Capstone Project: Notes-2 Report

By

Ritusri Mohan

CONTENTS

MODEL BUILDING

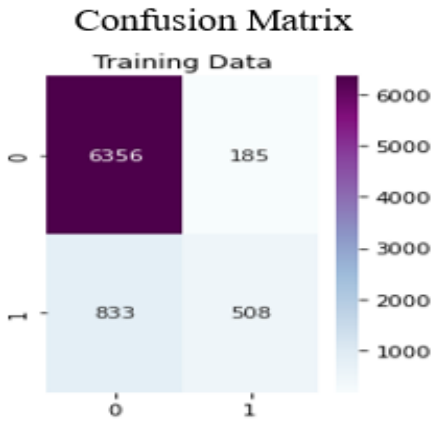
The dataset was split into two subsets known as training data and testing data in the ratio of 70:30.
In order for the machine learning model to find and learn patterns, training data is used which is a subset of our real dataset.
The unknown data is used to test your machine learning model after it has been constructed (using training data). This data, which is referred to as testing data, is used to assess the effectiveness and development of the training of the algorithms and to modify or optimize them for better outcomes.

Logistic Regression

A base model called 'lgmodel' was built using default parameters to be used as reference
The results of the base model on train data are as follows:

Classification Report

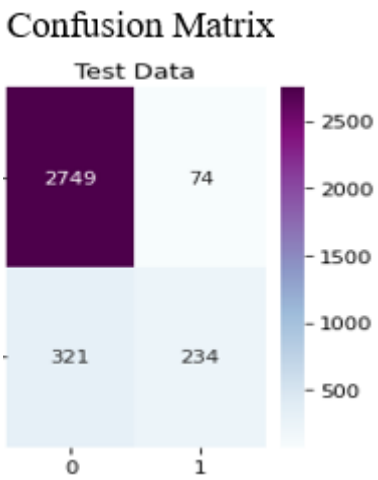
	precision	recall	f1-score	support
0	0.88	0.97	0.93	6541
1	0.73	0.38	0.50	1341
accuracy			0.87	7882
macro avg	0.81	0.68	0.71	7882
weighted avg	0.86	0.87	0.85	7882



The results of the base model on test data are as follows:

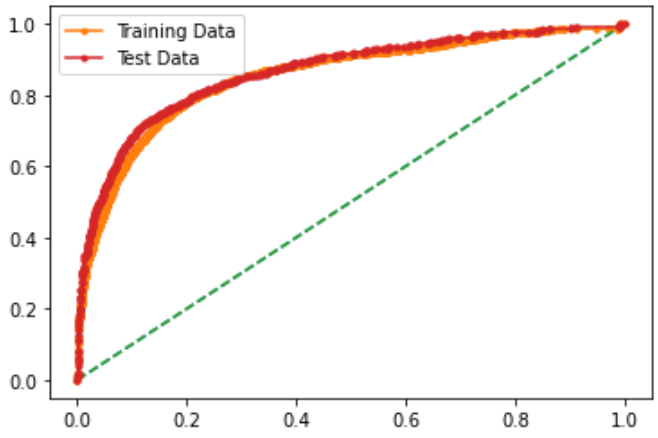
Classification Report

	precision	recall	f1-score	support
0	0.90	0.97	0.93	2823
1	0.76	0.42	0.54	555
accuracy			0.88	3378
macro avg	0.83	0.70	0.74	3378
weighted avg	0.87	0.88	0.87	3378



The ROC curve on both training and test data for base logistic regression model:

AUC for the Training Data: 0.854
AUC for the Test Data: 0.866



Observations:

- It can be seen that the recall value for both train and test data is 0.97 for majority class (i.e., 0).
- Since there is a huge difference in the recall values of the majority class (i.e., 0) and minority class (i.e., 1), it can be said that there is imbalance the model is biased towards majority class.
- The accuracy is not taken into consideration because there is imbalance in dataset.

A GridSearch model called ‘clf’ was built using hyperparameter tuning. The parameter grid to build the model is as follows:

```
GridSearchCV(cv=10, estimator=LogisticRegression(),
             param_grid={'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000],
                         'penalty': ['l1', 'l2'],
                         'solver': ['newton-cg', 'lbfgs', 'liblinear']},
             scoring='accuracy')
```

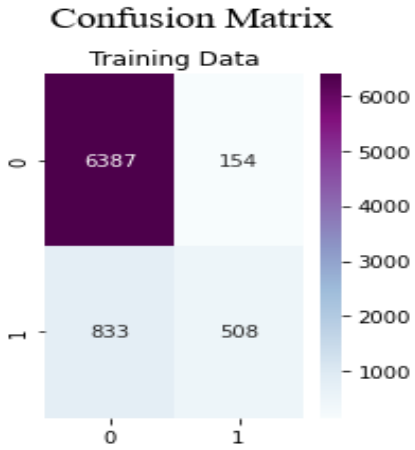
The tuned hyperparameters and accuracy of the best model called ‘logreg’ are as follows:

```
Tuned Hyperparameters : {'C': 10, 'penalty': 'l1', 'solver': 'liblinear'}
Accuracy : 0.8755391390502659
```

The results of the GridSearch model on train data are as follows:

Classification Report

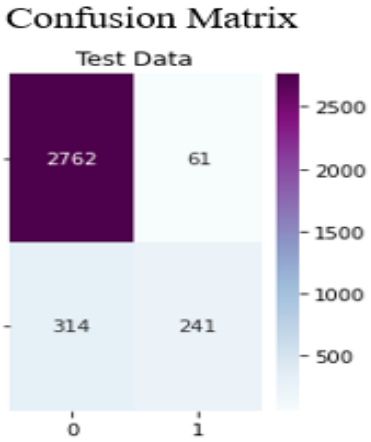
	precision	recall	f1-score	support
0	0.88	0.98	0.93	6541
1	0.77	0.38	0.51	1341
accuracy			0.87	7882
macro avg	0.83	0.68	0.72	7882
weighted avg	0.86	0.87	0.86	7882



The results of the GridSearch model on test data are as follows:

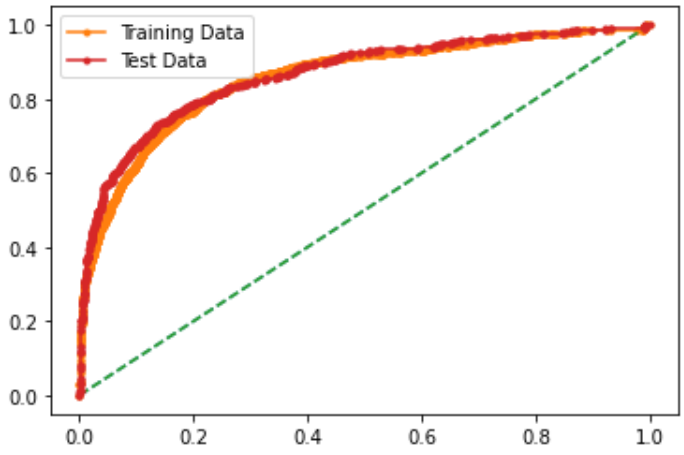
Classification Report

	precision	recall	f1-score	support
0	0.90	0.98	0.94	2823
1	0.80	0.43	0.56	555
accuracy			0.89	3378
macro avg	0.85	0.71	0.75	3378
weighted avg	0.88	0.89	0.87	3378



The ROC curve on both training and test data:

AUC for the Training Data: 0.858
AUC for the Test Data: 0.867



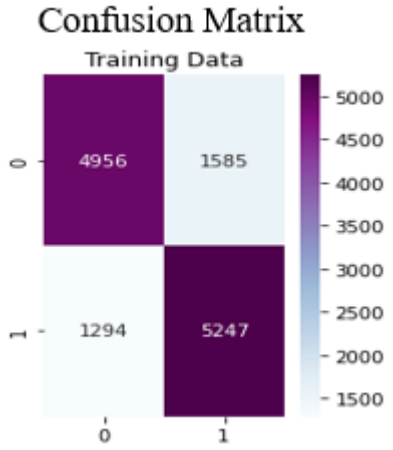
Observations:

- It can be seen that the recall value for both train and test data is 0.98 for majority class (i.e., 0)
- Since there is a huge difference in the recall values of the majority class (i.e., 0) and minority class (i.e., 1), it can be said that there is imbalance and the model is biased towards majority class.
- The accuracy is not taken into consideration because there is imbalance in dataset.

A SMOTE model called ‘lr1’ was built to overcome imbalance in dataset.
The results of the SMOTE model on train data are as follows:

Classification Report

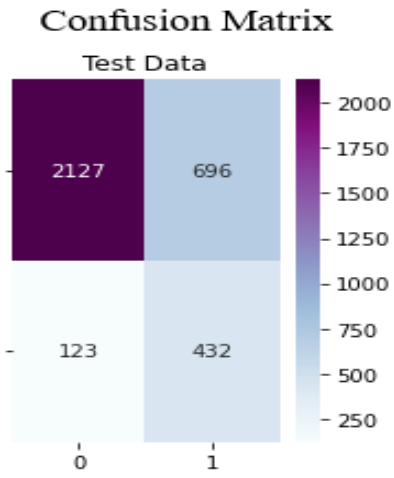
	precision	recall	f1-score	support
0	0.79	0.76	0.77	6541
1	0.77	0.80	0.78	6541
accuracy			0.78	13082
macro avg	0.78	0.78	0.78	13082
weighted avg	0.78	0.78	0.78	13082



The results of the SMOTE model on test data are as follows:

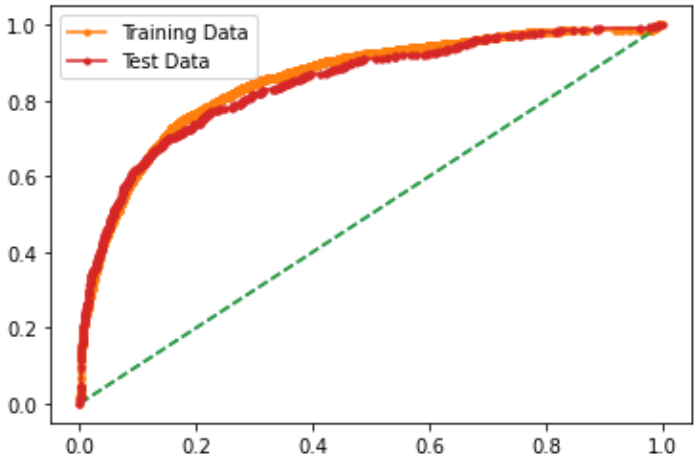
Classification Report

	precision	recall	f1-score	support
0	0.79	0.76	0.77	6541
1	0.77	0.80	0.78	6541
accuracy			0.78	13082
macro avg	0.78	0.78	0.78	13082
weighted avg	0.78	0.78	0.78	13082



The ROC curve on both training and test data:

AUC for the Training Data: 0.856
AUC for the Test Data: 0.845

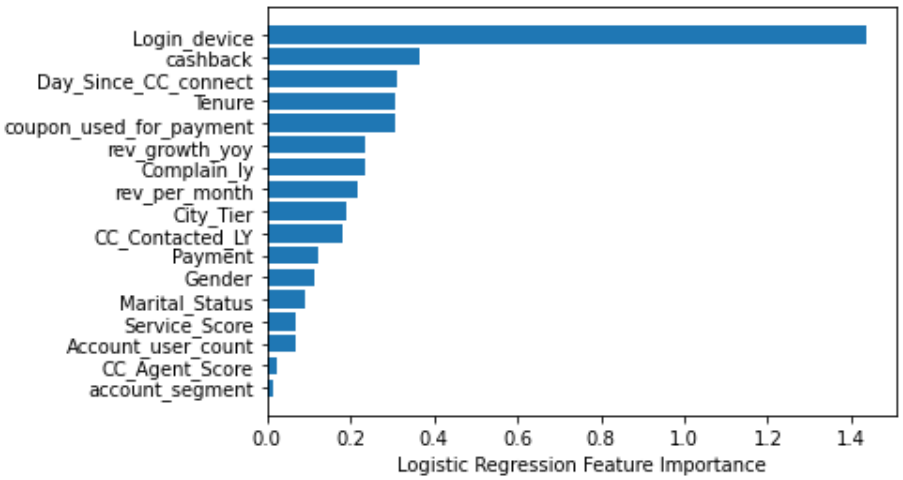


Observations:

- It can be seen that the recall value for both train and test data is 0.76.
- The accuracy of the model on both training and test is 0.78.
- Even though the accuracy of SMOTE model is lesser than the base model and GridSearch model, it chosen to as the better model because the problem of imbalance has been overcome.

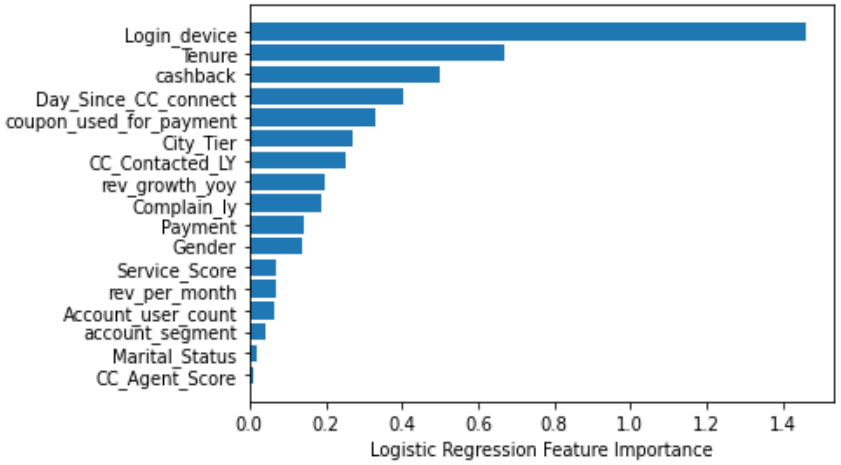
The top 6 important features of the dataset provided by logistic regression base model (before resampling) are:

- Login_device
- Cashback
- Day_Since_CC_connect
- Tenure
- Coupon_used for payment
- Rev_growth_yoy



The top 6 important features of the dataset provided by logistic regression SMOTE model (after resampling) are:

- Login_device
- Tenure
- Cashback
- Day_Since_CC_connect
- Coupon_used_for payment
- City_Tier

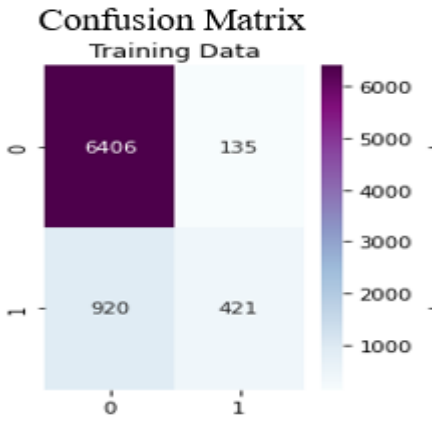


Linear Discriminant Analysis

A linear discriminant model called 'lda_model' was built using default parameters.
The results of the base model on train data are as follows:

Classification Report

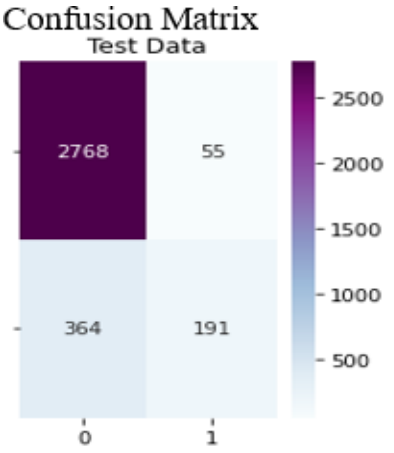
	precision	recall	f1-score	support
0	0.87	0.98	0.92	6541
1	0.76	0.31	0.44	1341
accuracy			0.87	7882
macro avg	0.82	0.65	0.68	7882
weighted avg	0.85	0.87	0.84	7882



The results of the base model on test data are as follows:

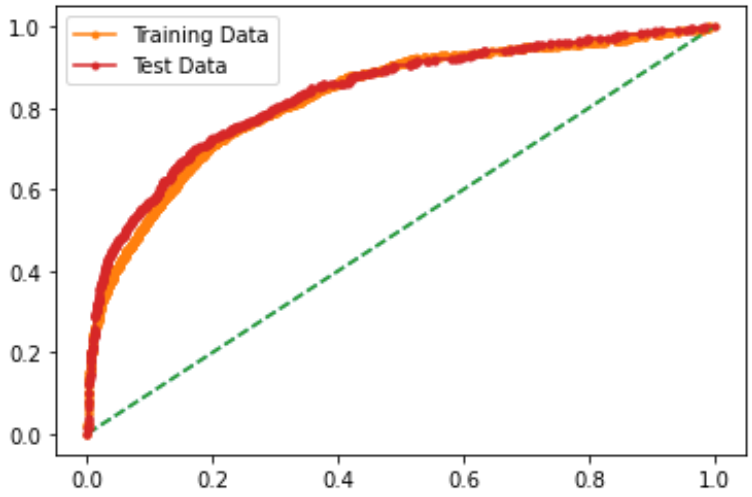
Classification Report

	precision	recall	f1-score	support
0	0.88	0.98	0.93	2823
1	0.78	0.34	0.48	555
accuracy			0.88	3378
macro avg	0.83	0.66	0.70	3378
weighted avg	0.87	0.88	0.86	3378



The ROC curve on both training and test data:

AUC for the Training Data: 0.826
AUC for the Test Data: 0.836



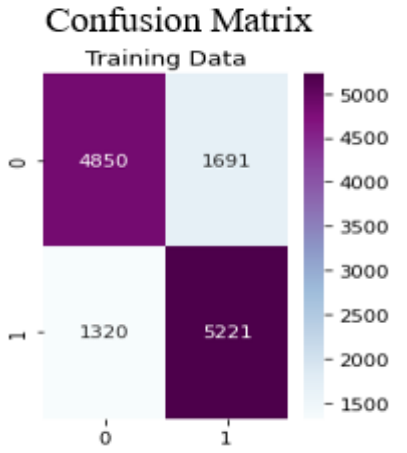
Observations:

- It can be seen that the recall value for both train and test data is 0.98 for majority class (i.e., 0).
- Since there is a huge difference in the recall values of the majority class (i.e., 0) and minority class (i.e., 1), it can be said that there is an imbalance and the model is biased towards majority class.
- The accuracy is not taken into consideration because there is imbalance in dataset.

A SMOTE model called ‘lda1’ was built to overcome imbalance in dataset.
The results of the SMOTE model on train data are as follows

Classification Report

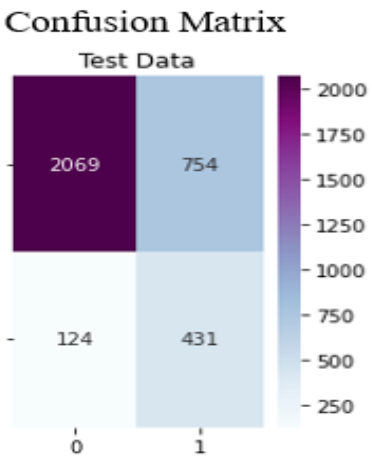
	precision	recall	f1-score	support
0	0.79	0.74	0.76	6541
1	0.76	0.80	0.78	6541
accuracy			0.77	13082
macro avg	0.77	0.77	0.77	13082
weighted avg	0.77	0.77	0.77	13082



The results of the SMOTE model on test data are as follows:

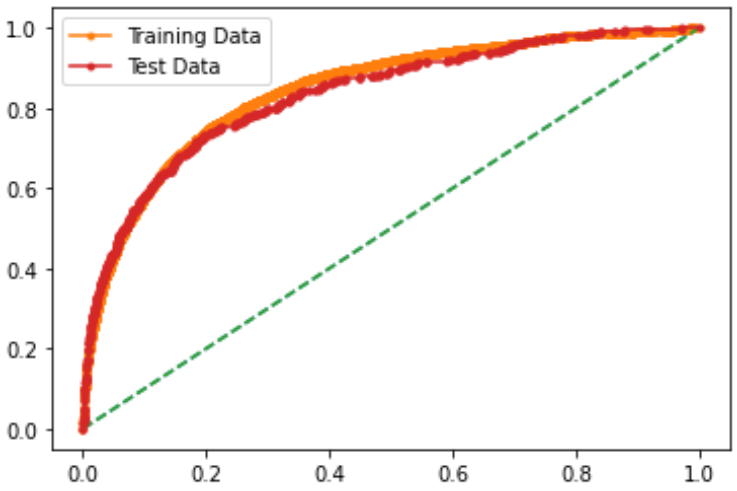
Classification Report

	precision	recall	f1-score	support
0	0.94	0.73	0.82	2823
1	0.36	0.78	0.50	555
accuracy			0.74	3378
macro avg	0.65	0.75	0.66	3378
weighted avg	0.85	0.74	0.77	3378



The ROC curve on both training and test data for SMOTE model:

AUC for the Training Data: 0.826
AUC for the Test Data: 0.836

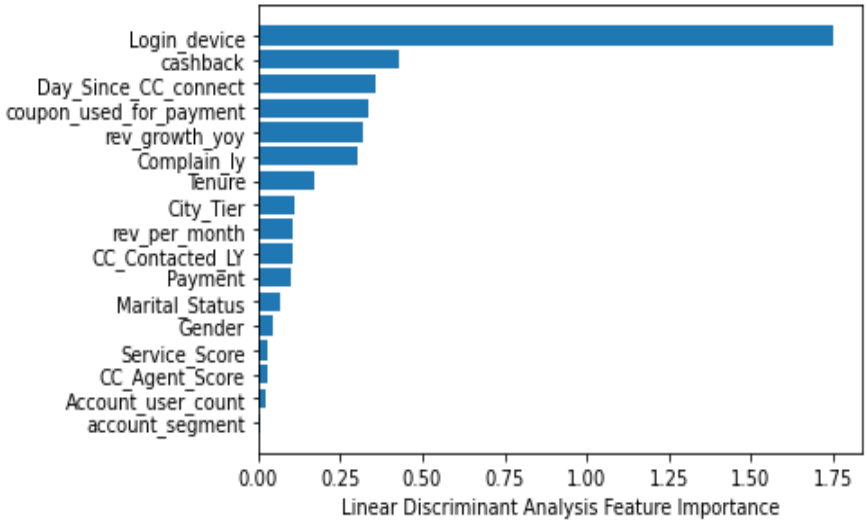


Observations:

- It can be seen that the recall value for train data is 0.74 and test data is 0.73 for the majority class (i.e., 0)
- The accuracy of the model on for training data is 0.77 and test data is 0.74.
- Even though the accuracy of SMOTE model is lesser than the base model, it is chosen to as the better model because the problem of imbalance has been overcome.

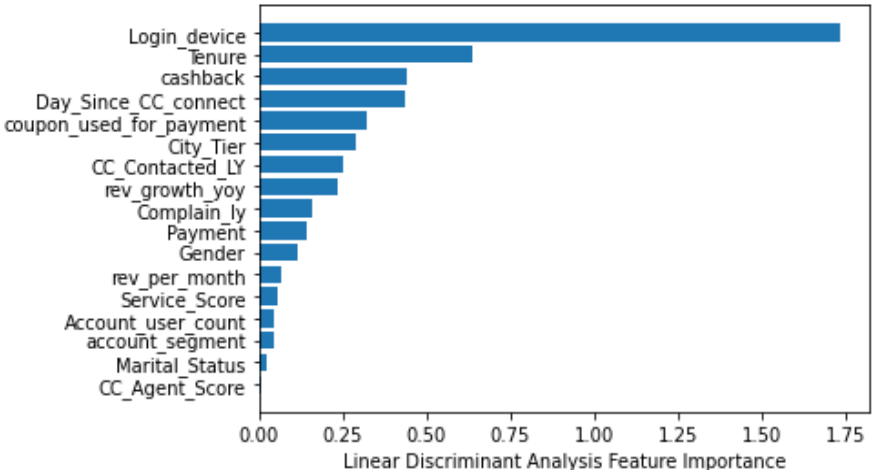
The top 6 important features of the dataset provided by linear discriminant analysis base (before resampling) are:

- Login_device
- Cashback
- Day_Since_CC_connect
- Coupon_used_for_payment
- Revv_growth_yoy
- Complain_ly



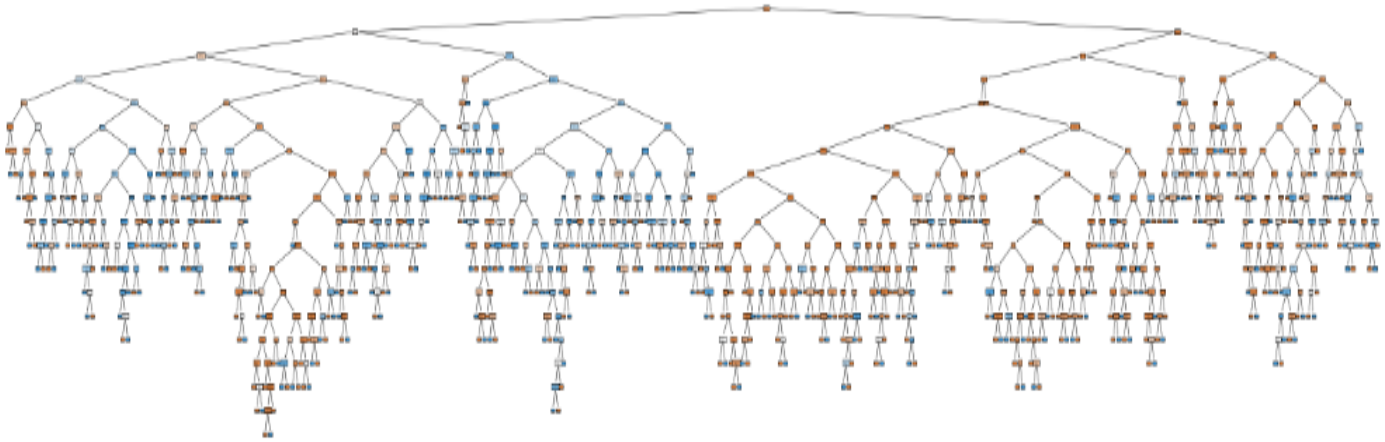
The top 6 important features of the dataset provided by linear discriminant analysis SMOTE model (after resampling) are:

- Login_device
- Tenure
- Cashback
- Day_Since_CC_connect
- Coupon_used_forpayment
- City_Tier



Decision Tree Classifier

The base decision tree with default parameters can be seen below:



GridSearchCV was used for hyperparameter. The following parameter grid was used to search the best parameters:

```
GridSearchCV(cv=5, estimator=DecisionTreeClassifier(random_state=1), n_jobs=-1,
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [10, 20, 30],
                         'min_samples_leaf': [9, 10, 11, 12],
                         'min_samples_split': [20, 25, 30]},
             verbose=1)
```

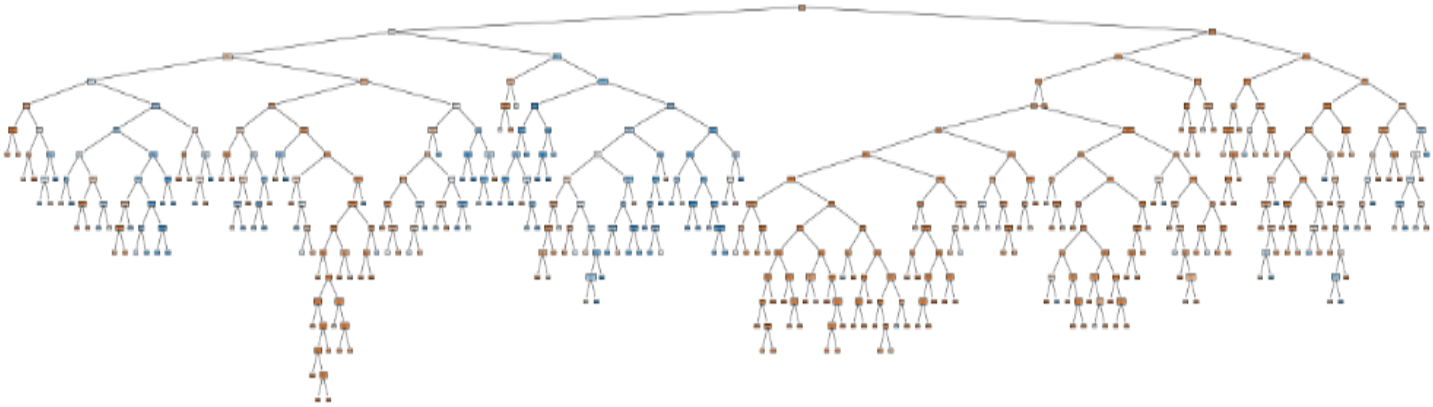
The best parameters and the best estimator were found from combination of given parameters from the parameter grid:

```
{'criterion': 'gini', 'max_depth': 20, 'min_samples_leaf': 10, 'min_samples_split': 20}

DecisionTreeClassifier(max_depth=20, min_samples_leaf=10, min_samples_split=20,
                       random_state=1)
```

The GridSearch decision tree model named ‘reg_dtcl’ was built using the best parameters and the best score was found to be 0. 903.

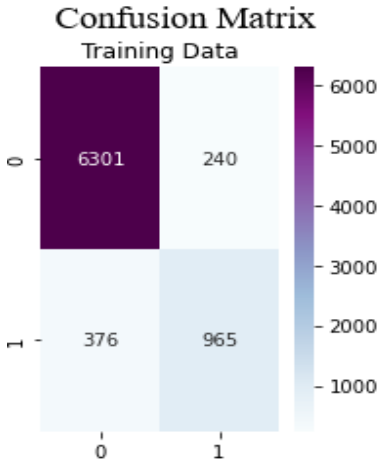
The tree of this model can be seen below.



The results of the GridSearch model on train data are as follows:

Classification Report

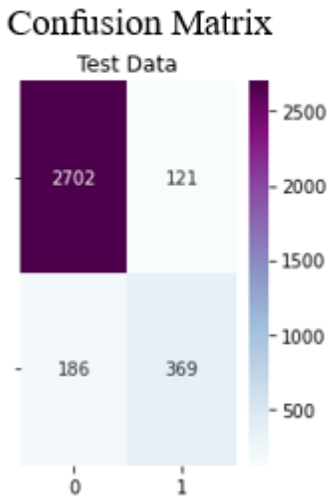
	precision	recall	f1-score	support
0	0.94	0.96	0.95	6541
1	0.80	0.72	0.76	1341
accuracy			0.92	7882
macro avg	0.87	0.84	0.86	7882
weighted avg	0.92	0.92	0.92	7882



The results of the GridSearch model on test data are as follows:

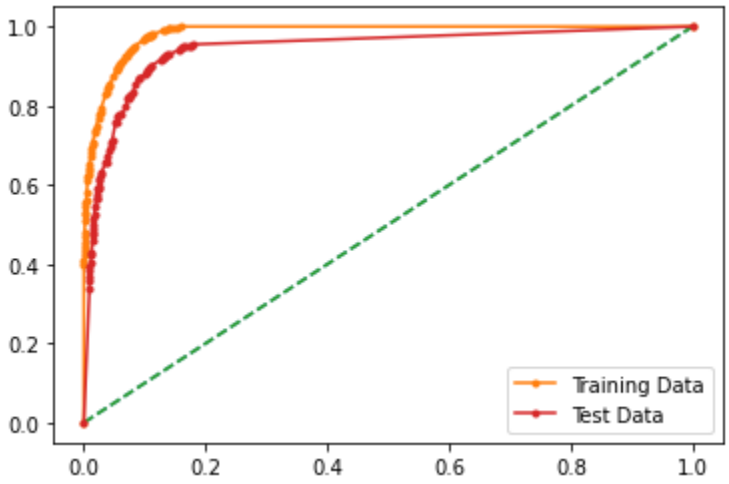
Classification Report

	precision	recall	f1-score	support
0	0.94	0.96	0.95	2823
1	0.75	0.66	0.71	555
accuracy			0.91	3378
macro avg	0.84	0.81	0.83	3378
weighted avg	0.91	0.91	0.91	3378



The ROC curve on both training and test data:

AUC for the Training Data: 0.983
AUC for the Test Data: 0.943



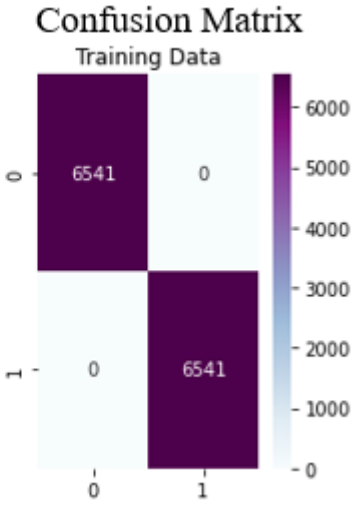
Observations:

- It can be seen that the recall value for both train and test data is 0.96 for majority class (i.e., 0)
- Since there is a huge difference in the recall values of the majority class (i.e., 0) and minority class (i.e., 1), it can be said that there is imbalance and the model is biased towards majority class.
- The accuracy is not taken into consideration because there is imbalance in dataset.

A SMOTE model called ‘dtcl1’ was built to overcome imbalance in dataset.
The results of the SMOTE model on train data are as follows:

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	6541
1	1.00	1.00	1.00	6541
accuracy			1.00	13082
macro avg	1.00	1.00	1.00	13082
weighted avg	1.00	1.00	1.00	13082

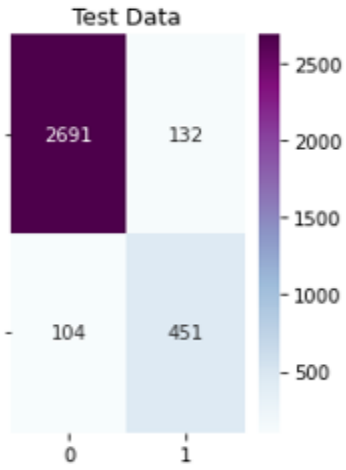


The results of the SMOTE model on test data are as follows:

Classification Report

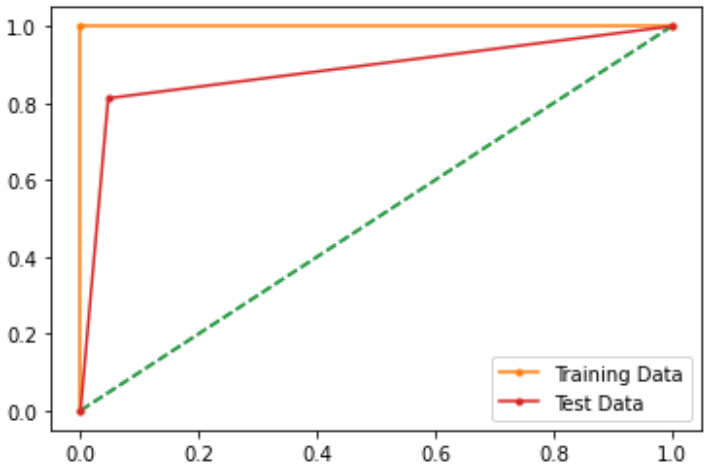
	precision	recall	f1-score	support
0	0.96	0.95	0.96	2823
1	0.77	0.81	0.79	555
accuracy			0.93	3378
macro avg	0.87	0.88	0.88	3378
weighted avg	0.93	0.93	0.93	3378

Confusion Matrix



The ROC curve on both training and test data for SMOTE model:

AUC for the Training Data: 1.00
AUC for the Test Data: 0.833



Observations:

- It can be seen that the recall value for train data is 1.00 and test data is 0.95 for the majority class (i.e., 0).
- The accuracy of the model on for training data is 1.0 and test data is 0.93 for the majority class.
- There seems to be a problem of overfitting as the recall value of minority class in testing data is 0.81 and that of training data is 1.0.

Another SMOTE model called ‘reg_dtcl2’ was built by using the hyperparameter tuning method known as GridSearchCV. It was built to overcome the overfitting problem occurring in the above model. The following parameter grid was used to search the best parameters:

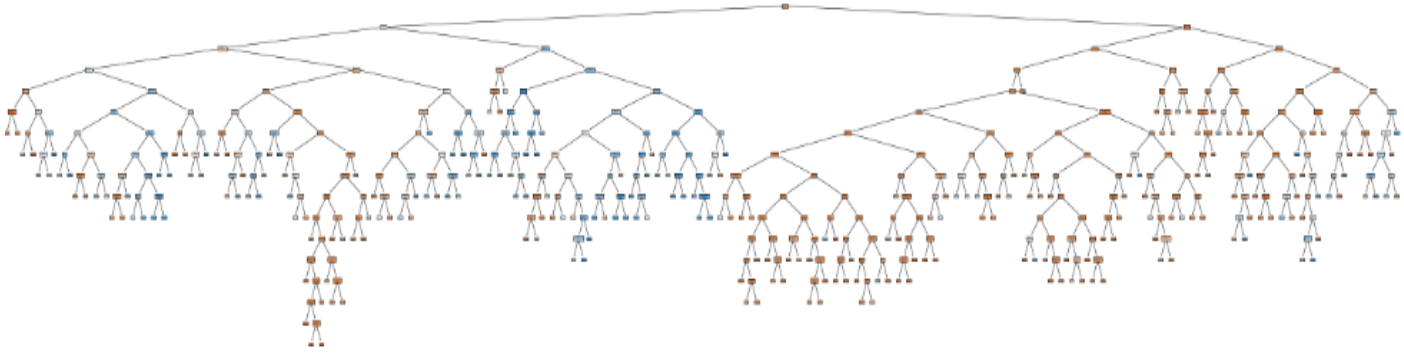
```
GridSearchCV(cv=5, estimator=DecisionTreeClassifier(), n_jobs=-1
             param_grid={'criterion': ['gini', 'entropy'],
                         'max_depth': [10, 20, 30],
                         'min_samples_leaf': [9, 10, 11, 12],
                         'min_samples_split': [20, 25, 30]},
             verbose=1)
```

After the several iterations, the following hyperparameters emerged as best:

```
{'criterion': 'entropy', 'max_depth': 20, 'min_samples_leaf': 9, 'min_samples_split': 20}
```

```
DecisionTreeClassifier(criterion='entropy', max_depth=20, min_samples_leaf=9,
                       min_samples_split=20)
```

The decision tree of the tuned hyperparameter SMOTE model is below:



The results of the tuned hyperparameter SMOTE model on train data are as follows:

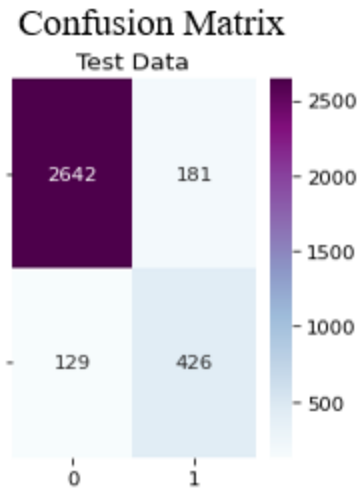
Classification Report

	precision	recall	f1-score	support
0	0.96	0.96	0.96	6541
1	0.96	0.96	0.96	6541
accuracy			0.96	13082
macro avg	0.96	0.96	0.96	13082
weighted avg	0.96	0.96	0.96	13082

The results of the tuned hyperparameter SMOTE model on test data are as follows:

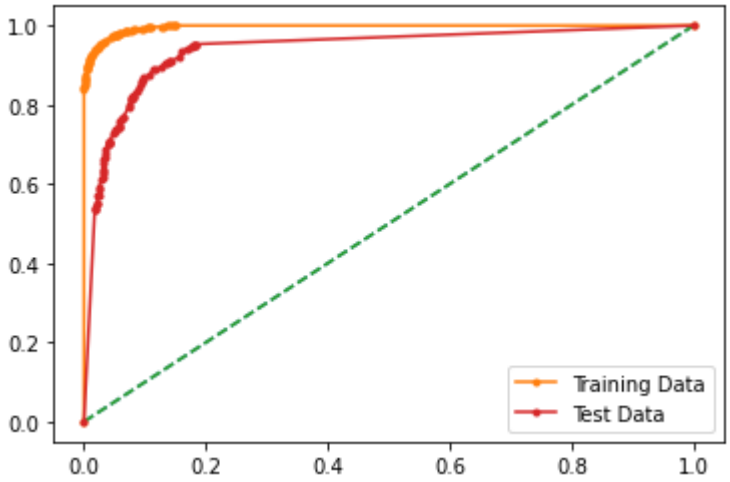
Classification Report

	precision	recall	f1-score	support
0	0.95	0.94	0.94	2823
1	0.70	0.77	0.73	555
accuracy			0.91	3378
macro avg	0.83	0.85	0.84	3378
weighted avg	0.91	0.91	0.91	3378



The ROC curve on both training and test data for hyperparameter tuned SMOTE model:

AUC for the Training Data: 0.996
AUC for the Test Data: 0.939

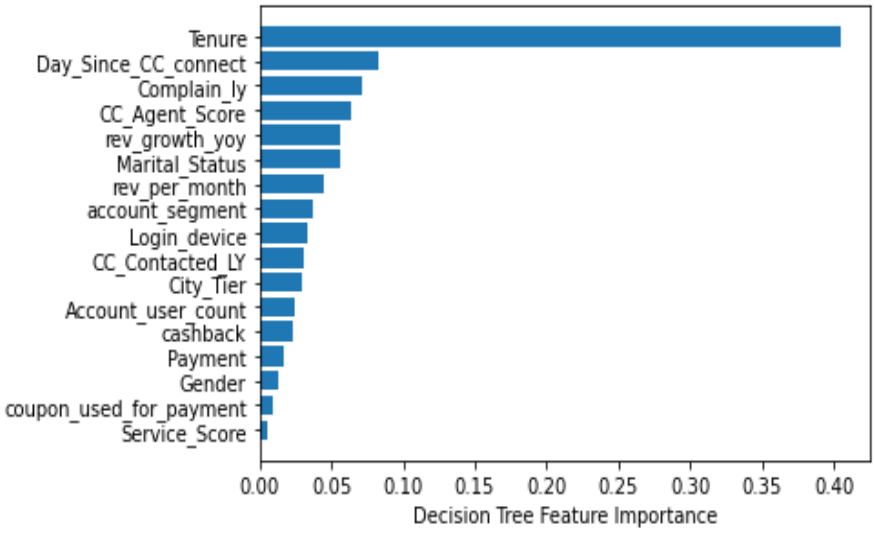


Observations:

- It can be seen that the recall value for train data is 0.96 and test data is 0.94 for the majority class (i.e., 0).
- The accuracy of the model on for training data is 0.96 and test data is 0.91.
- Even though there is some overfitting, the hyperparameter tuned SMOTE model is chosen as the recall values and AUC values for both training and test data are close.

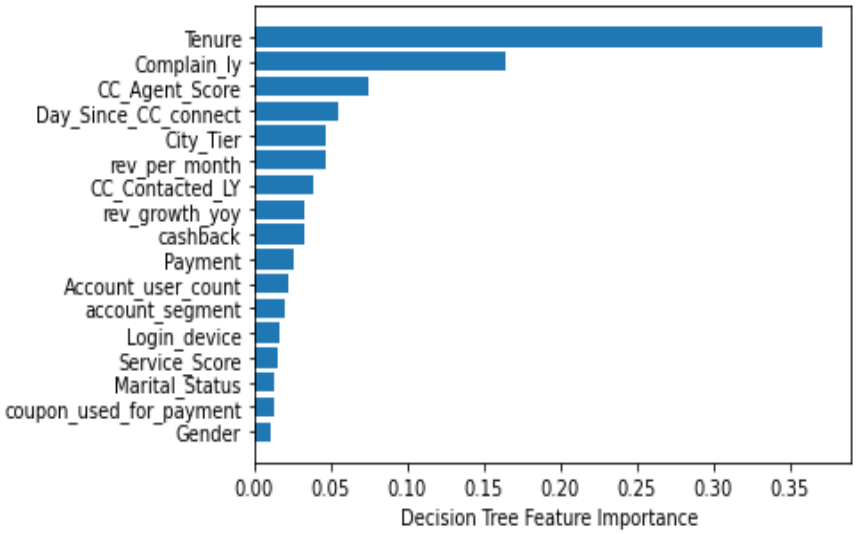
The top 6 important features of the dataset provided by decision tree grid base model (before resampling) are:

- Tenure
- Day_Since_CC_connect
- Complain_ly
- CC_Agent_Score
- Rev_growth_yoy
- Marital_Status



The top 6 important features of the dataset provided by decision tree SMOTE model (after resampling) are:

- Tenure
- Complain_ly
- CC_Agent_Score
- Day_Since_CC_connect
- City_Tier
- Rev_per_month

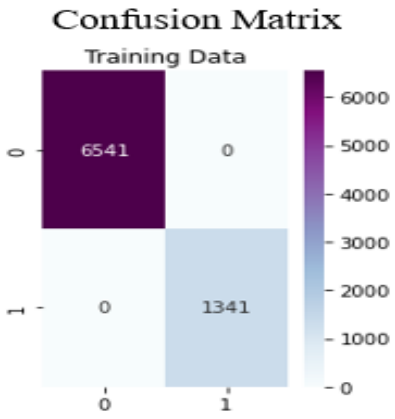


Random Forest Classifier

The base model called 'rfcl' was built with default parameters (i.e., n_estimators=100). The results of the base model on train data are as follows:

Classification Report

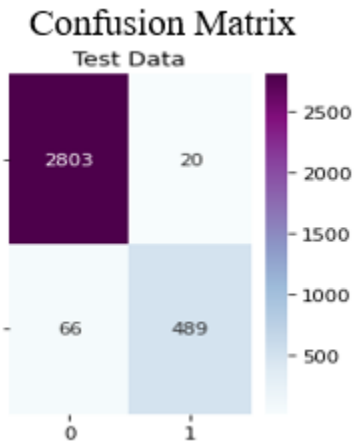
	precision	recall	f1-score	support
0	1.00	1.00	1.00	6541
1	1.00	1.00	1.00	1341
accuracy			1.00	7882
macro avg	1.00	1.00	1.00	7882
weighted avg	1.00	1.00	1.00	7882



The results of the base model on test data are as follows:

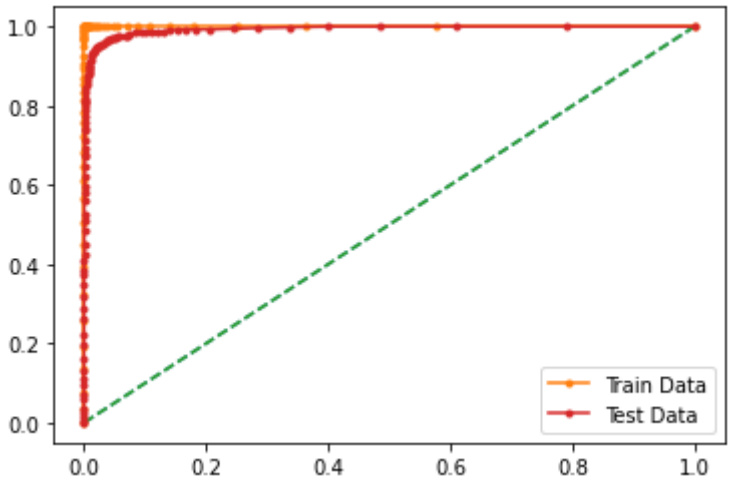
Classification Report

	precision	recall	f1-score	support
0	0.98	0.99	0.98	2823
1	0.96	0.88	0.92	555
accuracy			0.97	3378
macro avg	0.97	0.94	0.95	3378
weighted avg	0.97	0.97	0.97	3378



The ROC curve on both training and test data for base model:

AUC for the Training Data: 1.00
AUC for the Test Data: 0.994

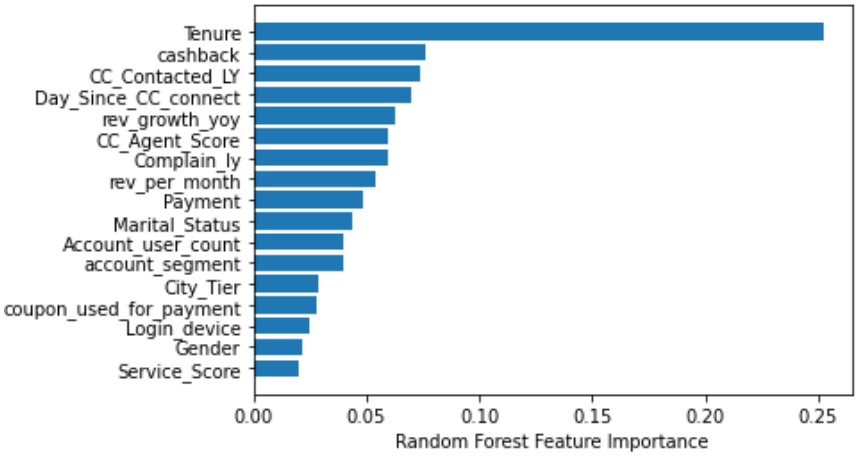


Observations:

- It can be seen that the recall value for train data is 1.00 and test data is 0.99 for the majority class (i.e., 0).
- The accuracy of the model on for training data is 1.00 and test data is 0.97.
- On observing the overall metrics of the base model 'rfcl' it can be seen that Random Forest Classifier emerged as one of the best fitting model even without correcting for class imbalance in the dataset.

The top 6 important features of the dataset provided by random forest model are –

- Tenure
- Cashback
- CC_Contacted_LY
- Day_Since_CC_connect
- Rev_growth_yoy
- CC_Agent_Score



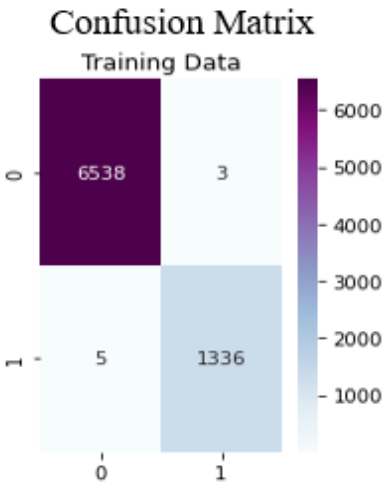
XG Boosting

An Extreme Gradient Boosting (XGB) base model called 'xgb' was built with default parameters.

The results of the base model on train data are as follows:

Classification Report

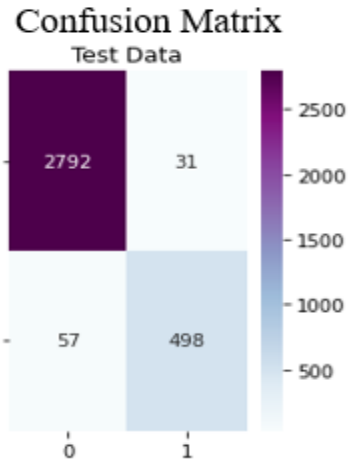
	precision	recall	f1-score	support
0	1.00	1.00	1.00	6541
1	1.00	1.00	1.00	1341
accuracy			1.00	7882
macro avg	1.00	1.00	1.00	7882
weighted avg	1.00	1.00	1.00	7882



The results of the base model on test data are as follows:

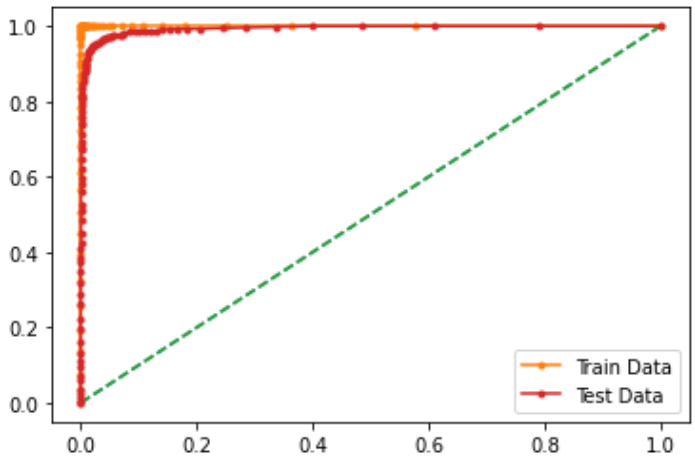
Classification Report

	precision	recall	f1-score	support
0	0.98	0.99	0.98	2823
1	0.94	0.90	0.92	555
accuracy			0.97	3378
macro avg	0.96	0.94	0.95	3378
weighted avg	0.97	0.97	0.97	3378



The ROC curve on both training and test data for base model:

AUC for the Training Data: 1.00
AUC for the Test Data: 0.994

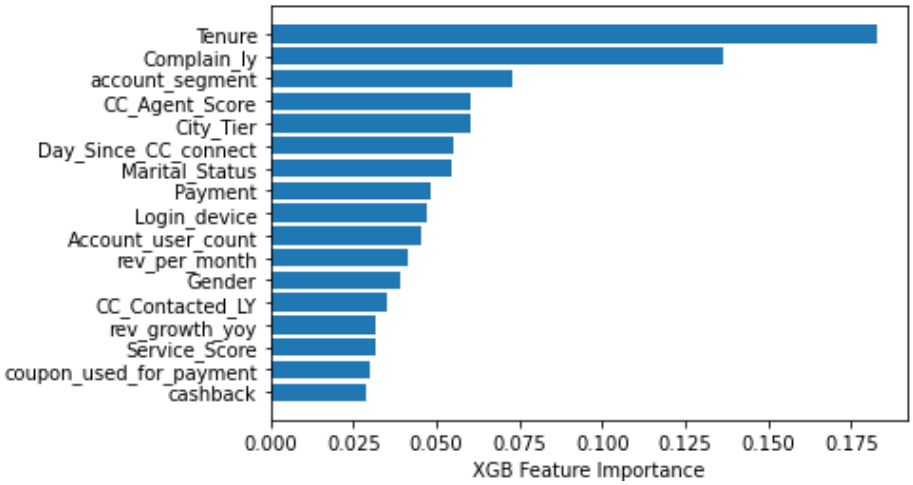


Observations:

- Out of all the customers that XGB model predicted to non-churn only 98% actually did.
- Out of all the customers that actually did not churn, the model only predicted this outcome correctly for 99% of those customers.
- Since f1-score of 0.943 is very close to 1, it tells us that the model does a very good job of predicting whether or not the companies will default.
- On observing the overall metrics of the base model ‘xgb’ it can be seen that XG Boosting emerged as one of the best fitting model even without correcting for class imbalance in the dataset.

The top 6 important features of the dataset provided by XGB model are –

- Tenure
- Complain_LY
- Account_segment
- CC_Agent_Score
- City_Tier
- Day_Since_CC_connect



COMPOSITE VIEW OF MODELS

Logistic Regression

SMOTE MODEL SELECTED!

	Log Base Train	Log Base Test	Log Grid Train	Log Grid Test	Log Smote Train	Log Smote Test
Accuracy	0.87	0.88	0.87	0.89	0.78	0.76
Recall	0.97	0.97	0.98	0.98	0.76	0.75
AUC	0.85	0.87	0.86	0.87	0.86	0.85
Precision	0.88	0.90	0.88	0.90	0.79	0.95
F1 Score	0.93	0.93	0.93	0.94	0.77	0.84

Linear Discriminant Analysis

SMOTE MODEL SELECTED!

	LDA Base Train	LDA Base Test	LDA Smote Train	LDA Smote Test
Accuracy	0.87	0.88	0.77	0.74
Recall	0.98	0.98	0.74	0.73
AUC	0.83	0.84	0.84	0.84
Precision	0.87	0.88	0.79	0.94
F1 Score	0.92	0.93	0.76	0.82

Decision Tree Classifier

HYPERPARAMETER TUNED SMOTE MODEL SELECTED!

	DTCL Grid Train	DTCL Grid Test	DTCL Smote Train	DTCL Smote Test	DTCL Grid Smote Train	DTCL Grid Smote Test
Accuracy	0.94	0.92	1.0	0.93	0.96	0.91
Recall	0.97	0.95	1.0	0.95	0.96	0.93
AUC	0.98	0.94	1.0	0.89	1.00	0.93
Precision	0.96	0.94	1.0	0.96	0.96	0.95
F1 Score	0.97	0.95	1.0	0.96	0.96	0.94

Random Forest Classifier

BASE MODEL SELECTED!

	RF Base Train	RF Base Test
Accuracy	1.00	0.97
Recall	0.97	0.95
AUC	1.00	0.99
Precision	0.96	0.94
F1 Score	0.97	0.95

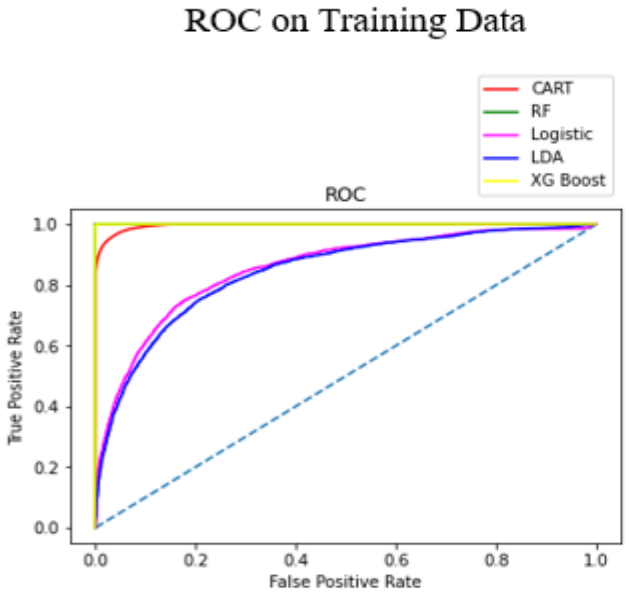
XG Boosting

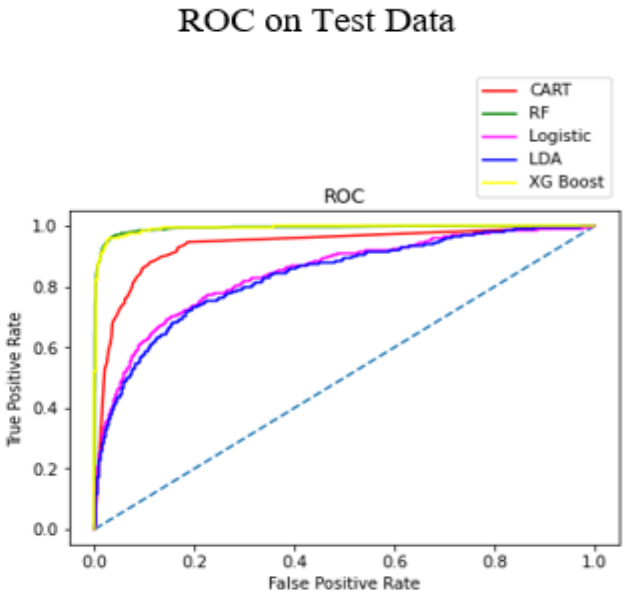
BASE MODEL SELECTED!

	XGB Base Train	XGB Base Test
Accuracy	1.0	0.97
Recall	1.0	0.99
AUC	1.0	0.99
Precision	1.0	0.98
F1 Score	1.0	0.98

BEST MODEL SELECTION: PERFORMANCE METRICS

	Log Smote Train	Log Smote Test	LDA Smote Train	LDA Smote Test	CART Grid Smote Train	CART Grid Smote Test	RF Base Train	RF Base Test	XGB Base Train	XGB Base Test
Accuracy	0.78	0.76	0.77	0.74	0.96	0.91	1.00	0.97	1.0	0.97
Recall	0.76	0.75	0.74	0.73	0.96	0.93	0.97	0.95	1.0	0.99
AUC	0.86	0.85	0.84	0.84	1.00	0.93	1.00	0.99	1.0	0.99
Precision	0.79	0.95	0.79	0.94	0.96	0.95	0.96	0.94	1.0	0.98
F1 Score	0.77	0.84	0.76	0.82	0.96	0.94	0.97	0.95	1.0	0.98





In the above table it can be seen that after resampling,

- Logistic regression model gives an accuracy of 0.76
- Linear discriminant analysis model gives an accuracy of 0.74
- Cart model gives an accuracy of 0.91.

But Random forest classifier and Extreme gradient boosting(XGB) give the highest model accuracy of 0.97 on test data. On comparing the recall and precision and F1 score values for the two models, it can be seen that XGB model has higher values which means that it is the better model.

Therefore, the **XGB model** is the best model for the given data.

INSIGHTS & RECOMMENDATIONS:

- The SMOTE models for Logistic regression, Linear discriminant analysis and Decision tree classifier showed better result when compared to their respective base models. This was due to imbalance in the data points. To obtain better results more data for the minority class (i.e, churn class) are preferred.
- The base model of Random forest classifier established itself as a good model even though there was an imbalance.
- From the XGB model, it can be seen that ‘Tenure’, ‘Complain_LY’, ‘CC_Agent_Score’, ‘Account_segment’, ‘City_Tier’ have a very strong influence on customer churn.
- It is important to attend to the needs of customers, especially those who have a tenure of 1-2 years as there is always a risk to losing them to competition.
- A regular survey should be conducted in order to know the problems faced by the customers and their interaction with Customer care.

THE END