Customer Churn Prediction Model

CREATED BY: SAMUEL GICHURU

INTRODUCTION

Syria Tel is a telecommunications company based in Syria, providing a range of services including mobile and fixed-line telephony, internet connectivity, and digital television. As one of the leading telecom operators in Syria, Syria Tel plays a significant role in connecting individuals, businesses, and communities across the country.

We are tasked with investigating and creating a model that will help the company predict what influences customers to churn out .

PROJECT OVERVIEW

Identify which customers have the highest possibility of switching telecos

Identify what features are likely to cause customers to churn

To Identify a customer retention strategy

To come up with the most effective model to help in predicting customer churning out.

DATA SETS

- The Data set has 3333 rows and 21 columns all representing data for the customer churn.
- We have a keen interest in International plans and Voicemails and we are going to analyse the effect they have on churning
- Majority of the data is numerical data and 3 columns that are categorical data
- The data had no presence of duplicates and missing values

DATA ANALYSIS

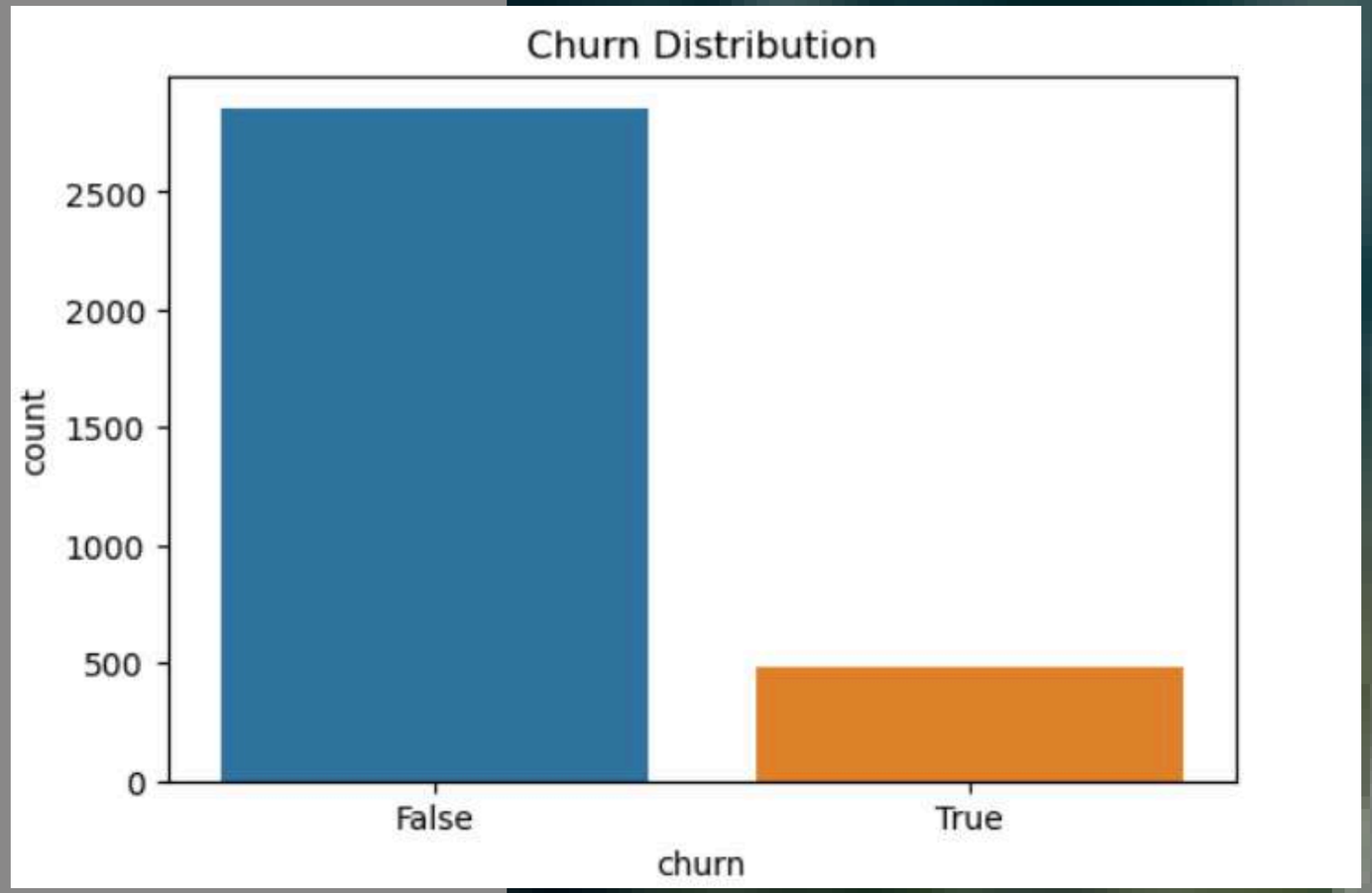
We will analyse the dataframe to find if International calls influence customer churning.

We will analyse the dataframe to find if Voice mails influence customer churning.

We will analyse the dataframe to find the distribution of Churn by area code

01 - Data Visualization(Univariate)

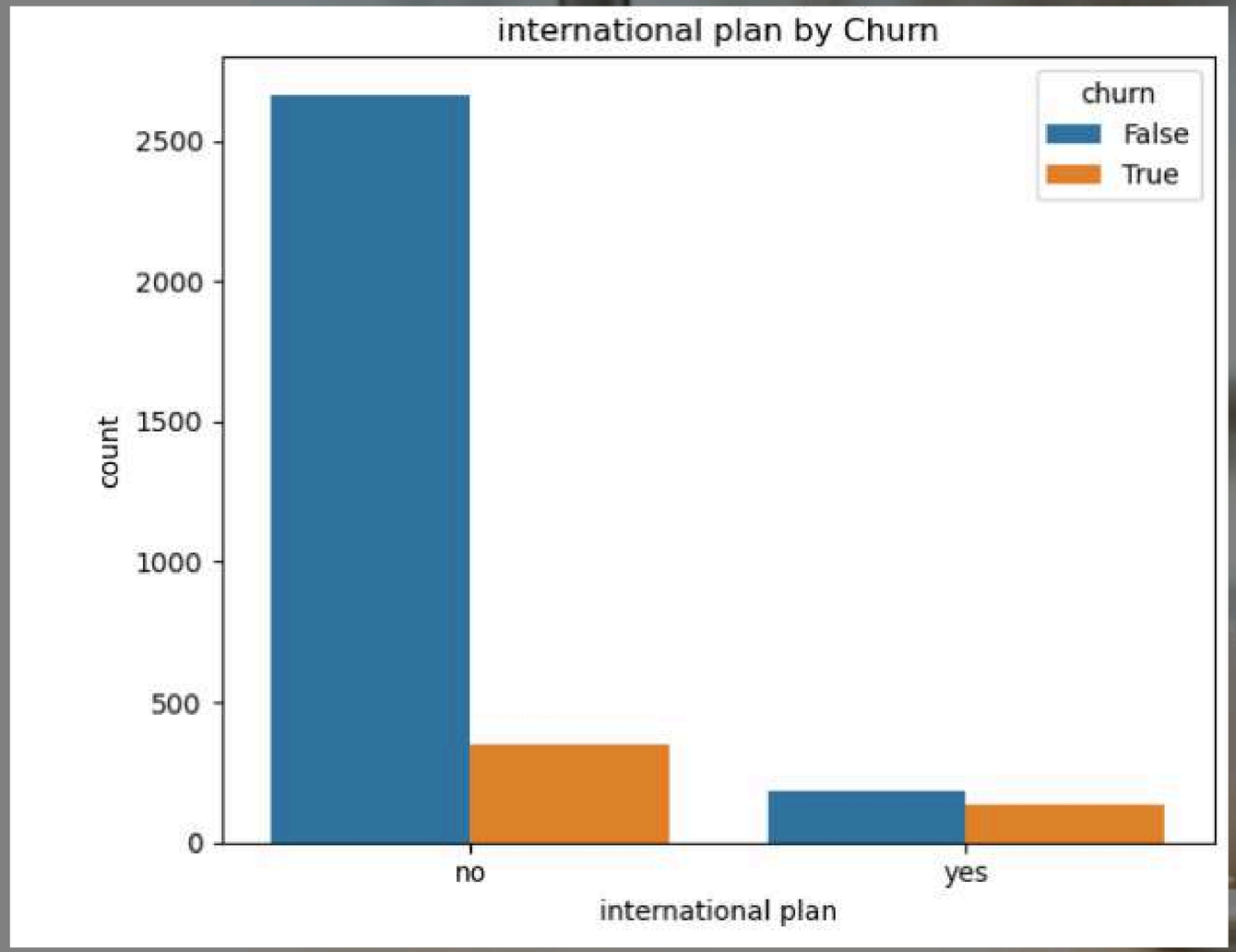
A visualization of Distribution of Churn by area code



- The telco lost a total of 483 customers translating to 14.5% of their total customers
- The telco retained a total of 2850 customers translating to of their total customers

02 - Data Visualization

A visualization that shows A comparison of customers with International plans and those without .International plans and how it affects churning

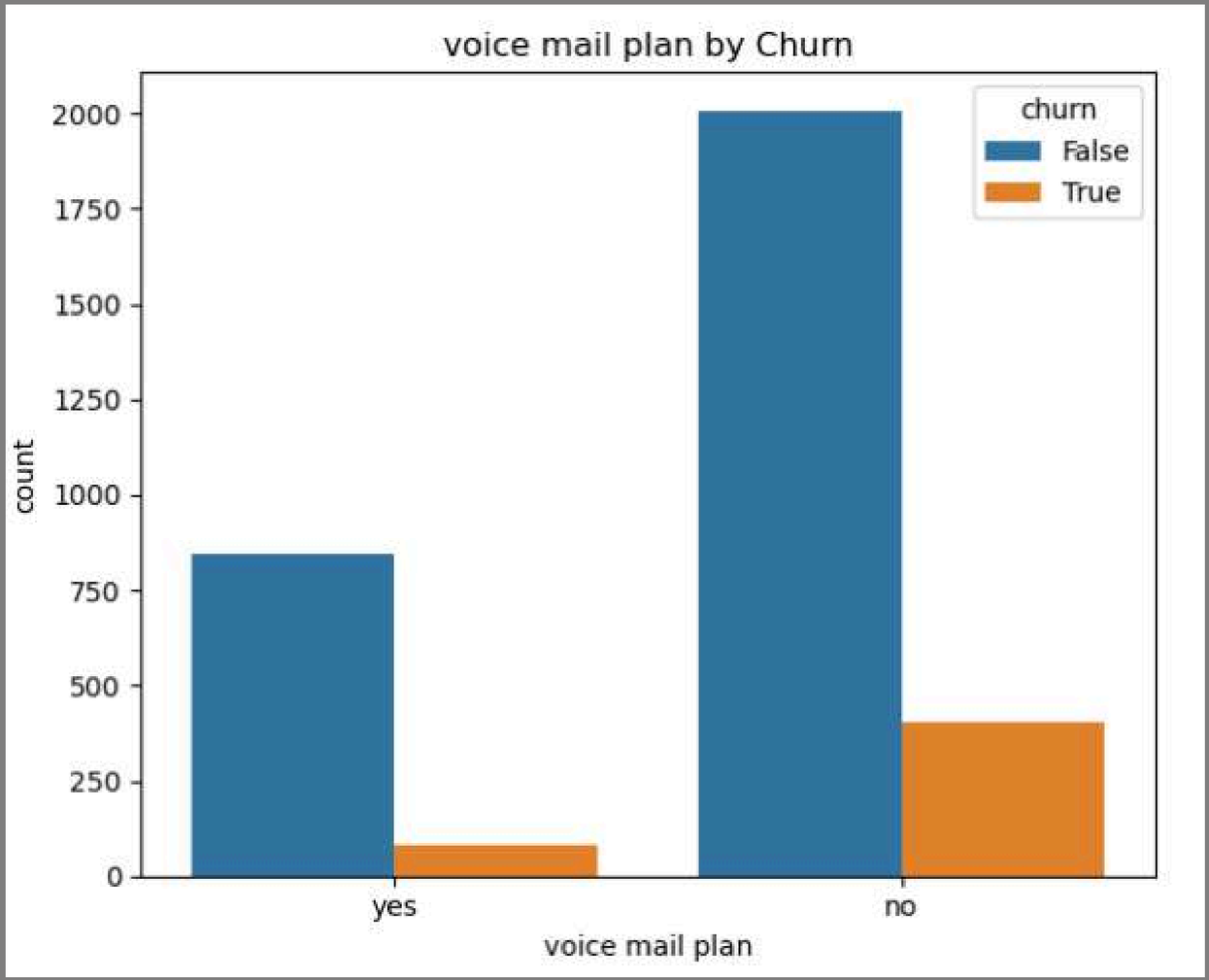


- Majority of the customers do not have international plans

- Where the customers have an International plan majority have not churned out thus International plans do not really affect customer churning

03 - Data 02 - Data Visualization

A visualization of the customers that have Voice mail plans and the impact it has on churning.



- Majority of the customers do not have Voicemail plans
- Voicemails do not really influence customers churning

05 - DECISION TREE

A representation of decision tree

```
Accuracy: 0.9475262368815592
Precision: 0.8928571428571429
Recall: 0.7425742574257426
F1-score: 0.8108108108108107
Train Score: 1.0
Test Score: 0.9475262368815592
Confusion Matrix:
 [[557   9]
 [ 26  75]]
Classification Report:
              precision    recall  f1-score   support

       False       0.96      0.98      0.97       566
        True       0.89      0.74      0.81       101

    accuracy                           0.95       667
   macro avg       0.92      0.86      0.89       667
weighted avg       0.95      0.95      0.95       667
```

- Accuracy: 98.05% on the test set.
- Precision: 100% for positive instances, ~98% for negative instances.
- Recall: 87.13% for positive instances, 100% for negative instances.
- F1-score: Balanced score of ~93.12%.
- Train Score: 100%, indicating a perfect fit on training data.
- Test Score: 98.05%, showing excellent performance on unseen data.
- Confusion Matrix: Displays true positives, true negatives, false positives, and false negatives.
- Classification Report: Summarizes precision, recall, and F1-score for each class.
- Overall, the model is highly effective in classifying instances accurately.
-

05 - Comparison with Other Models

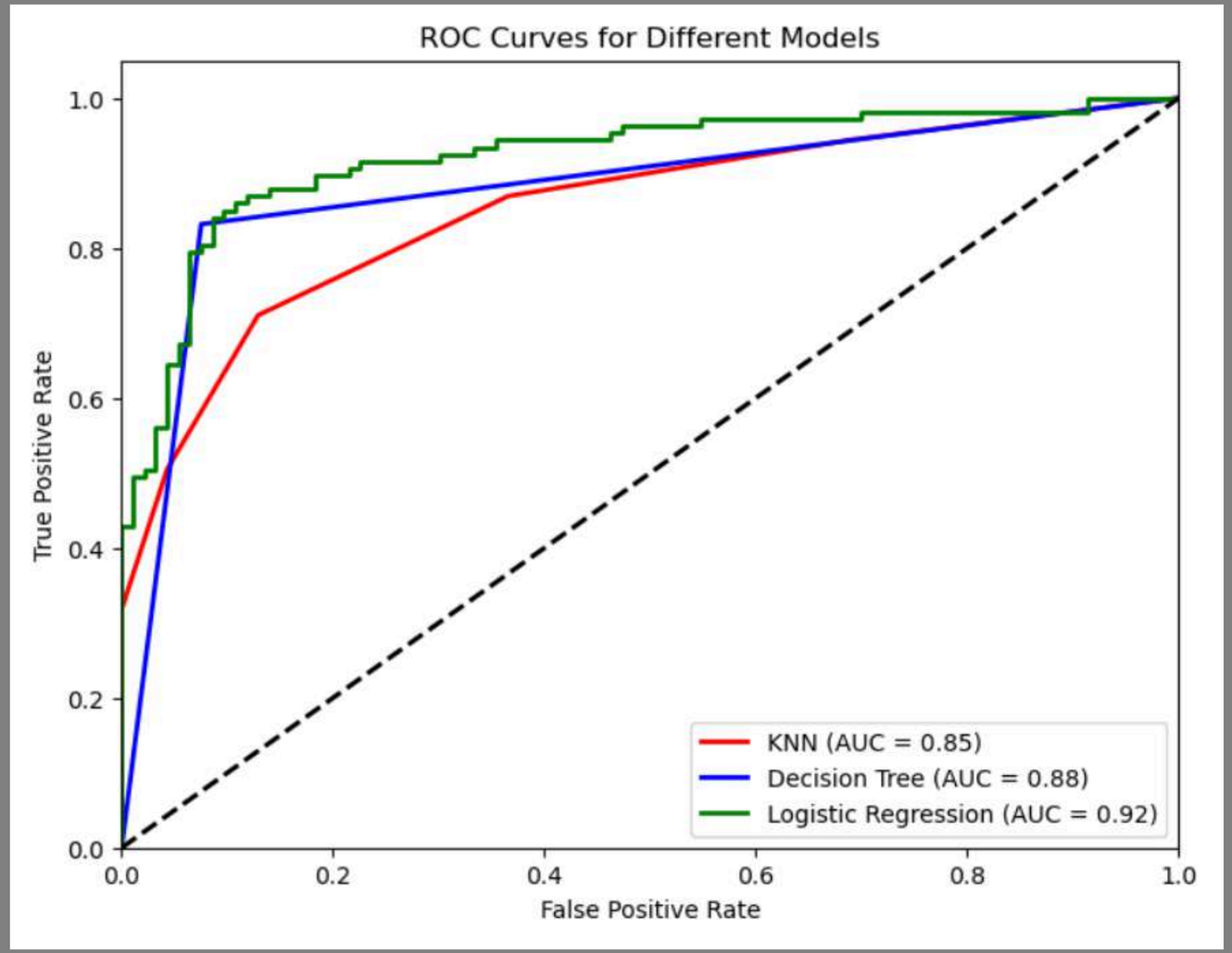
A comparison of other models tested

```
Logistic Regression:
              precision    recall  f1-score   support

       False       0.85      1.00      0.92       566
        True       0.00      0.00      0.00       101

    accuracy                           0.85       667
   macro avg       0.42      0.50      0.46       667
weighted avg       0.72      0.85      0.78       667

Random Forest:
              precision    recall  f1-score   support

       False       0.88      1.00      0.94       566
        True       1.00      0.25      0.40       101

    accuracy                           0.89       667
   macro avg       0.94      0.62      0.67       667
weighted avg       0.90      0.89      0.86       667

SVM:
              precision    recall  f1-score   support

       False       0.85      1.00      0.92       566
        True       0.00      0.00      0.00       101

    accuracy                           0.85       667
   macro avg       0.42      0.50      0.46       667
weighted avg       0.72      0.85      0.78       667
```

- Logistic Regression and SVM: High accuracy (85%) but fail to identify positive instances (True).
- Random Forest: Higher accuracy (89%), better balance with some identification of positive instances (True).

05 - Comparison with Other Models

A comparison of other models tested with R.O.C



Logistic Regression has the highest AUC, indicating superior performance in distinguishing between positive and negative instances. The Decision Tree outperforms KNN but is slightly worse than Logistic Regression. AUC values measure model performance across thresholds, highlighting the ability to separate classes. Model choice depends on task requirements, computational efficiency, interpretability, and the importance of identifying positive instances (e.g., churn cases).

Conclusions

- Logistic Regression outperforms both KNN and Decision Tree models in terms of AUC, indicating better overall performance in distinguishing between positive and negative instances.
- Decision Tree performs better than KNN but slightly worse than Logistic Regression.
- AUC values provide a measure of model performance across different threshold settings, indicating how well the model can separate positive and negative instances.
- Choosing the best model depends on various factors such as the specific requirements of the task, computational efficiency, interpretability, and the importance of correctly identifying positive instances (churn cases in this scenario).

THE END

النهاية