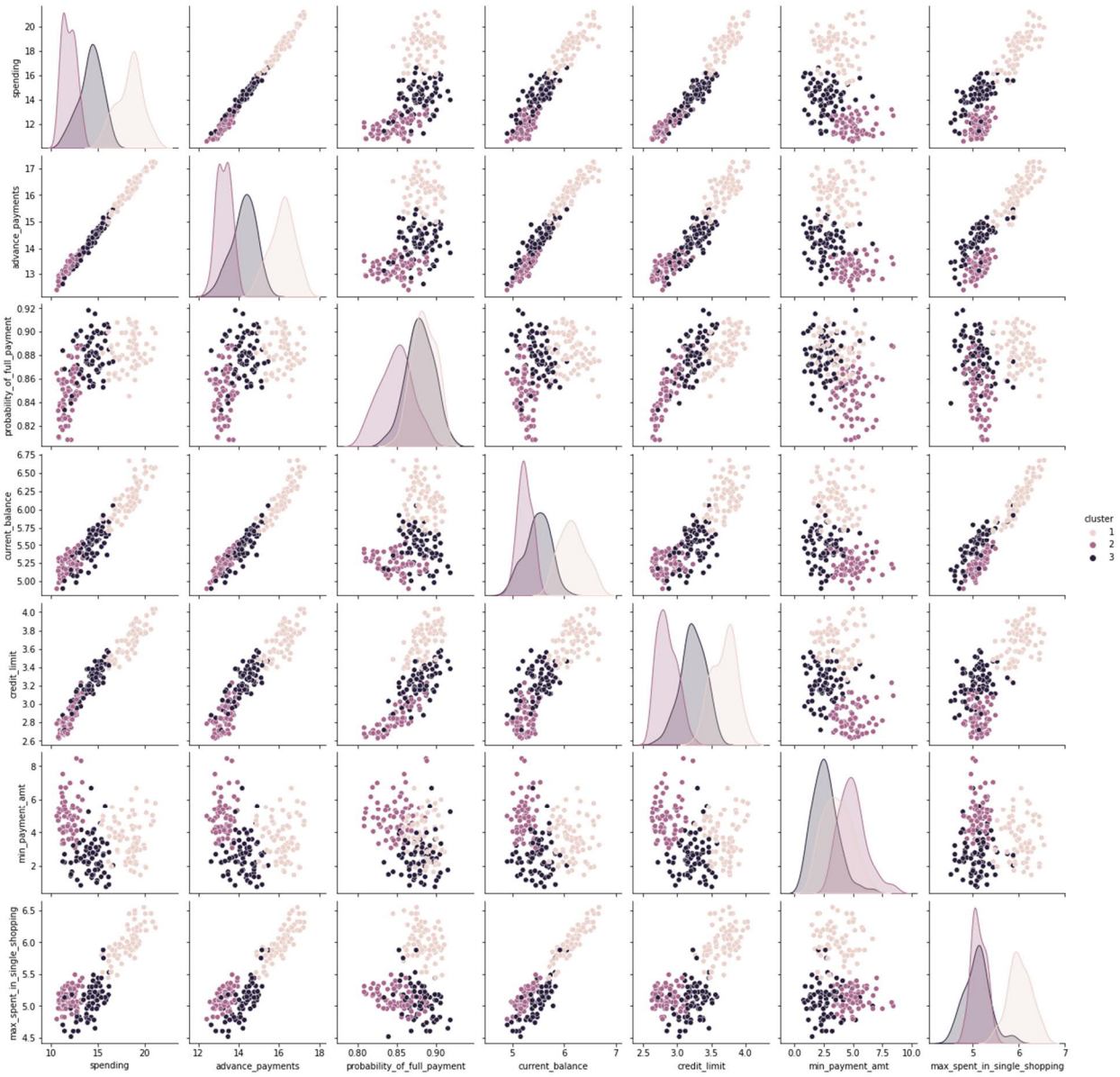


Figure 8: Pairplot with the hue of cluster labels

Now, the following observations can be drawn from this plot:

1. Cluster 1 seems to be having high values for every feature in the dataset and most heterogeneous. By looking closely, we can say that these are the customers who use **their credit cards frequently** and also pay their due amount on time.

2. Users in Cluster 2, spend a lesser amount but their minimum payment amount is high which means they **do not use the credit card very often**. Their advance payment amount is lower as well as the probability of full payment and current balance.
3. The third cluster seems to be existing in between the above two clusters which we can observe clearly in credit limit distribution in the above plot. An interesting point to observe is most of the users from this cluster tend to pay somewhat lesser amounts using the credit card (which can be observed by looking at the min payment amount and max amount in single shopping). But still, they have high probabilities for full payments. Hence, we can say **users from this cluster are using their cards moderately.**

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.

Sol:

K-Means Clustering

K-Means method is a distance-based method hence we have to perform scaling. Here we have used the normalization method. Now, for obtaining the number of optimum clusters we can look at the Elbow curve which represents the within-sum of squares value for each number of clusters.

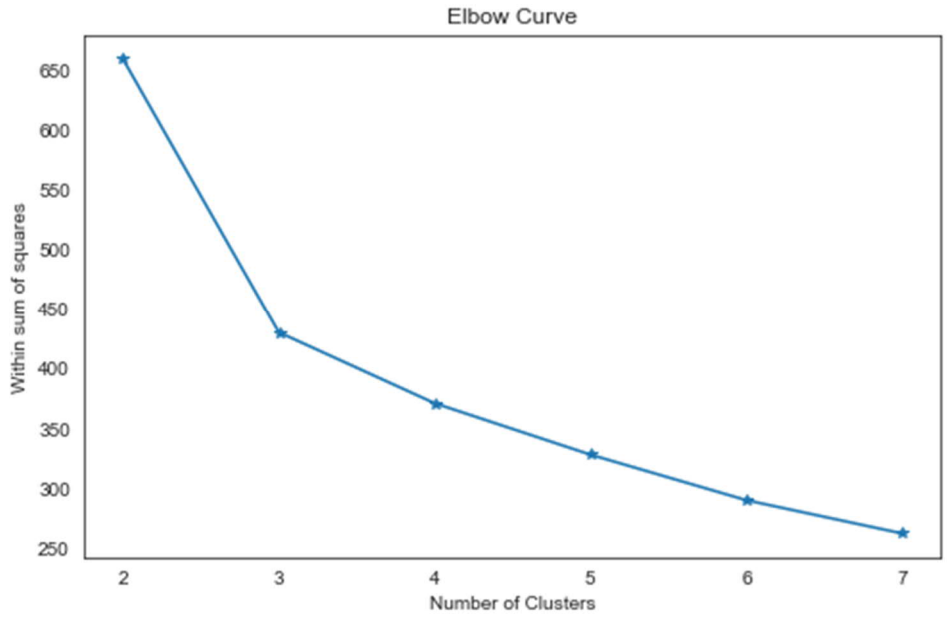


Figure 9: Elbow Curve

As the drop of within sum of squares is quite significant on increasing the number of clusters 2 to 3. We can take the 3 as the optimum number of clusters. Hence taking this value we can perform K-Means clustering.

After this, we can look at the silhouette score which in our case was 0.40. Since it is coming out to be greater than zero, we can say the clustering was fairly good. We can visualize the silhouette width distribution-

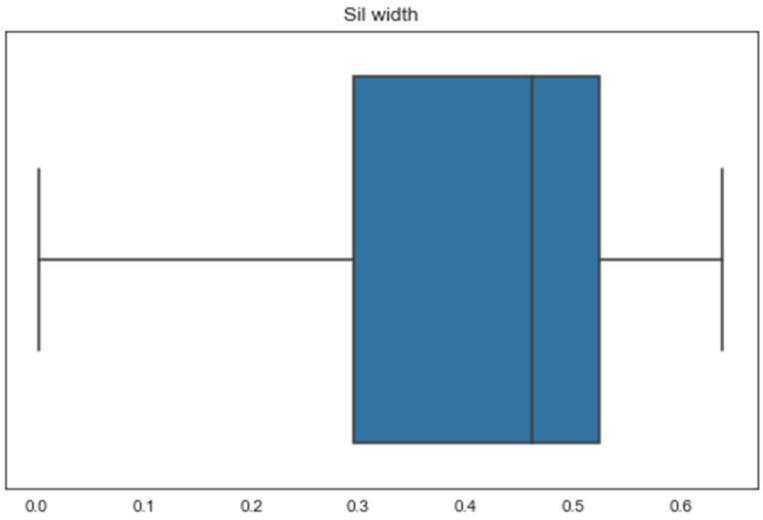


Figure 10: Silhouette Width distribution

As it is visible the entire distribution lies between 0 to 0.7. We can conclude clustering results are good.

For visualization of these clusters one can perform PCA for 2 components and can obtain the following visualization results:

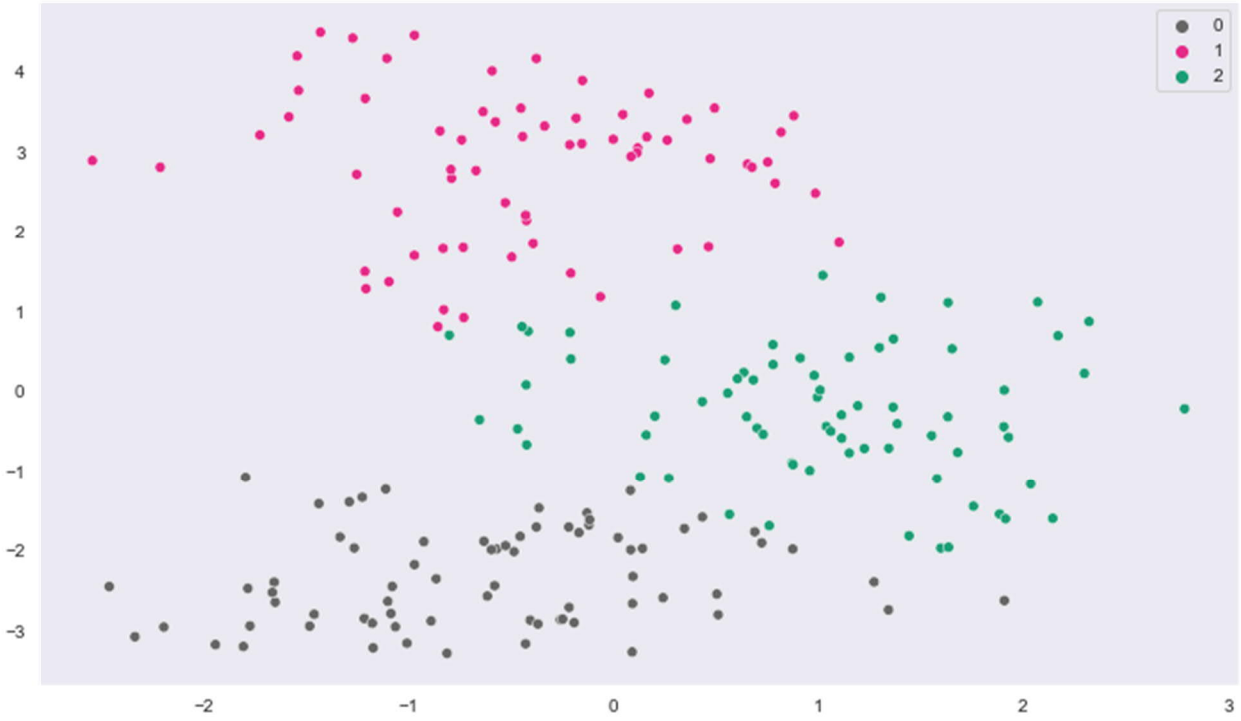


Figure 11: Visualization of clusters (using PCA)

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

Sol:

For this purpose, we can again go for pair plot with the hue of cluster label as we did in the earlier clustering model. Now, since the number of clusters is the same as we obtained in the earlier case most of the observations are quite similar here.

The observations drawn from this are mentioned below-

- Cluster 1 seems to be having high values for every feature in the dataset and most heterogeneous. By looking closely, we can say that these are the customers who use **their credit cards frequently** and also pay their due amount on time.
- Users in Cluster 0, spend a lesser amount but their minimum payment amount is high, and the maximum amount in single shopping is low which means they **do not use the credit card very often**. Their advance payment amount is lower as well as the probability of full payment and current balance.

- Cluster 2 seems to be existing in between the above two clusters which we can observe clearly in credit limit distribution in the above plot. An interesting point to observe is most of the users from this cluster tend to pay somewhat lesser amounts using the credit card (which can be observed by looking at the min payment amount and max amount in single shopping). But still, they have high probabilities for full payments. Hence, we can say **users from this cluster are using their cards moderately.**

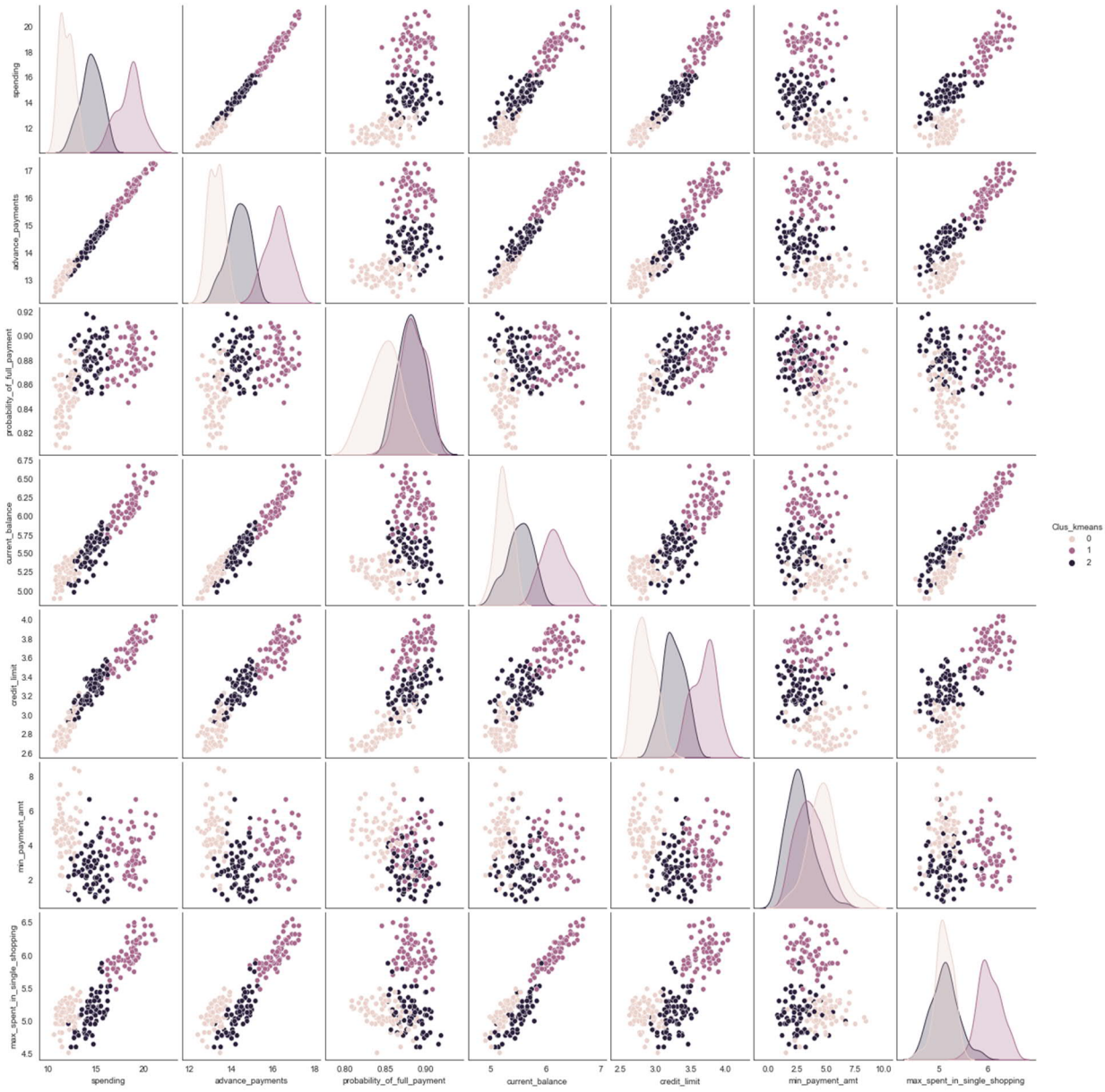


Figure 12: Cluster Profiles obtained from K-Means Clustering

Recommendations for Promotional Strategies

- For targeting the clients who fit the Cluster profile **1** (who are using their credit cards moderately and paying their due amounts regularly) following promotional strategies can be there-
    - By offering some perks like discounts/cashback on advance payments, shopping, meals, or other payments.
    - By offering better security options since these customers tend to have a comparatively higher current balance.
    - By offering better loan options on the credit card to these customers as they are more likely to pay their due amounts.

- For users fitting the cluster profiles **2** and **0**, our main target is increasing credit card usage. Now, for this purpose we can go following strategies:
    - By reaching out to users if there are any technical or transactional they are facing and resolving the same.
    - One reason for the lower usage of credit cards could be fear of fraud. This problem can also be addressed by contacting clients.

Problem 2

Information about the Dataset

An insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. The dataset contains the following attributes:

1. 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**

We are supposed to create a model using the provided dataset to predict the claimed status and provide recommendations to the management.

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

Sol:

EDA

Sample of dataset

	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 13: Sample of Insurance dataset

```
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)
```

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

Figure 14: Feature info and null count of the dataset

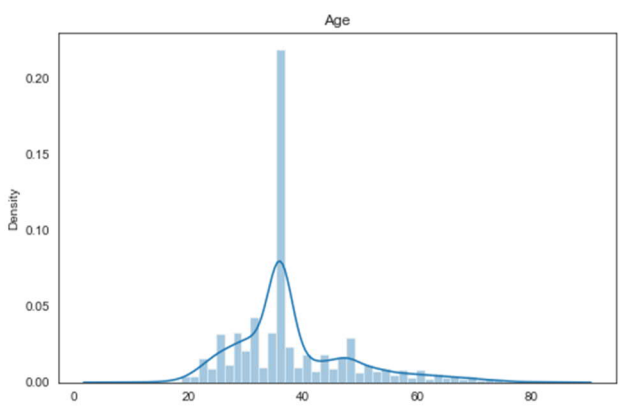
As we can observe that the dataset has a total of 3000 entries and no null values. There is a total of 6 object type variables including our target variable with is 'Claimed'. The individual value counts for these object type variables are following:

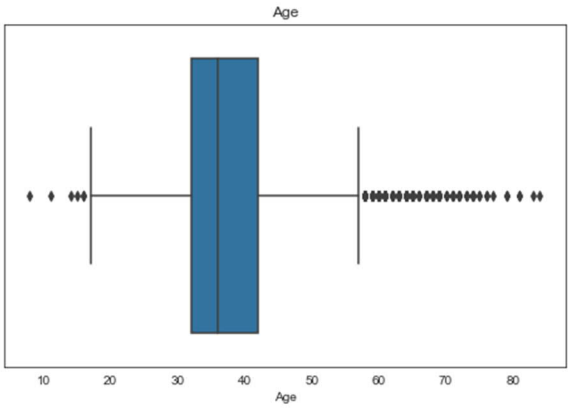
```
EPX    1365
C2B     924
CWT     472
JZI     239
Name: Agency_Code, dtype: int64

Travel Agency    1837
Airlines         1163
Name: Type, dtype: int64

No     2076
Yes     924
Name: Claimed, dtype: int64

Online     2954
Offline      46
Name: Channel, dtype: int64

Customised Plan      1136
Cancellation Plan     678
Bronze Plan           650
Silver Plan           427
Gold Plan             109
Name: Product Name, dtype: int64

ASIA        2465
Americas     320
EUROPE       215
Name: Destination, dtype: int64
```

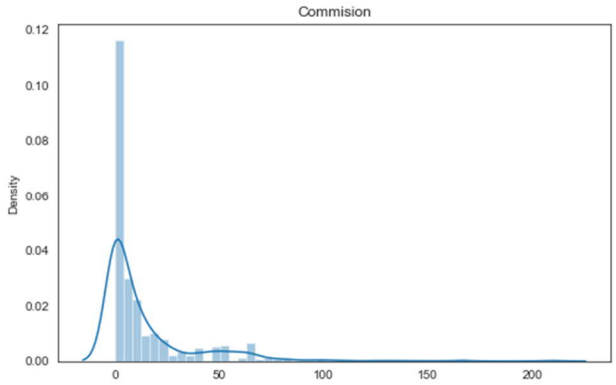
Figure 15: Value counts for each attribute in the dataset

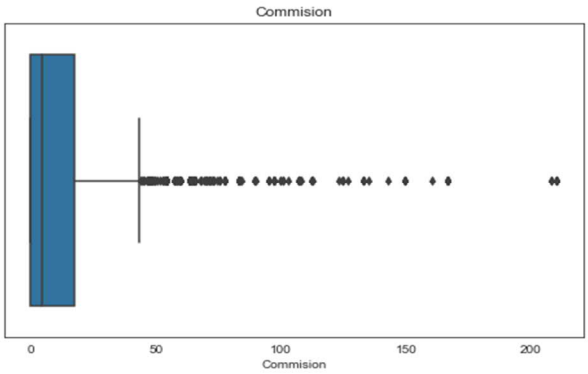
Univariate Analysis

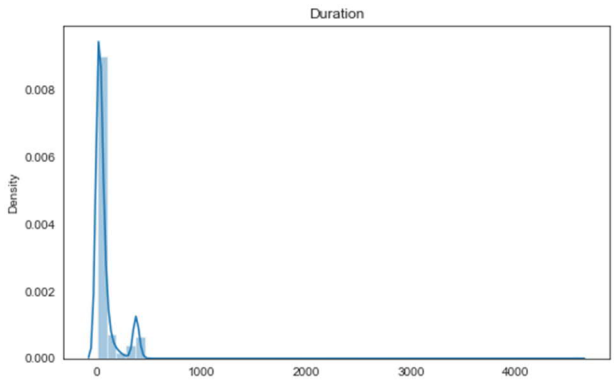
For univariate analysis we can look at distribution plot and boxplots of the attributes present in the dataset:

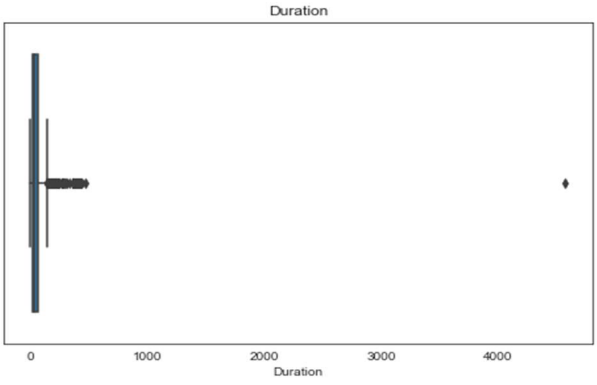












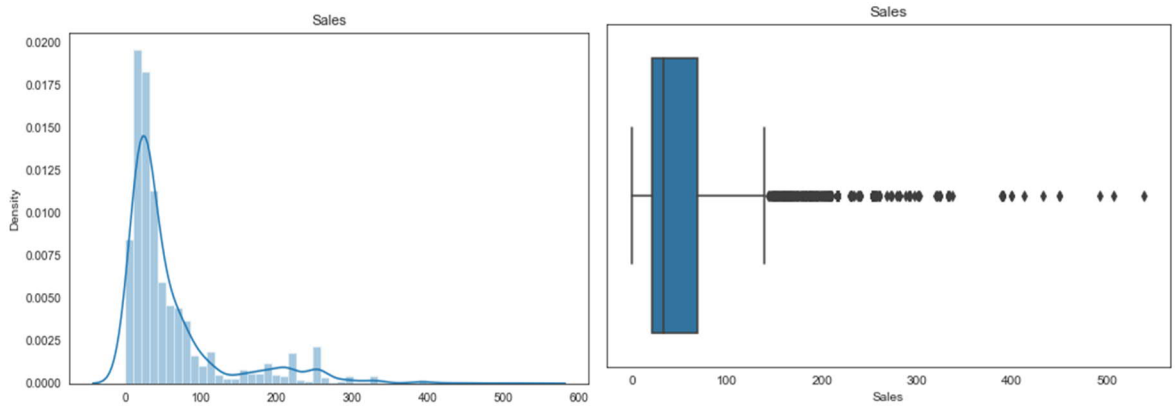
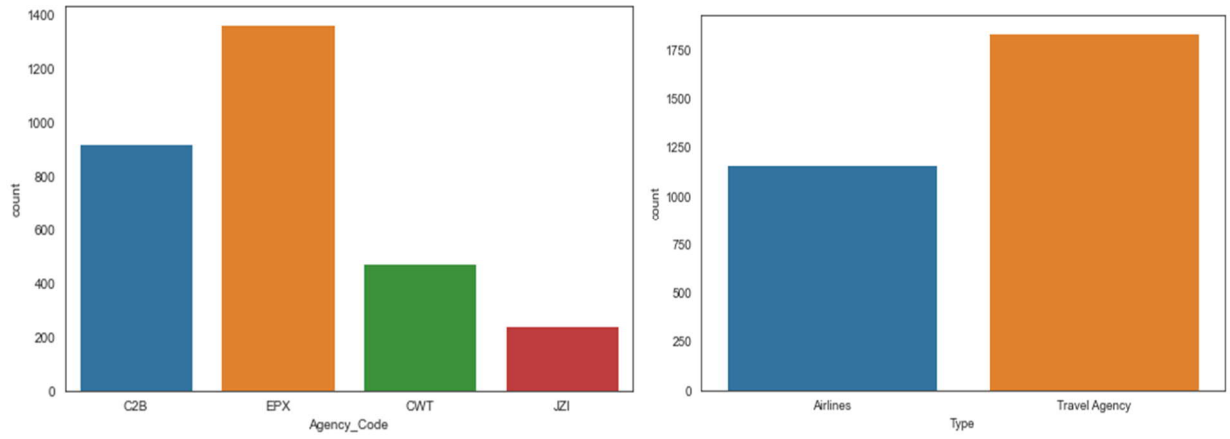


Figure 16: Univariate Analysis for continuous variables

Observations-

- All of the continuous variables are right-skewed variables and a high number of outliers are present in each of these attributes.
- In 'Duration' there is one outlier which is having an extreme value of 4580 hours but since this attribute is the duration of the tour it is one possible value.
- If we look closely at the distribution/boxplot plot of the age variable we can see that most of the people age 30 to 40.

Now, let's look at the count plot for each object type variable present in the dataset.



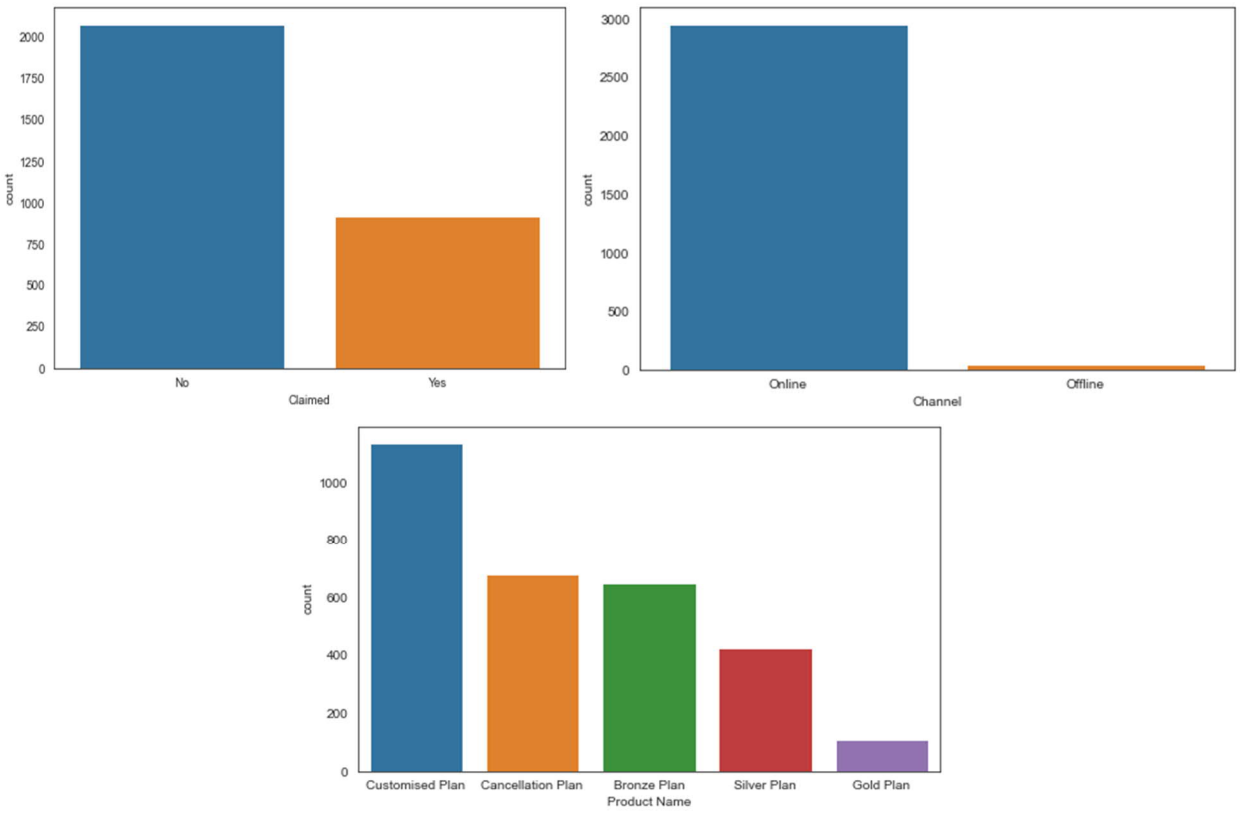


Figure 17: Count plots for object type attributes

Observations:

- **EPX** agency has a maximum number of data points.
- People are preferring the Travel Agency over the airlines.
- Customers prefer the online mode of booking over the offline mode.
- Customized plans are more popular than any of the offered plans by agencies
- Total claimed insurance is almost 1/3rd of the whole sample which is surprisingly high.

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

Sol:

Now, before moving to train test split we must change the object type attributes to integer values because, for Decision Tree and Random Forest, the scikit-learn package requires the data to be in numeric values only. After converting the data to numeric values, we can go for the train test split. In this analysis, we have chosen a 70-30 (percent of data) split for train and test data. Also, both Train and Test set contains almost 70 percent of class 'No' and 30 percent of class 'Yes' of the target variable ('Claimed').

For the modeling part please refer to the Notebook file attached with this document.

Decision Tree Model

The Decision tree model we have made is using the Gini Impurity as split criteria. For optimizing the performance of the model, we have used the Grid search method which simply runs several models in one go and returns the best model among those. After running the grid search we can use this object to predict values in test data. The hyperparameters which we obtained after running the grid search are following:

1. 'max_depth': 5
2. 'max_features': 8
3. 'min_samples_leaf': 20
4. 'min_samples_split': 200

The minimum number of samples required to split an internal node is represented as min_samples_split. The minimum number of samples required to be at a leaf node is represented as min_samples_leaf.

The model we obtain using these hyperparameters gets an accuracy of 79 percent on the training set and 76 percent on the test set.

Random Forest Model

In Random Forest Model, after running the grid search for hyperparameters we obtained the following values as best parameters:

1. 'max_depth': 5,
2. 'max_features': 4,
3. 'min_samples_leaf': 300,
4. 'min_samples_split': 10,
5. 'n_estimators': 200

The n_estimators parameter represents the number of trees in Random Forest. Rest all other parameters have the same meaning as it was mentioned in the Decision Tree model. The accuracy score for this model is 77 percent for the train set and 76 percent for the test set.

Neural Network Model

Now for training our Neural Network Model, we have scaled our data first. In earlier models, it wasn't necessary because Random Forest and Decision Trees broadly are nothing but a long set of if-else

statements. The tolerance limit for training the model was chosen to be 0.0001. Other hyperparameters obtained from the grid search method are-

1. 'hidden_layer_sizes': 50
2. 'max_iter': 5000
3. 'solver': 'adam'

Here, the 'max_iter' parameter limits the number of iterations for which the model will try to converge to its tolerance limit. The accuracy score for the train set was 79 and 77 percent for the test set using this model.

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.

Sol:

Performance Metrics

For Decision Tree Model-

When a full-grown decision tree is trained it always performs best on the training set but it struggles to perform on the test set. It was the same case in our analysis too. The full-grown Decision tree has an accuracy score of 100 percent on the train set and 71 percent on the test set.

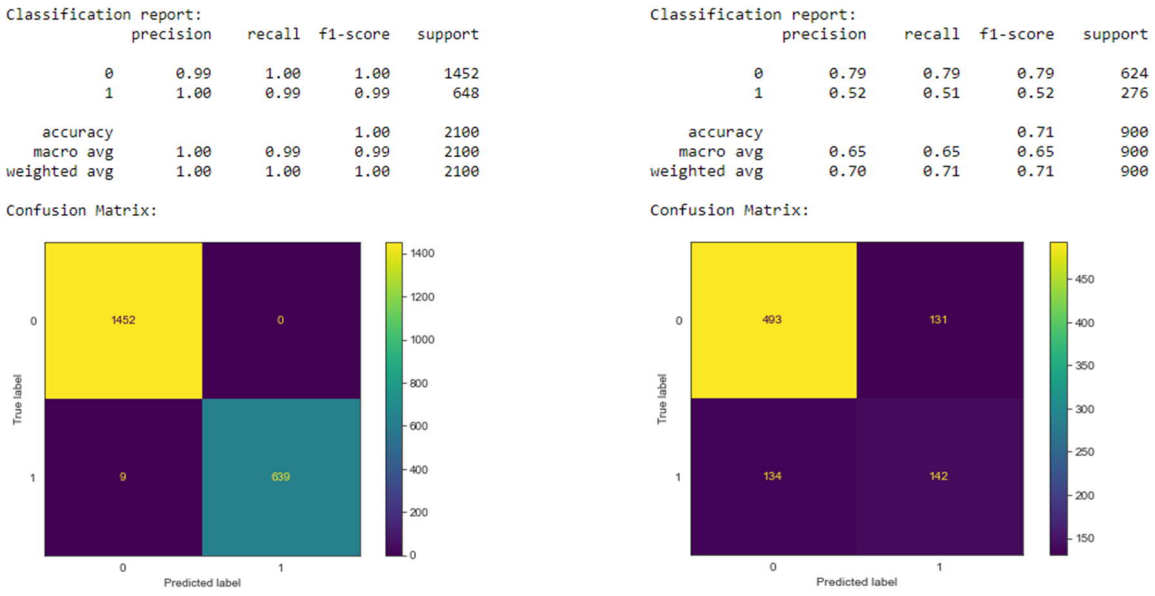


Figure 18: Performance metrics for Decision Tree Model (full-grown) on the train (left) and test set (right)

After tuning this model, the model was able to perform slightly better on the test set. Performance metrics are depicted in the following images:

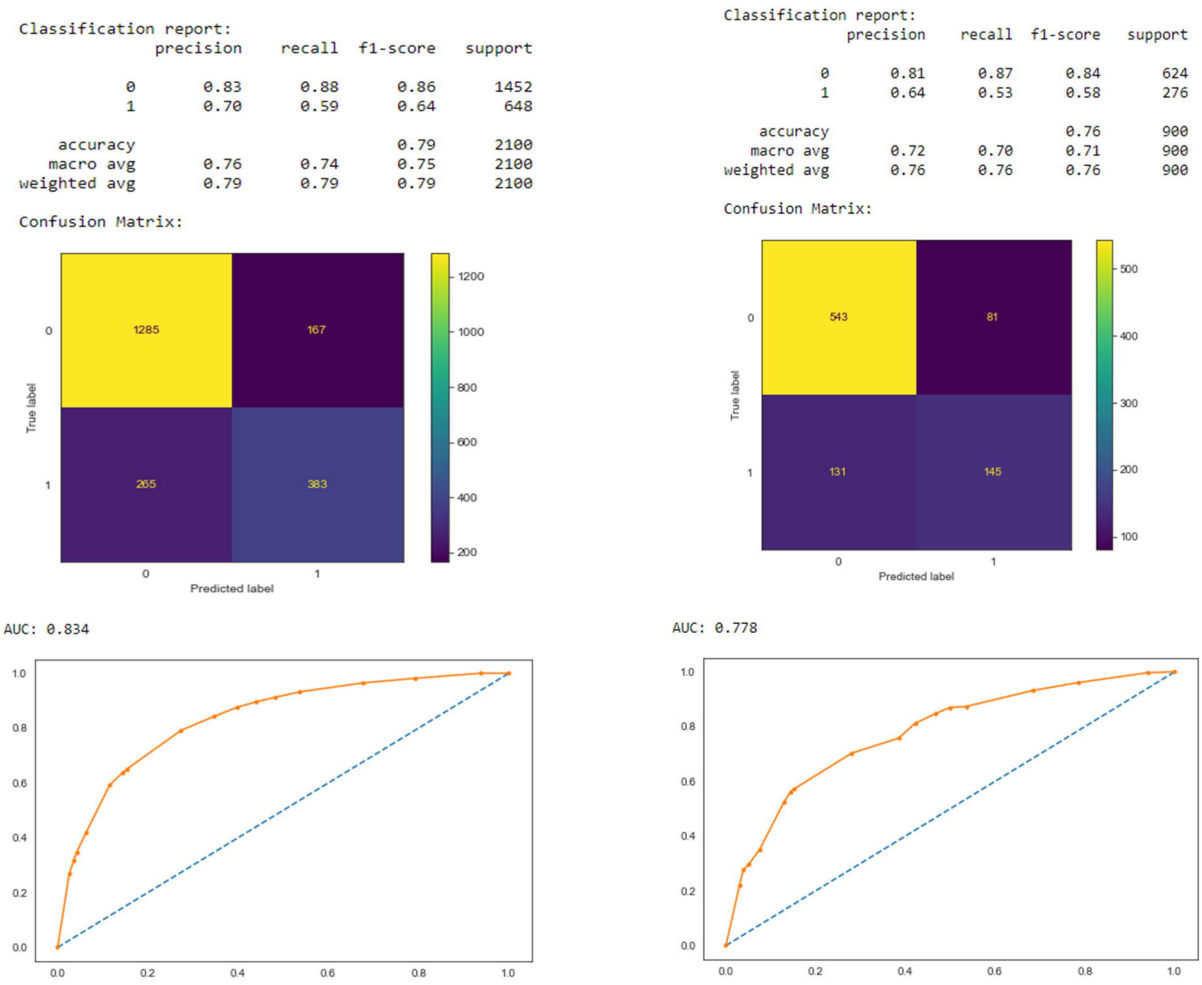


Figure 19: Performance metrics for Decision Tree Model (tuned) on the train (left) and test set (right)

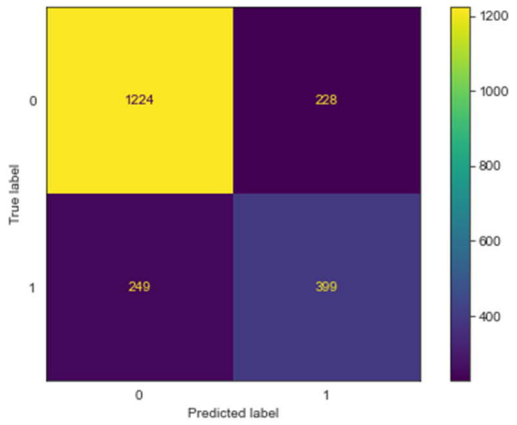
<u>For Random Forest Model-</u>

Although Random Forest performs significantly better in most cases but in the present study it wasn't able to perform as expected. The performance metrics achieved using this model are given as:

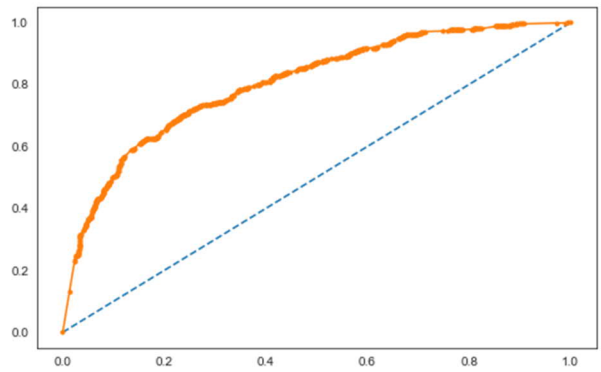
Classification Report:

	precision	recall	f1-score	support
0	0.83	0.84	0.84	1452
1	0.64	0.62	0.63	648
accuracy			0.77	2100
macro avg	0.73	0.73	0.73	2100
weighted avg	0.77	0.77	0.77	2100

Confusion Matrix:



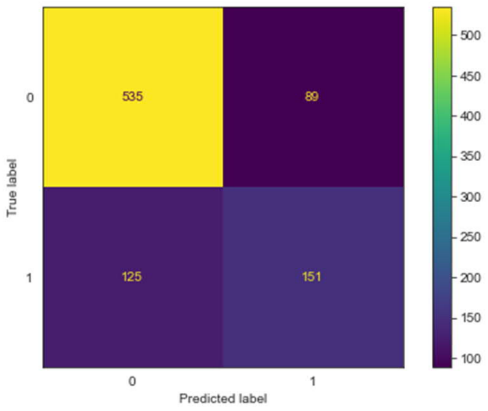
AUC: 0.803



Classification Report:

	precision	recall	f1-score	support
0	0.81	0.86	0.83	624
1	0.63	0.55	0.59	276
accuracy			0.76	900
macro avg	0.72	0.70	0.71	900
weighted avg	0.75	0.76	0.76	900

Confusion Matrix:



AUC: 0.780

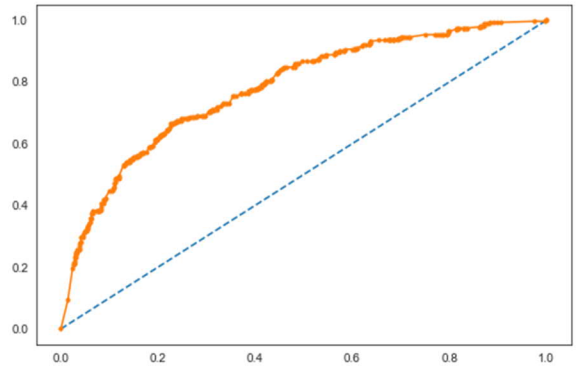


Figure 20: Performance metrics for Random Forest Model (tuned) on the train (left) and test set (right)

For Neural Network Model-

This model was able to perform slightly better than both of the other models. Performance metrics are given as:

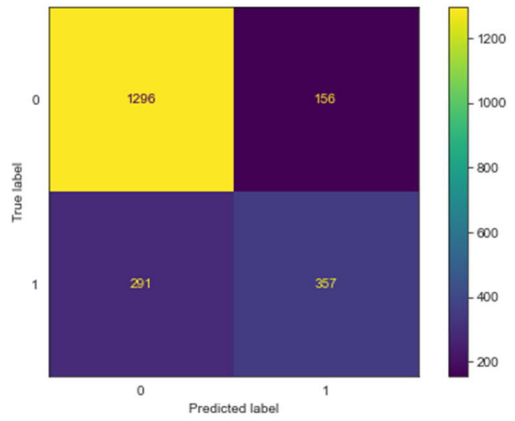
Classification Report:

	precision	recall	f1-score	support
0	0.82	0.89	0.85	1452
1	0.70	0.55	0.61	648
accuracy			0.79	2100
macro avg	0.76	0.72	0.73	2100
weighted avg	0.78	0.79	0.78	2100

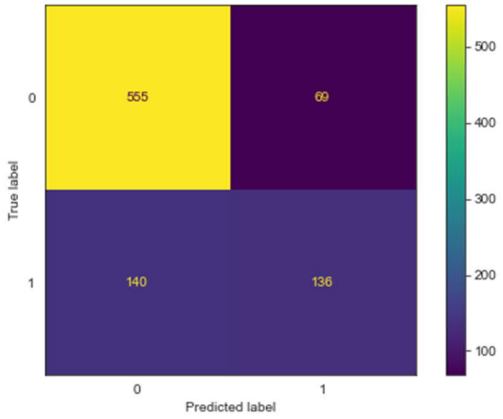
Classification Report:

	precision	recall	f1-score	support
0	0.80	0.89	0.84	624
1	0.66	0.49	0.57	276
accuracy			0.77	900
macro avg	0.73	0.69	0.70	900
weighted avg	0.76	0.77	0.76	900

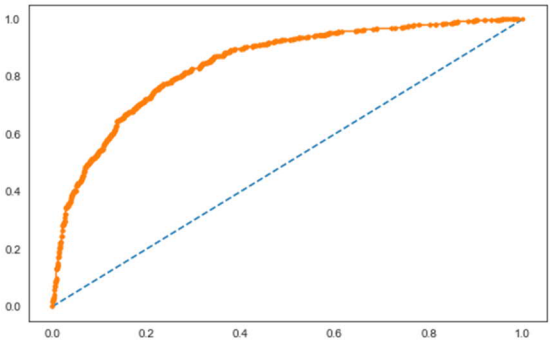
Confusion Matrix:



Confusion Matrix:



AUC: 0.844



AUC: 0.803

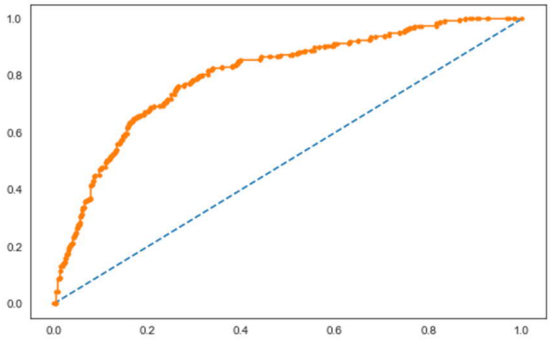


Figure 21: Performance metrics for Neural Network model(tuned) on the train (left) and test set (right)

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

Sol:

Now, if we compare all the above models, the Neural Network Model performs slightly better than the rest of the two (based on accuracy). In this analysis, the False Negative plays a more important role than False

positives because it is more important to predict ‘Claimed’ insurance than predicting ‘Not Claimed’ from the insurance firm’s perspective. Hence, Recall plays a major role in performance metrics than precision. Though, in this analysis, the recall values for class 1 were not good enough. Considering the above three models one should go for the Random Forest model because it was showing somewhat better results for recall values.

The reasons for low recall values may be the class imbalance, fewer data points available, or lack of optimization. For class imbalance, one can go for oversampling. Tuning of models can be improved by using the randomized search or running grid search for bigger ranges. In this analysis due to computational limits, the grid search was not run for huge ranges of values for tuning the hyperparameters.

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

Inferences and Recommendations

Based on this analysis following conclusions can be drawn:

- The Random Forest model performed best considering the business problem.
- Considering Accuracy and ROC-AUC metrics Neural Network model performed better than the other two
- Further tuning of hyperparameters for Random Forest is required though it is computationally expensive and time-consuming
- The data contains a large number of outliers which could have impacted the performance of the Neural Network.