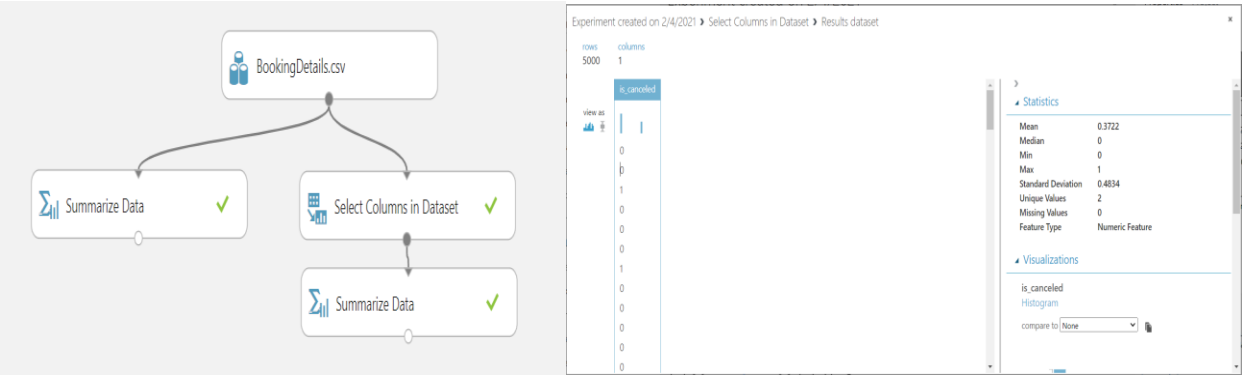
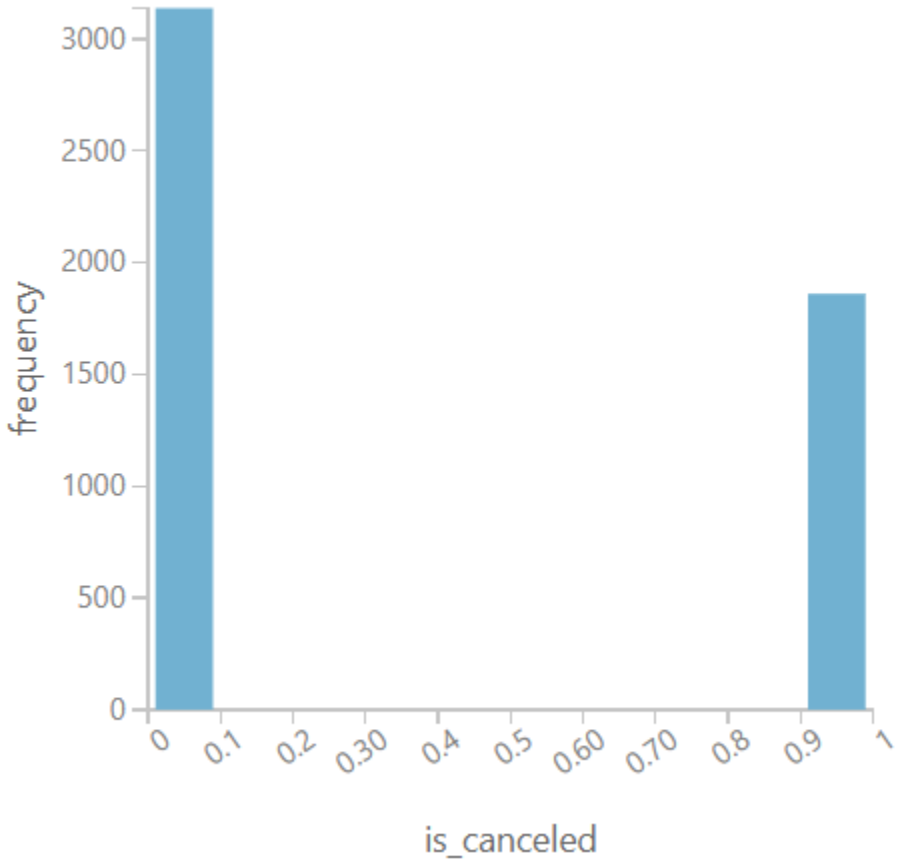
PROJECT – 3

Q1. Read the dataset and visualize the target(i.e. is_cancel). State whether it is imbalanced or not. How we can deal with class imbalance, state briefly.(3 point)

Ans.



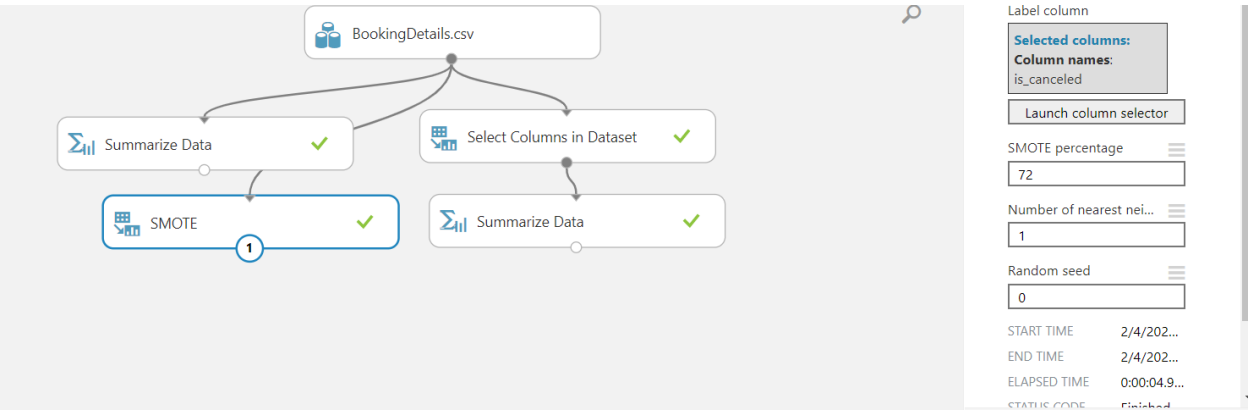


Yes, the given data is imbalanced and the percentages of data is 63% and 37% stating is_cancelled or not which can result in bias of the model.

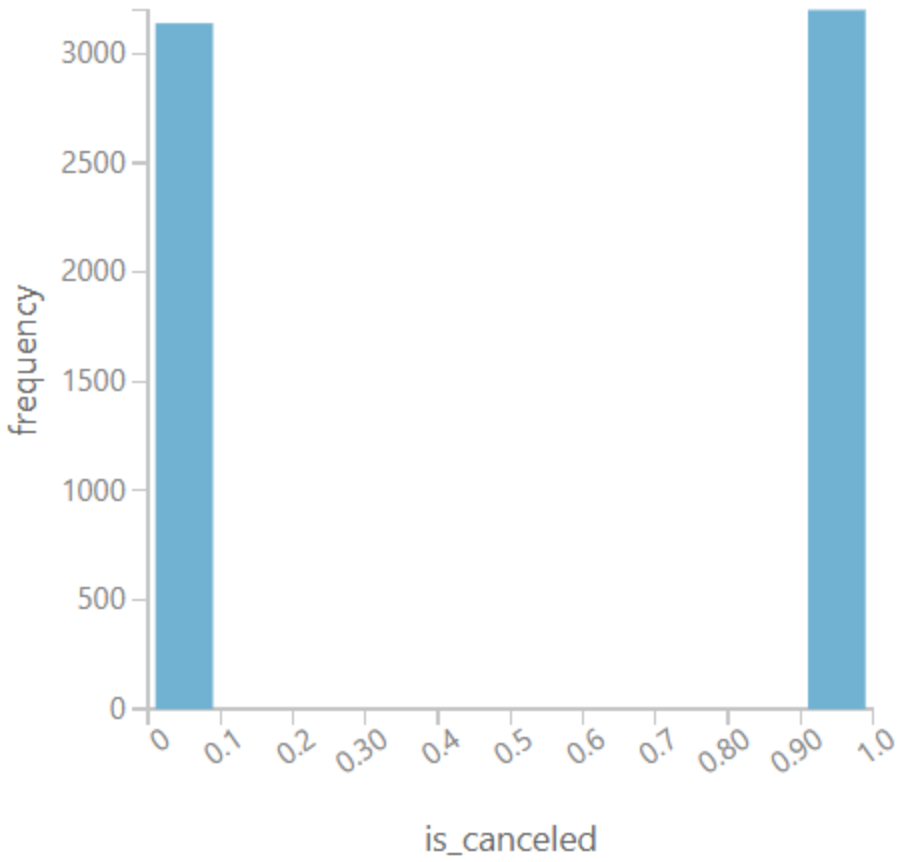
We have to balance the data.

The main objective of balancing classes is to either increasing the frequency of the minority class or decreasing the frequency of the majority class. This is done in order to obtain approximately the same number of instances for both the classes. Few techniques are:

1. Change the performance metric
2. Change the algorithm
3. Oversample minority class
4. Undersample majority class
5. Generate synthetic samples



I used SMOTE to correct the imbalance. Percentage used is 72% to get get balanced data.

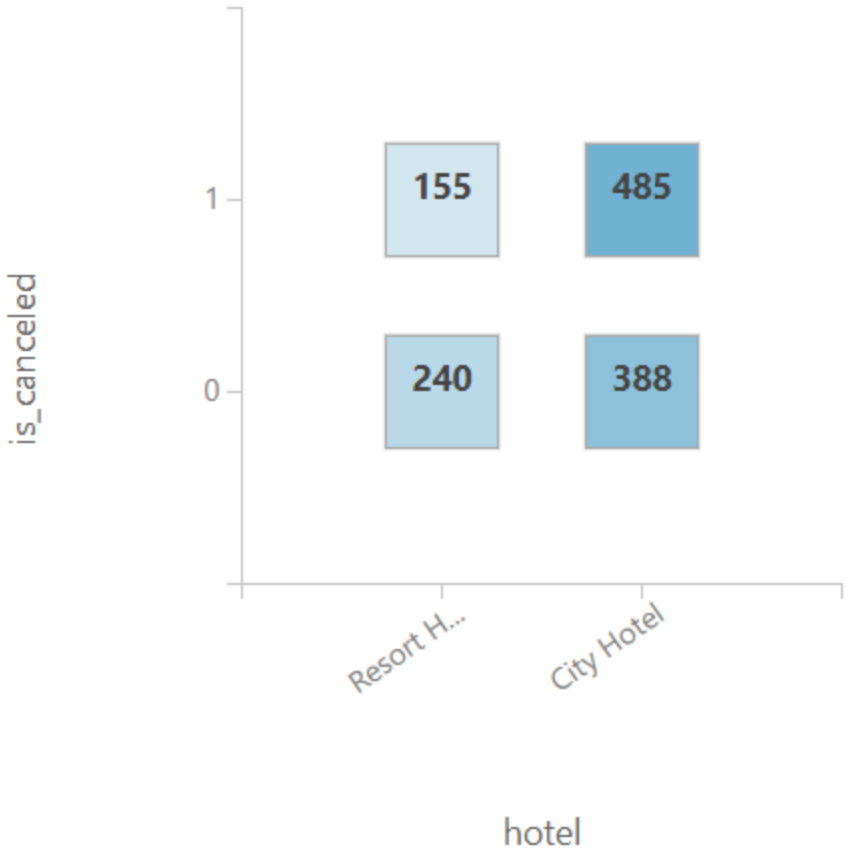


SMOTE is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.

Split the dataset into 80:20 ratio.(80% training and 20% testing).Visualize the test data. NOTE: ALL QUESTIONS FROM Q2. to Q6. will be answered on the test data.

Q2. In the test data in which type of hotel the cancelation is more?(2 points)

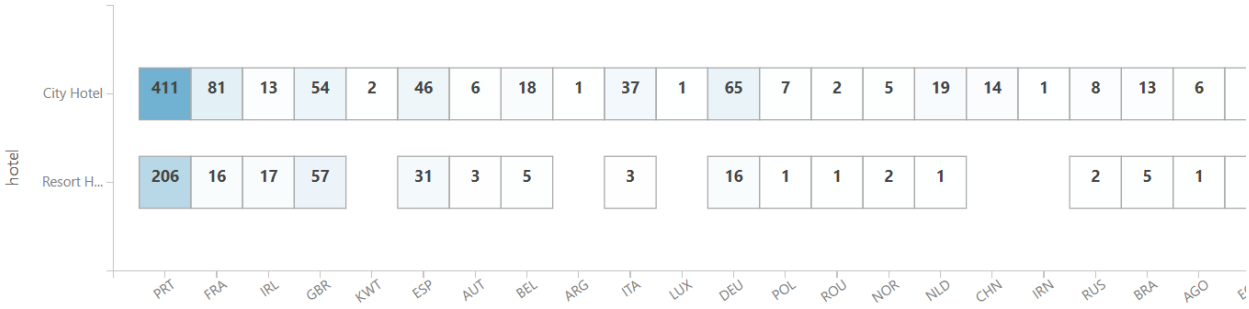
Ans.

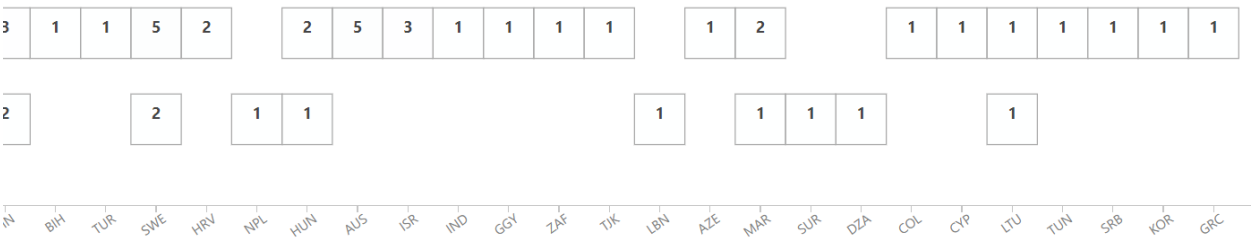


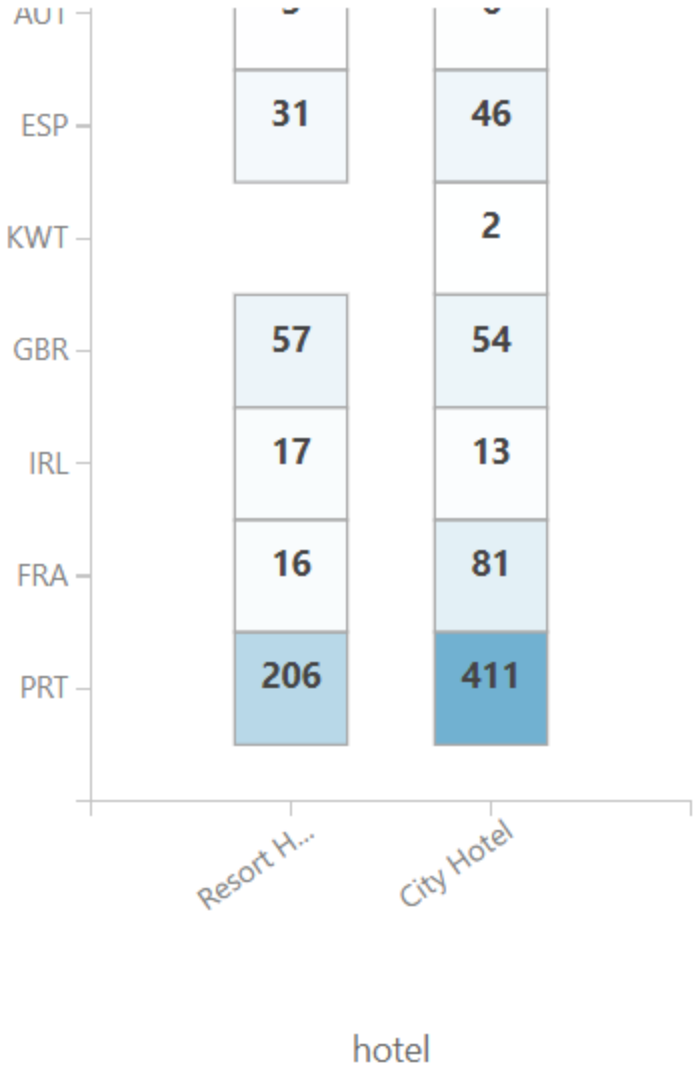
We can see from the plot that city Hotel have higher cancellation cases than Resort Hotel.

Q3. Which Country has more resort hotels and which country has more city hotels?(3 points)

Ans.





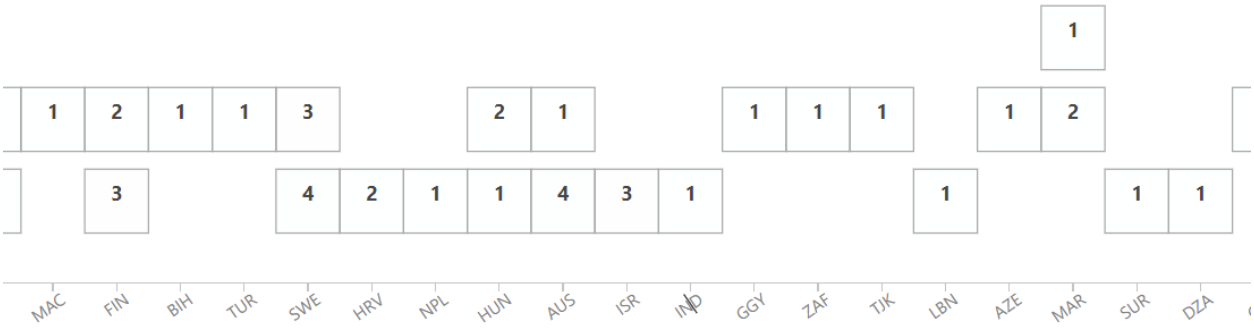


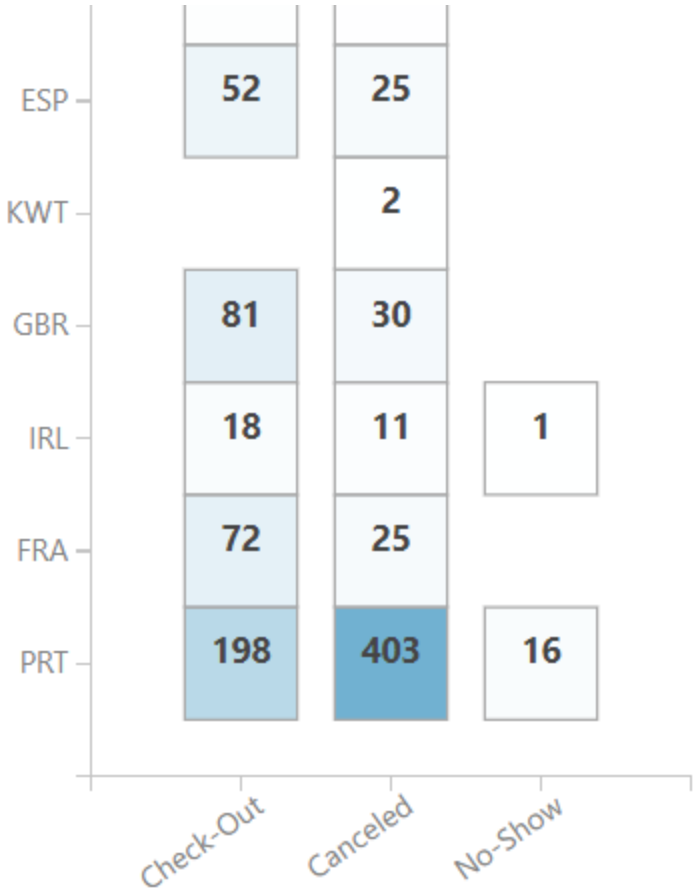
PRT has maximum number of city hotels and resort hotel being 411 and 206 respectively. PRT is the three-letter country abbreviation for Portugal. So Portugal has maximum hotels.

This a Crosstab.

Q4. How many check-outs has been done in hotels in India, Here IDN, refers to India?(2 Points)

Ans.



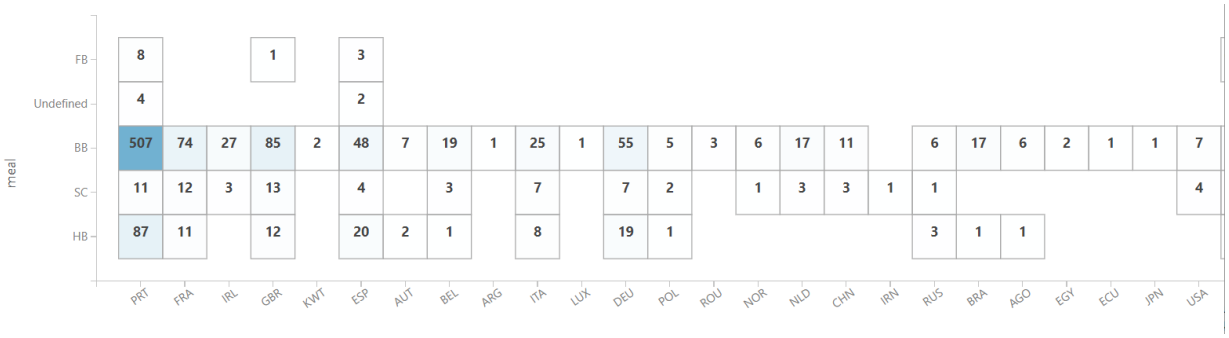


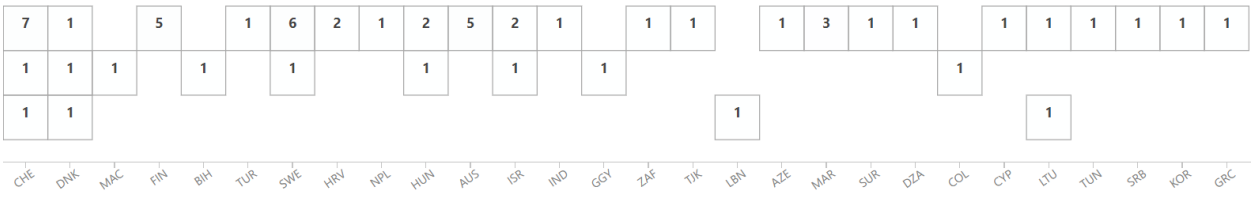
There is only 1 checkout done from hotel from India (IDN).

This a Crosstab.

Q5. In which countries the minimum number of BB & SC meals have been booked?(3points)

Ans.





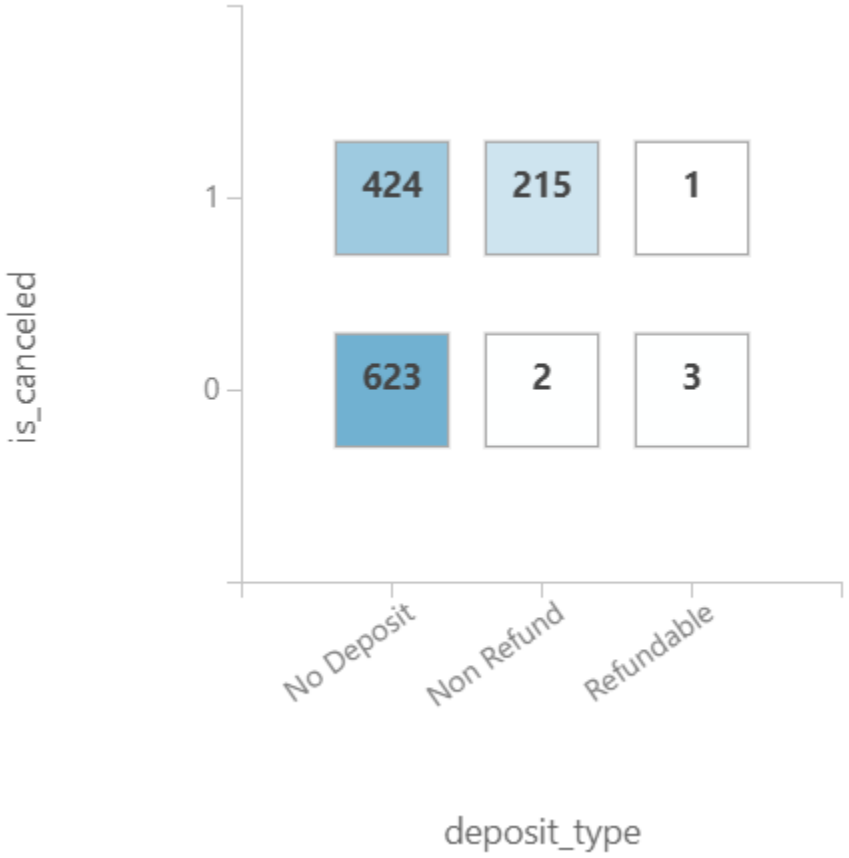
We can see from the plot that LBN has minimum i.e. no SC or BB meals.

Other countries having minimum SC and BB meal are GRC, KOR, SRB, TUN, LTU, CYP, COL, DZA, SUR, AZE, LBN, etc...

This a Crosstab.

Q6. State the inferences between deposit_type and is_cancel.(2 points)

Ans.



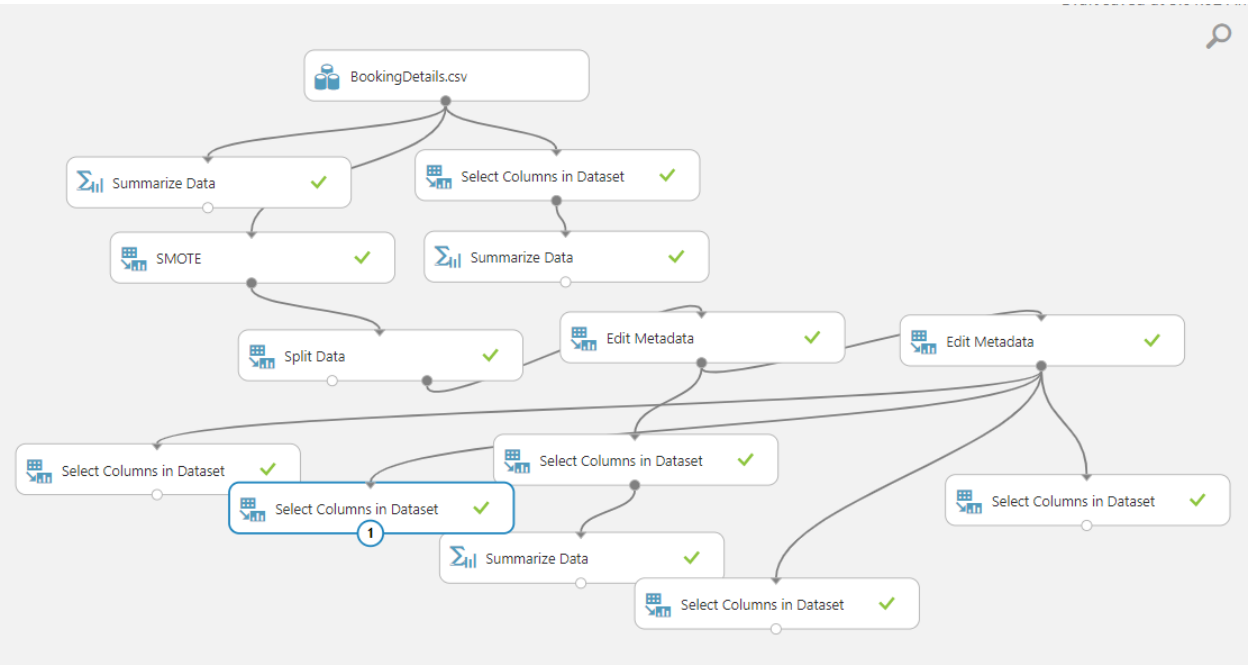
We can see from the plot that chances of reservation being cancelled is highest when there is non-refundable deposit made and is lowest when its refundable deposit type. Here we are not considering the number but we are taking the ratio of most to list the ratio tells the exact results.

Following are the percentages of hotel being cancelled wrt to deposit type:

40.49 % - no deposit

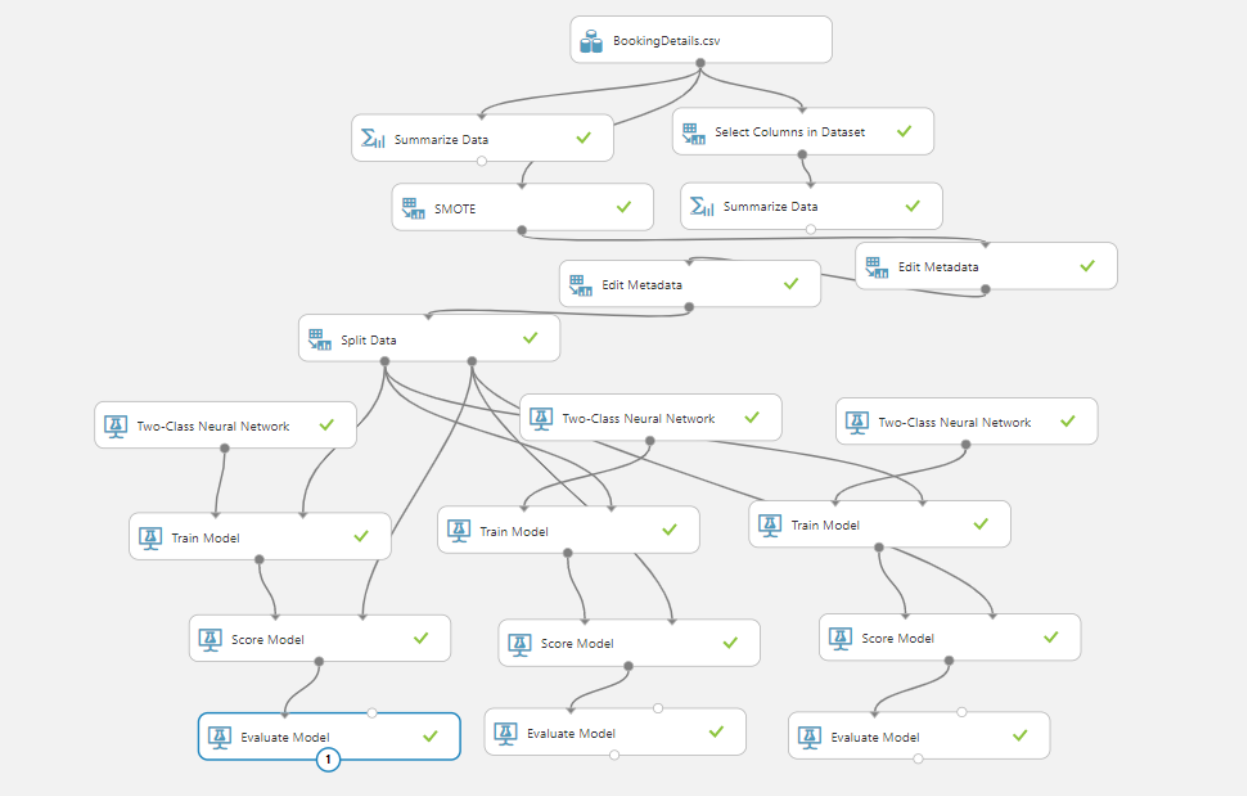
99.07% - non refundable

25% - refundable



Q7. Please follow below steps

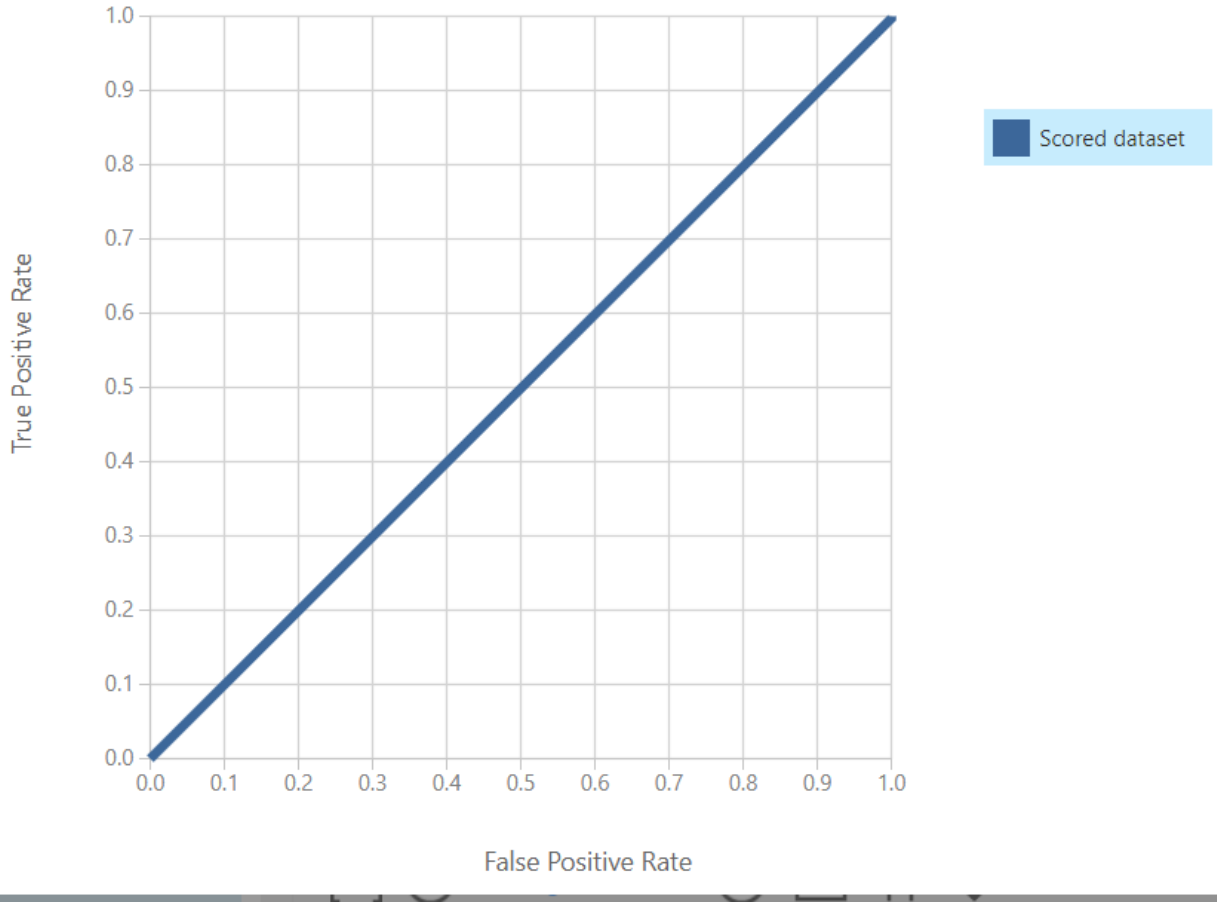
7.1 Experiment with different neural networks with parameters combinations given in the below table and run each model. (4 points)



No. of Hidden Nodes	Learning Rate	Momentum	Normalizer
128	0.0001	0.3	Do not Normalize
64	0.001	0.2	Mini-Max Normalizer
32	0.03	0.5	Gaussian Normalizer

These are the 3 models I am using trained on same training data and tested on same test data and following are the inferences and result.

1. FIRST Model

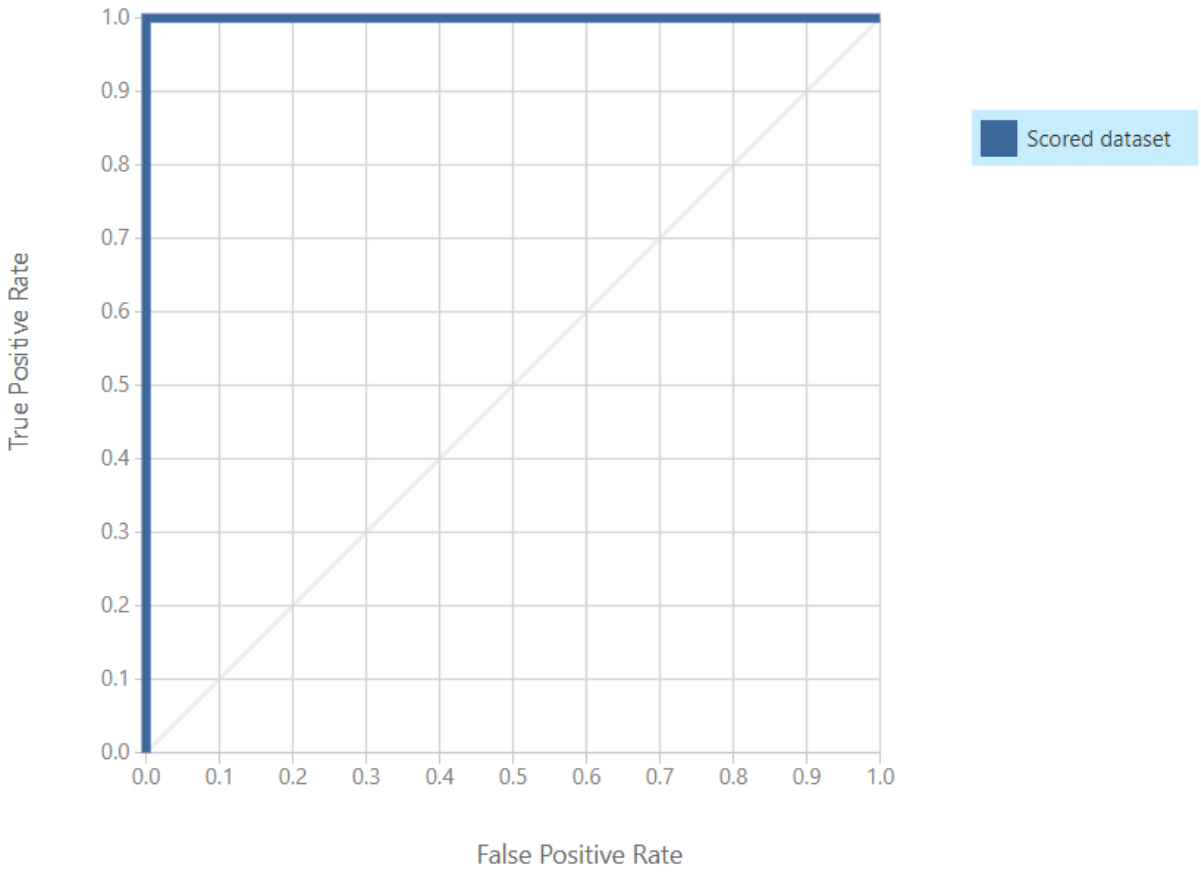


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
0	640	0.495	1.000	0.5	0.540

False Positive	True Negative	Recall	F1 Score
0	628	0.000	0.000

Positive Label	Negative Label
1	0

2. SECOND Model

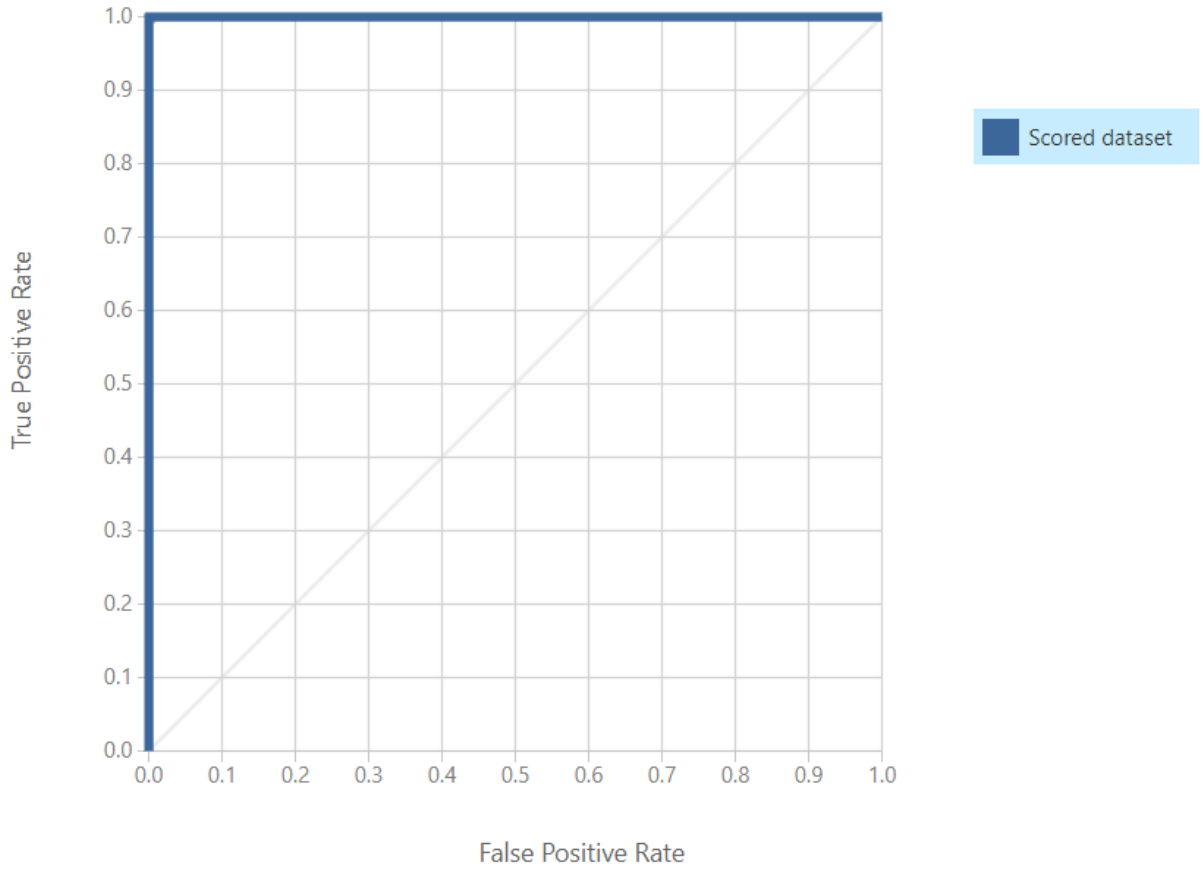


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
640	0	1.000	1.000	0.5	1.000

False Positive	True Negative	Recall	F1 Score
0	628	1.000	1.000

Positive Label	Negative Label
1	0

3. THIRD Model



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
640	0	0.998	0.997	0.5	1.000

False Positive	True Negative	Recall	F1 Score
2	626	1.000	0.998

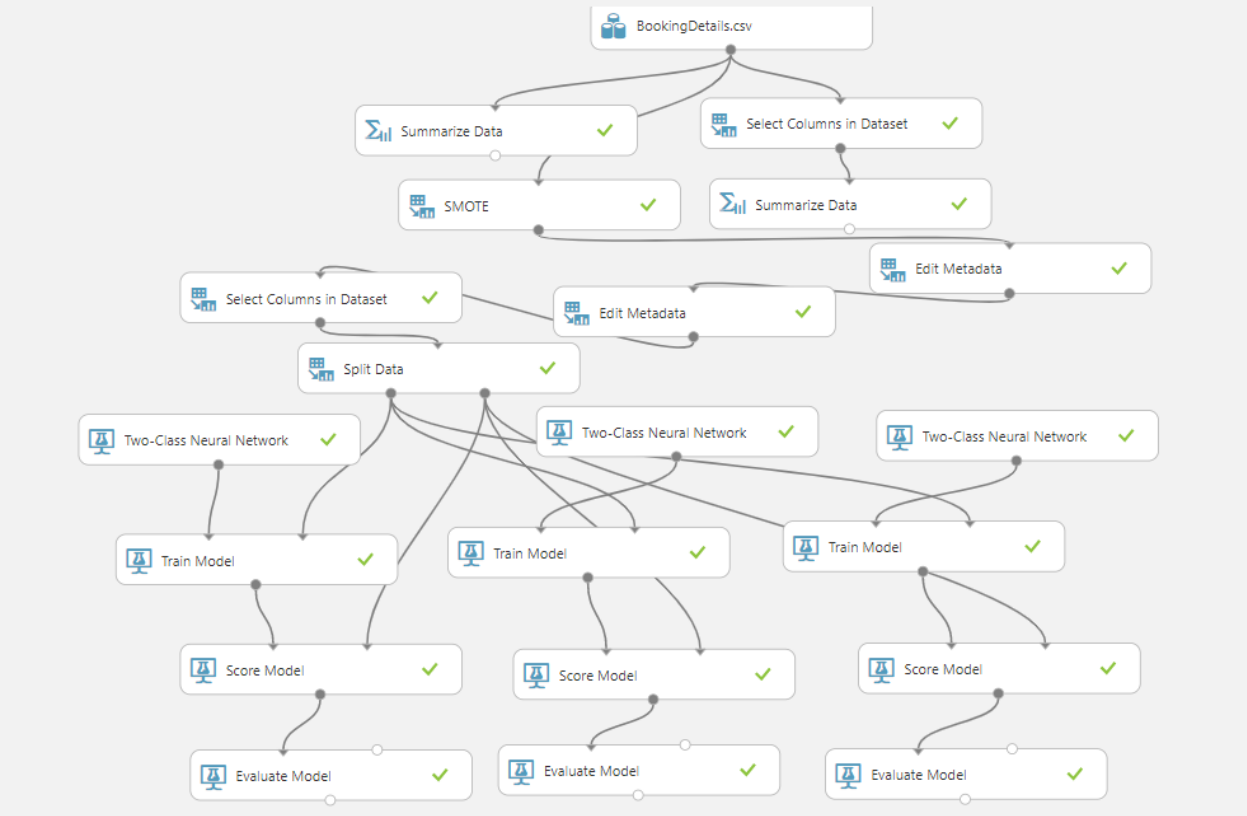
Positive Label	Negative Label
1	0

7.2 State the difference between results obtained in train and test dataset clearly with reasons. (2 points)

Ans.

We can see increasing the hidden layer more is causing over fitting while taking the number of hidden layers too less there is underfitting and normalizing the data is really important.

7.3 Please drop the “reservation status” column and follow the below steps. (1 point)

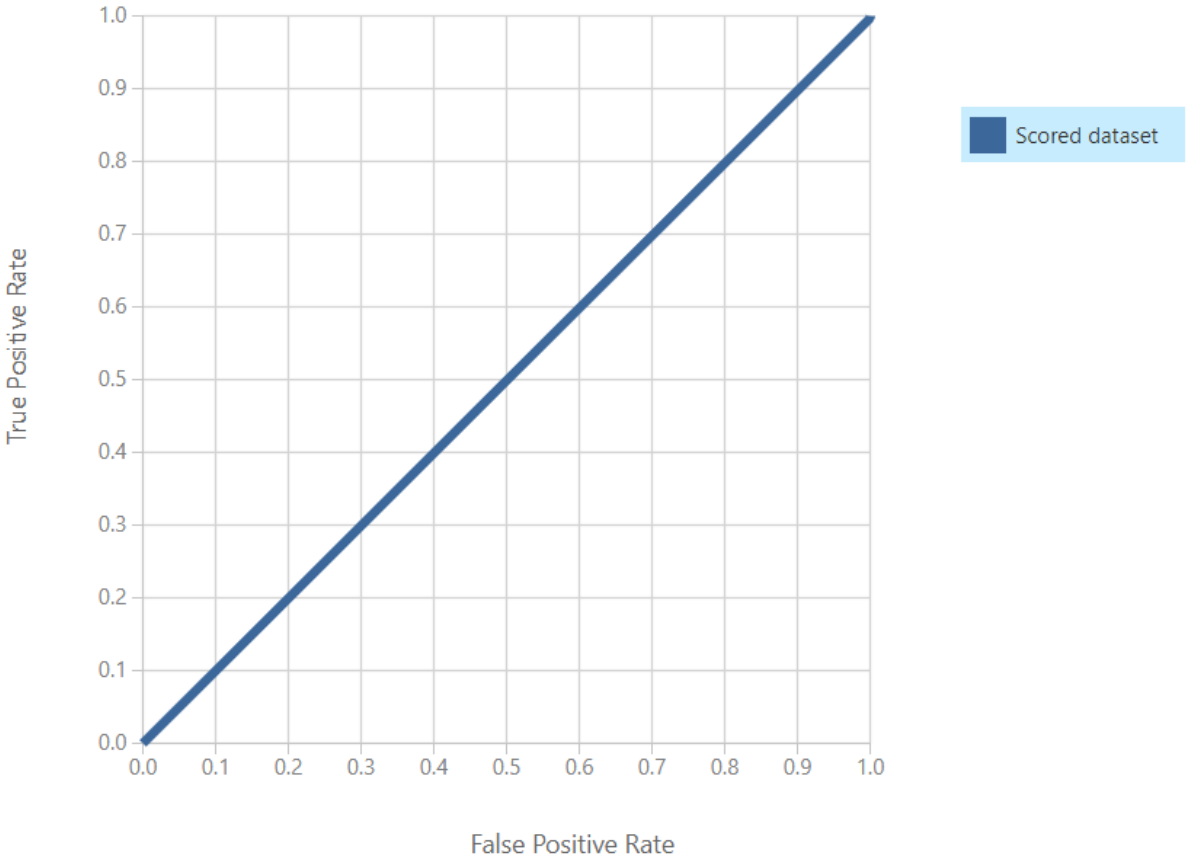


Dropped reservation status column

7.4 Experiment with different neural networks with parameters combinations given in the below below and run each model. (Note:- the dataset is without “reservation status”) (4 points)

No. of Hidden Nodes	Learning Rate	Momentum	Normalizer
128	0.0001	0.3	Do not Normalize
64	0.001	0.2	Mini-Max Normalizer
32	0.03	0.5	Gaussian Normalizer

1. FIRST Model

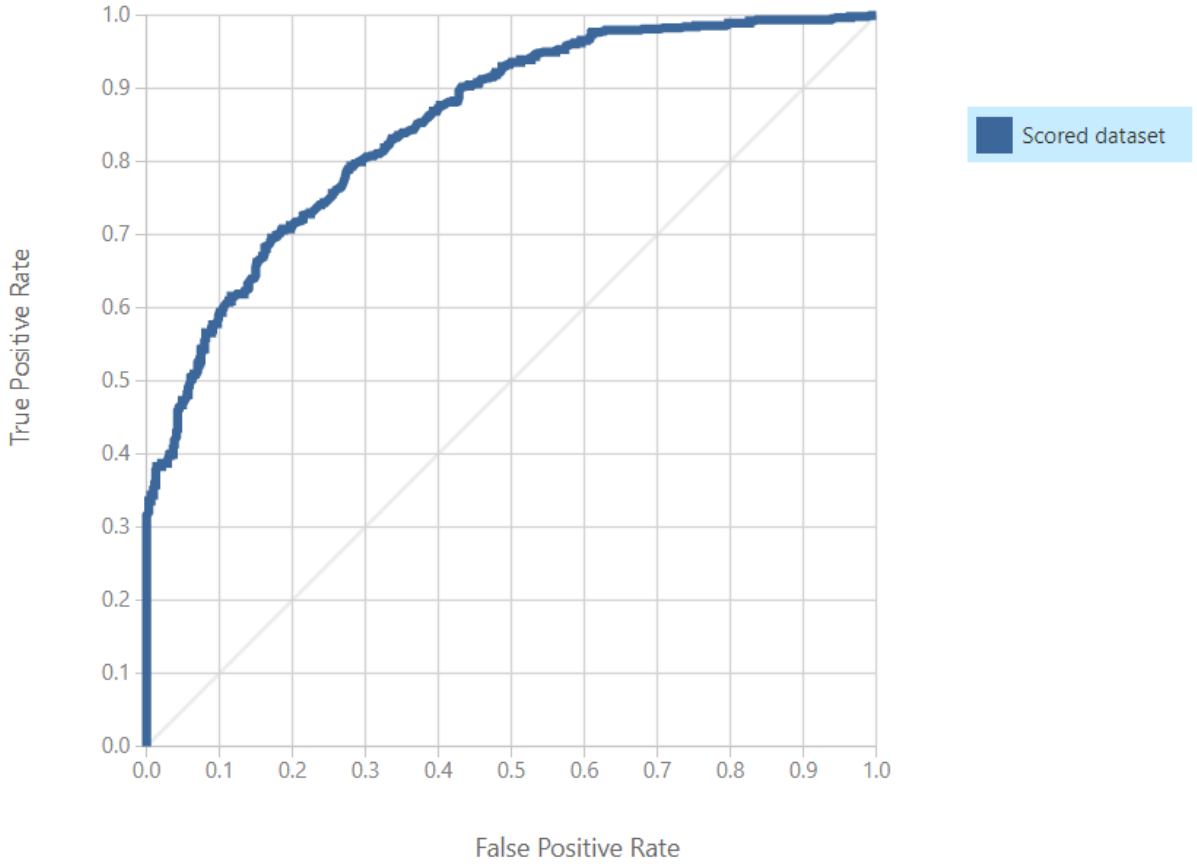


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
0	640	0.495	1.000	0.5	0.509

False Positive	True Negative	Recall	F1 Score
0	628	0.000	0.000

Positive Label	Negative Label
1	0

2. SECOND Model

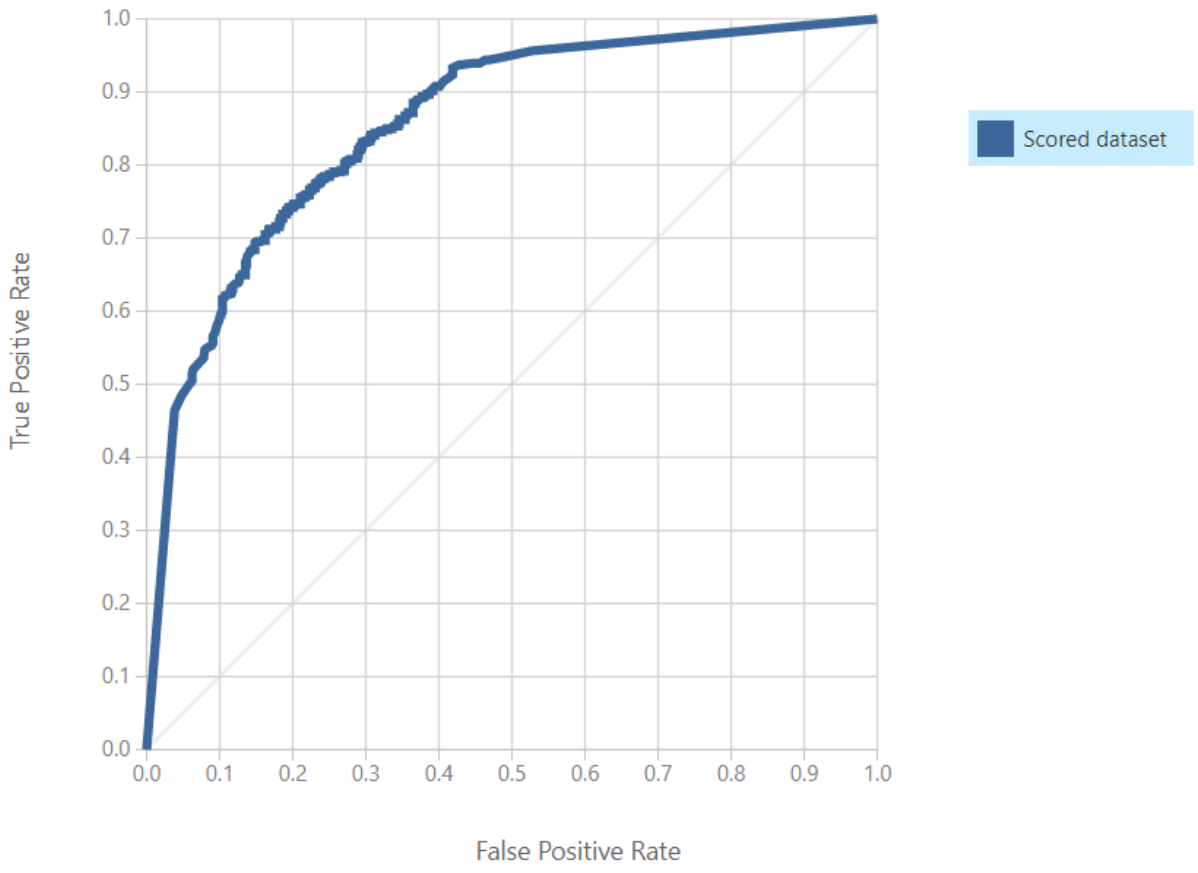


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
458	182	0.756	0.783	0.5	0.851

False Positive	True Negative	Recall	F1 Score
127	501	0.716	0.748

Positive Label	Negative Label
1	0

3. THIRD Model



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
504	136	0.769	0.762	0.5	0.859

False Positive	True Negative	Recall	F1 Score
157	471	0.787	0.775

Positive Label	Negative Label
1	0

7.5 State the difference between results obtained in train and test dataset ("excluding reservation status") clearly with reasons. State which model we should use for production. (4 points)

Ans.

We can see over fitting still persists but in case of other qualities we can observe that now reservation status was one of the crucial factors and is effecting the classification very much. This help us infer

reservation status playing a crucial feature in classification of this model. We can also see change in graph and accuracy of the models with less layers as well.

No. of Hidden Nodes	Learning Rate	Momentum	Normalizer
128	0.0001	0.3	Do not Normalize
64	0.001	0.2	Mini-Max Normalizer
32	0.03	0.5	Gaussian Normalizer

We can state this clearly that a model with 64 hidden layers and normalized one should be used for production and reservation status should be included.

If we have to exclude reservation status then we can adjust the model with 32 layers to get the following results.