CHURN PREDICTION FOR THE TELECOMMUNICATION SECTOR

Background

- According to Wikipidea the telecommunications industries within the sector of information and communication technology is made up of all telecommunications/telephone companies and internet service providers and plays a crucial role in the evolution of mobile communications and the information society.

- Traditional telephone calls continue to be the industry's biggest revenue generator, but thanks to advances in network technology, telecom today is less about voice and increasingly about text (messaging, email) and images (e.g. video streaming). High-speed internet access for computer-based data applications such as broadband information services and interactive entertainment is pervasive. Digital subscriber line (DSL) is the main broadband telecom technology. The fastest growth comes from (value-added) services delivered over mobile networks.

Problem Statement

- The telecommunication industry is quite capital intensive and requires recruitment and retention of a wide customer base in order to spread overheads. The industry guidance is that the cost of acquiring a new customer is more than the cost of retaining an existing customer. Consequently, churn prediction ,which is detecting which customers are likely to leave , is imperative because the company can then focus on retention of existing customers who are likely to leave.

Objective

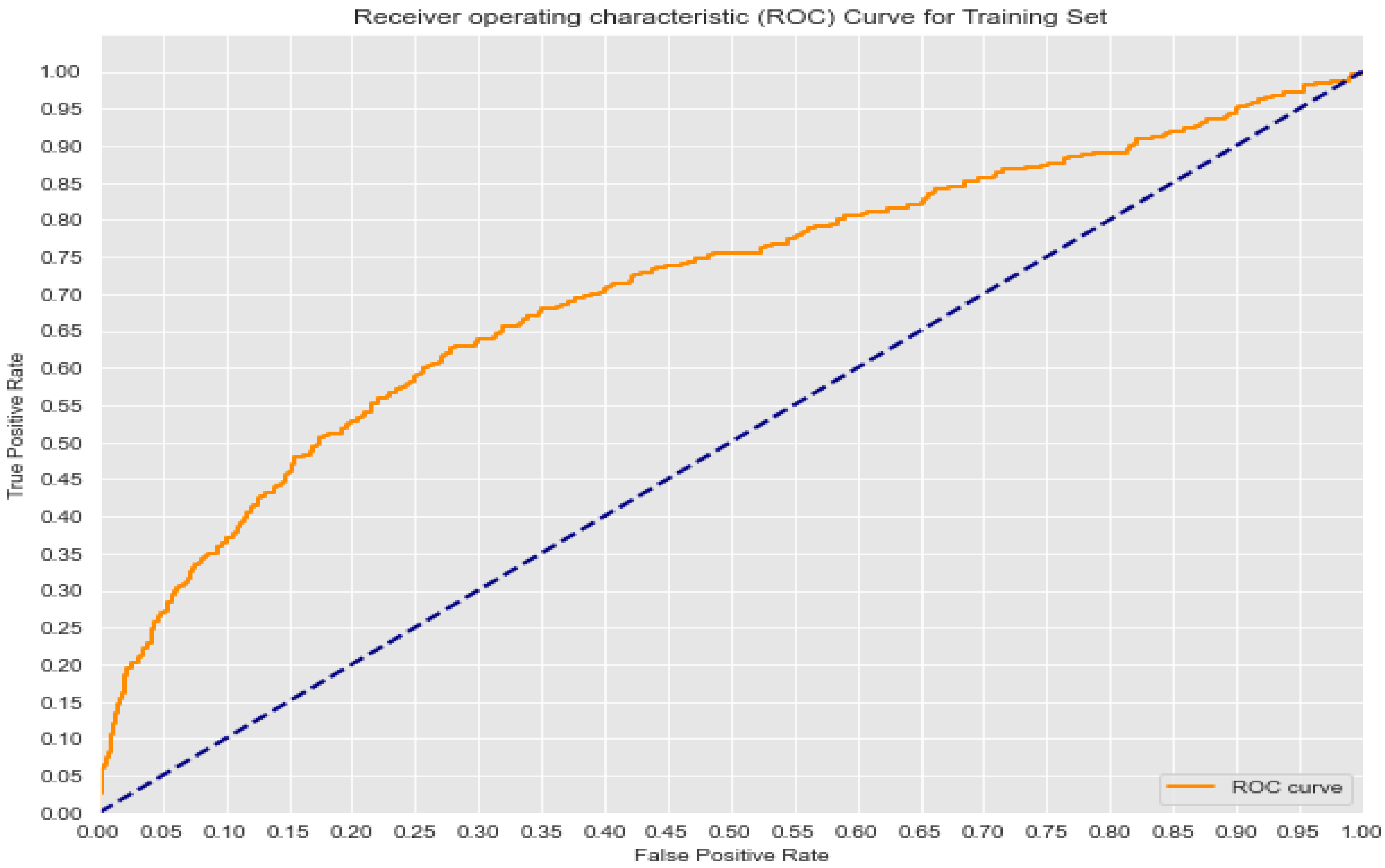
- Develop a classification model for churn prediction in order to guide customer retention strategies.

Data Understanding and Strategy.

- The data comprises 3,333 records and 20 features. The features comprise static data of state ,area code phone number. The other features a dynamic data on number of calls, minutes and charges recorded during the daytime, evening, night and international calls.
- We have conducted two classification models on the data . These are: Logistic regression and Decision Tree.These models are evaluated based on their AUC score and accuracy scores.

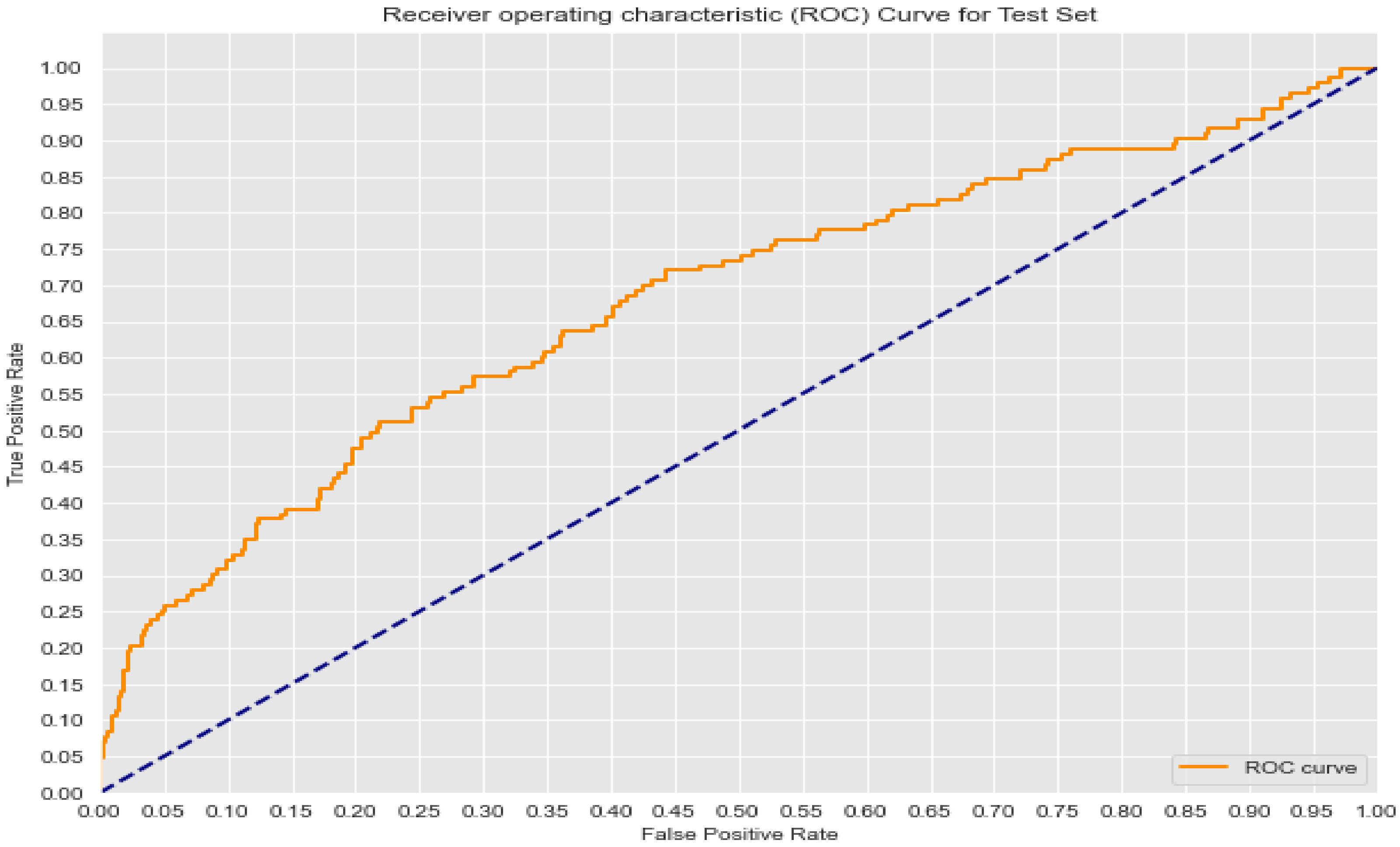
LOGISTIC REGRESSION MODELS.

Logistic Regression Train Data Scores.



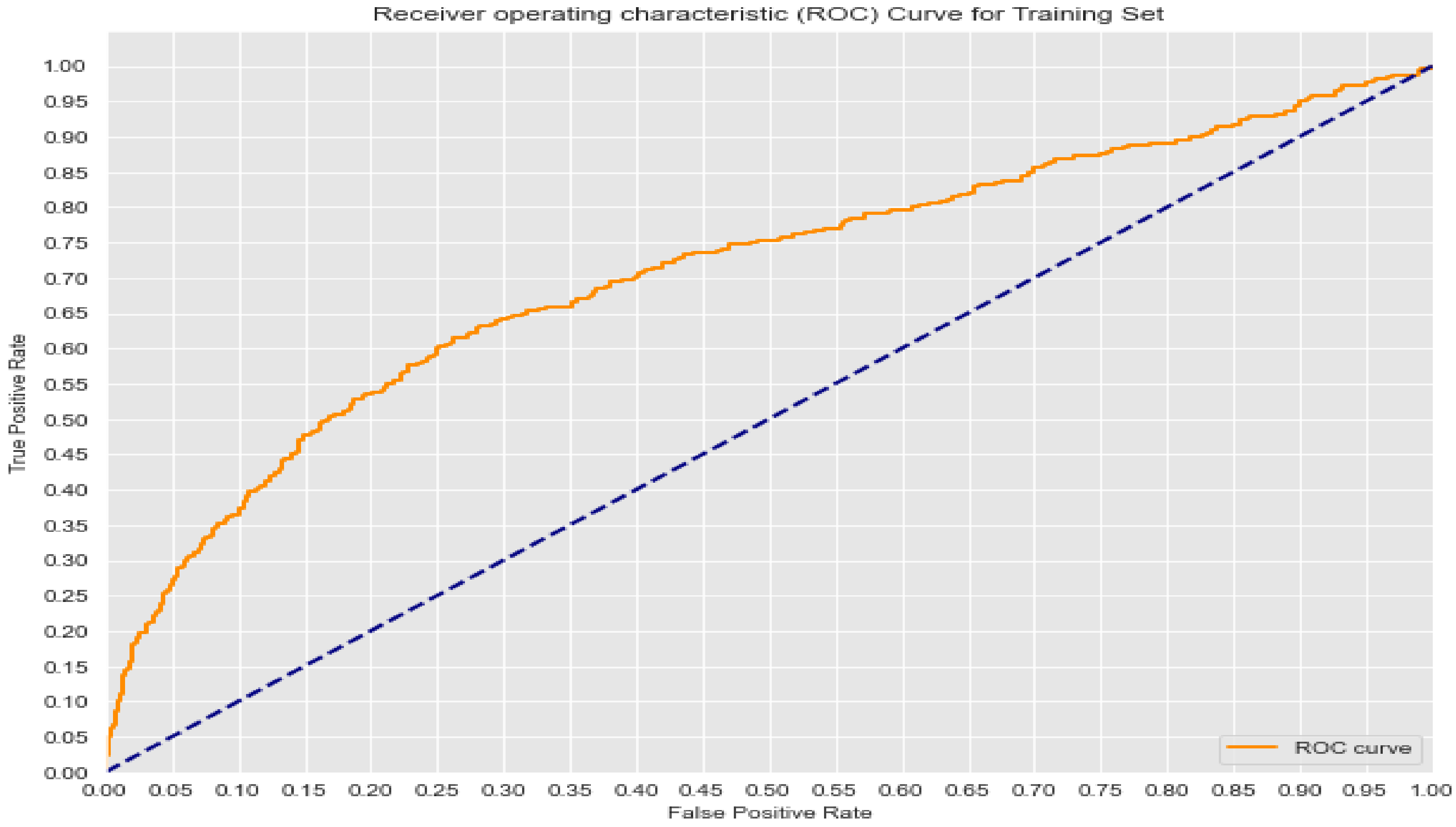
- The logistic regression train data set has the following scores:
- AUC- 0.707
- Accuracy score – 0.861

Logistic Regression Test Data Scores.



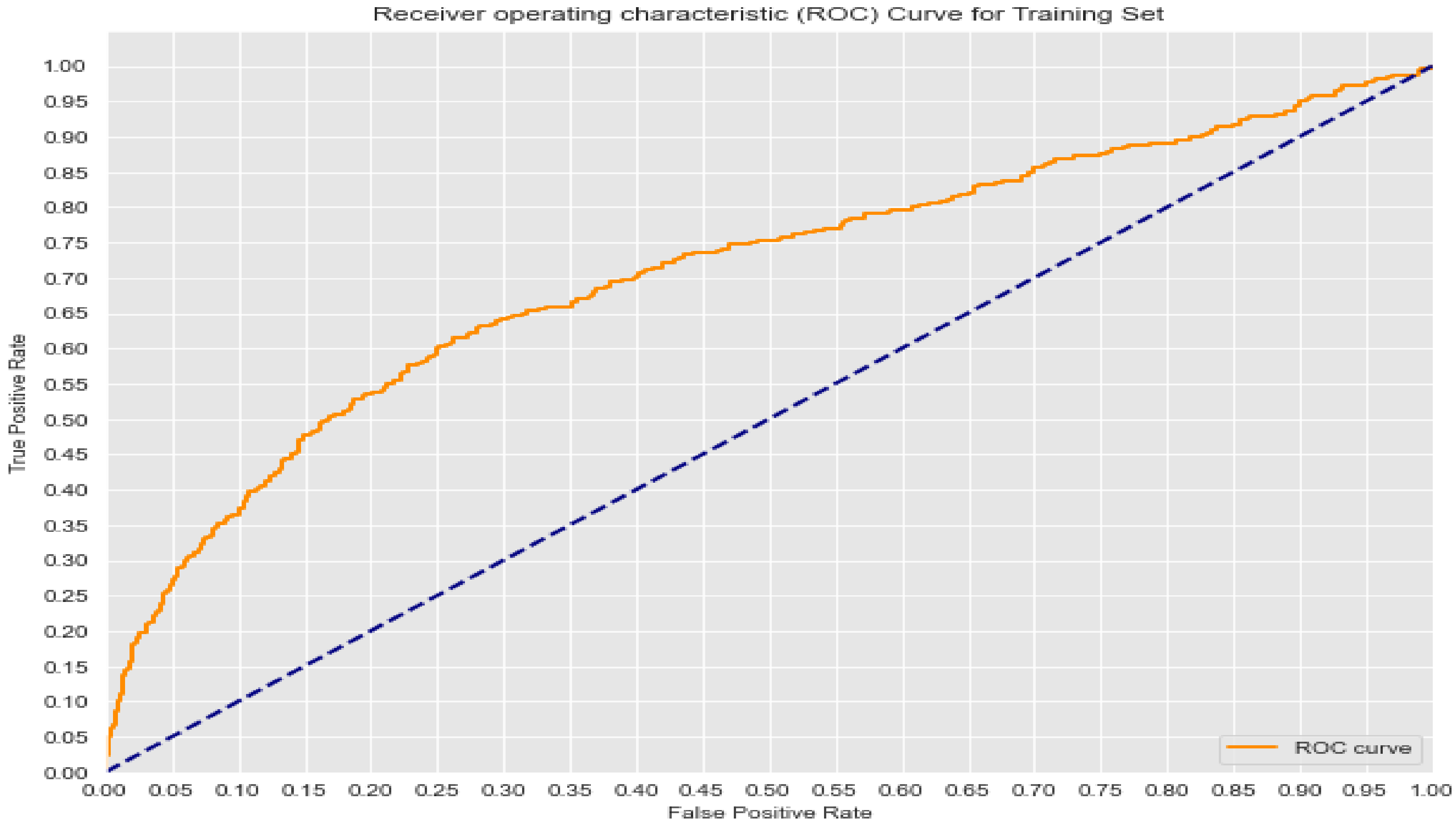
- The logistic regression test data set has the following scores:
- AUC- 0.678
- Accuracy score – 0.862

Logistic Regression Class weighted Train Data Scores.



- The logistic regression test data set has the following scores:
- AUC- 0.706
- Accuracy score – 0.753

Logistic Regression Class weighted Test Data Scores.



- The logistic regression test data set has the following scores:
- AUC- 0.678
- Accuracy score – 0.738

Logistic Regression Model: Baseline vs Class weighted.

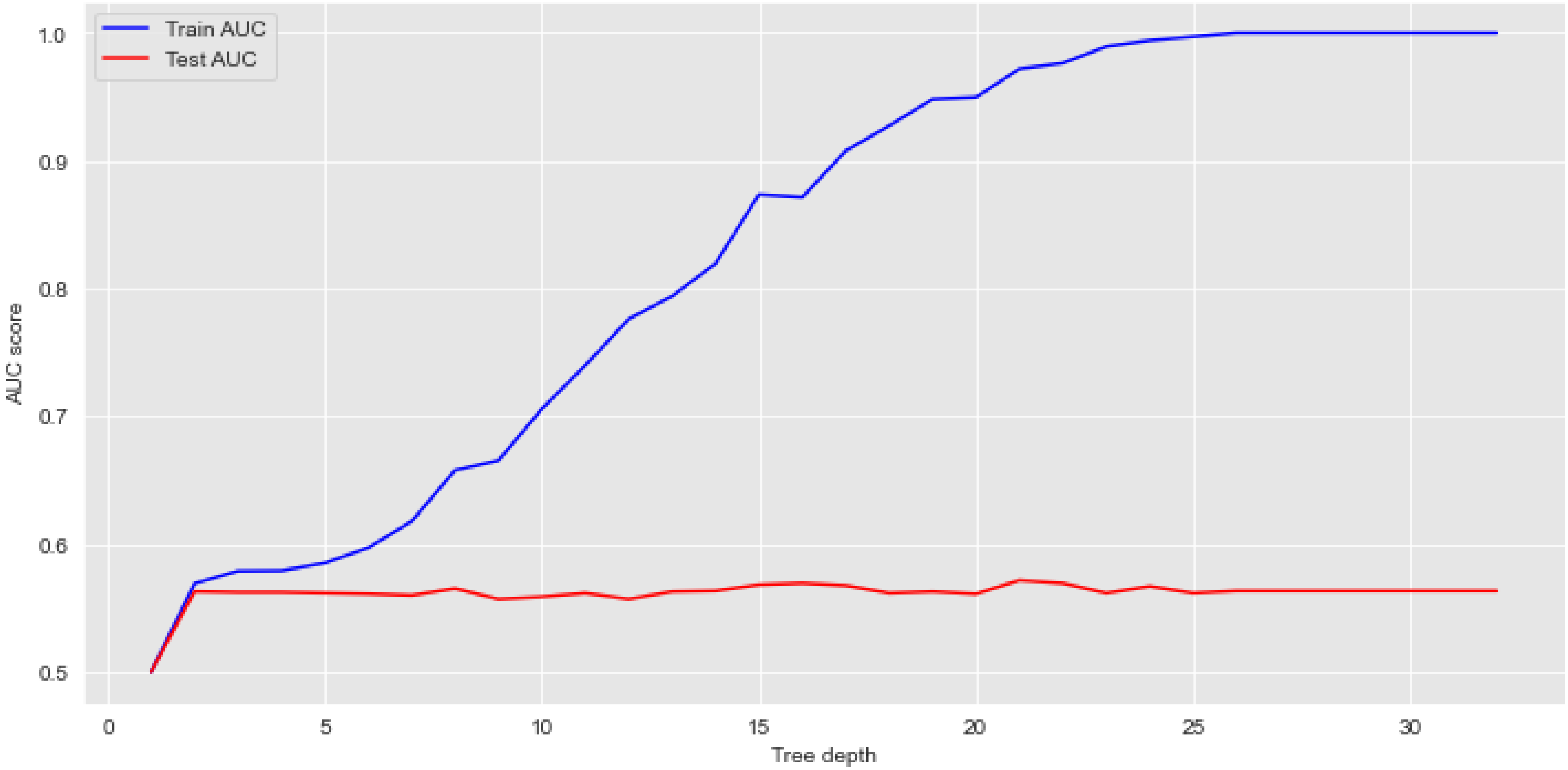
- The initial train and test logistic regressions has an almost identical accuracy score of 0.86 and AUC train and test score of 0.707 and 0.678 respectively. After class weighting to mitigate class imbalance the train and test accuracy scores reduced to 0.75 and 0.73 respectively. The train and test AUC scores , however, changed slightly to 0.706 and 0.678.
- Based on the data we are unable to conclude on whether the initial model or weighted model is more suitable because the accuracy scores have reduced but the AUC scores are unchanged.
- The performance of the logistic model has not improved hence we cannot adopt it for prediction.

DECISION TREE MODELS.

Decision Tree Initial Model Train & Test Data Scores.

- The decision tree train accuracy score is 1 and the test score is 0.78 .These suggests overfitting because the model has performed really well on the train data but a bit poorly on the test data.
- We would need to optimize the decision tree to determine if it would predict suitably.

Decision Tree Pruned Model Train & Test Data Scores.



- The decision tree model is pruned through limiting the tree depth.
- The optimal tree depth was determined as 2 being the point beyond which test AUC does not increase.
- After pruning our train accuracy score is more realistic at 0.876 versus an initial score of 1. Our model is not overfitting.

Decision Tree Initial Model vs Pruned Model: Train & Test Data Scores.

- We now have a trained accuracy score of 0.876 and a test accuracy score of 0.874 . This shows the model is not overfitted nor is it underfitted. It is at an optimal level to make predictions.

Recommendation.

1. The model to be considered to predict customer churn would be the Pruned decision tree which has a minimal difference between its train and test accuracy scores.

2. The pruned decision tree model's accuracy scores are higher than the logistic regression model.