REDUCING CUSTOMER CHURN: USING MACHINE LEARNING TO PREDICT CUSTOMER RETENTION AT SYRIATEL MOBILE TELECOM.

- Student Name: Eugene Asengi Ukiru
- Student pace: Hybrid
- Instructor name: Maryann Mwikali

INTRODUCTION.

- Business growth and development are one of the main motivators of organizational decision-making and policy-making. Every business leader wants to increase sales, clientele, and profitability, but they also have to do everything in their power to prevent losses. In recent years, business experts and leaders have recognized customer satisfaction as a critical component in guaranteeing such growth and development. A business cannot record any cash inflows in terms of revenues, make any sales, or turn a profit if it does not have customers. This emphasizes how important it is for businesses to put policies in place that keep their current clientele.
- Recent technological developments have also exacerbated business rivalry. When combined with increased market saturation, this competition means that it has become harder and more expensive for businesses in most sectors to acquire new clients, which means they must shift their focus to cementing relationships with existing customers.
- Through this project, I am building a prediction model that identifies customer churning patterns, which can help develop mitigation strategies. The project is structured as follows:

1. Business Understanding.
2. Data Understanding.
3. Data preparation.
4. Exploratory Data Analysis.
5. Modelling.
6. Model Evaluation.
7. Recommendations and Conclusions.

BUSINESS UNDERSTANDING.

- Syriatel Mobile Telecom has emphasized the need to increase customer satisfaction and maintain its 8 million clientele in light of a growing combination of factors in the telecommunications markets, including competition, technological advancements, and globalization.
- The multinational telecom company from Syria restates its resolve to hold onto its market share by building "its reputation by focusing on customer satisfaction and social responsibility." Even though these initiatives have been successful over the years, the business must be more dedicated to lowering customer attrition rates as they could jeopardize its position in the market, financial success, and expansion as a whole. The company can cut expenses, prevent losses, and boost sales by keeping its 8 million customers.

- Primary stakeholder:

❖ Syriatel Mobile Telecom

- Other stakeholders:

❖ Shareholders
❖ Employees
❖ Customers

Research Objectives:

1.To determine the most suitable model to predict customer churn.
2.To establish a customer retention strategy to reduce churn.

Research questions:

1.Which machine learning model is the most suitable for predicting customer churn?
2.What strategies can Syriatel Mobile Telecom implement to retain customers and reduce churn rates?

DATA UNDERSTANDING.

- The Churn in Telecom's dataset from Kaggle contains information about customer activity and whether or not they cancelled their subscription with the Telecom firm. The goal of this dataset is to develop predictive models that can help the telecom business reduce the amount of money lost due to customers who don't stick around for very long.
- The dataset contains 3333 entries and 21 columns, including information about the state, account length, area code, phone number, international plan, voice mail plan, number of voice mail messages, total day minutes, total day calls, total day charge, total evening minutes, total evening calls, total evening charge, total night minutes, total night calls, total night charge, total international minutes, total international calls, total international charge, customer service calls and churn.

DATA PREPARATION.

The analysis performed on the dataset included the following steps:

•Data Cleaning: The dataset was checked per column to find missing values, duplicates, and outliers and dealt with them accordingly.

•Data Transformation: Categorical data in the churn column was converted into numerical data.

•Exploratory Data Analysis: Analysis was done on the data to check the distribution of features and the target and to identify possible correlations between features.

•Feature Engineering: Certain columns were transformed to enhance their usefulness. This included; Encoding categorical variables into numerical representations e.g area code, international plan, voice_mail_plan Normalization and Scaling features to a consistent range using the Standard Scaler

By performing these steps, I aimed to gain a comprehensive understanding of the dataset and prepare it for further analysis and modelling.

MODELLING.

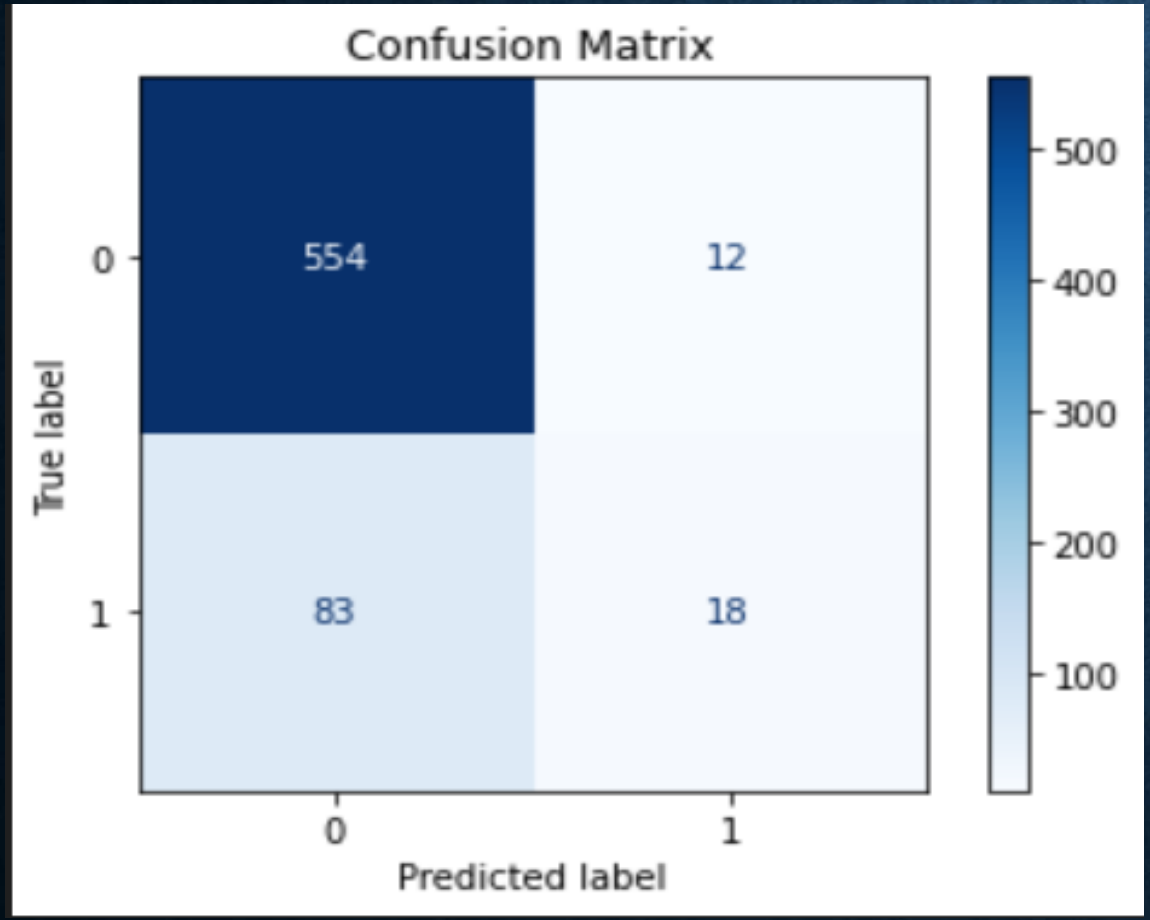
- During the analysis, I developed and evaluated two classification models to gain insights and make predictions.

1. **Model 1: Logistic regression model.**
- Baseline Model The first logistic regression model was built by default parameters. The model accrued 86% training accuracy and 86% as well for test accuracy. The recall was 18%, precision 60% and 27% F1 score.

2. **Model 2: Decision Tree Classifier.**
- Baseline Model The baseline model used the default parameters to train and test the model. The model has an accuracy of 92%, precision of 72%, recall of 75% and F1 score of 73%

VISUALISATIONS

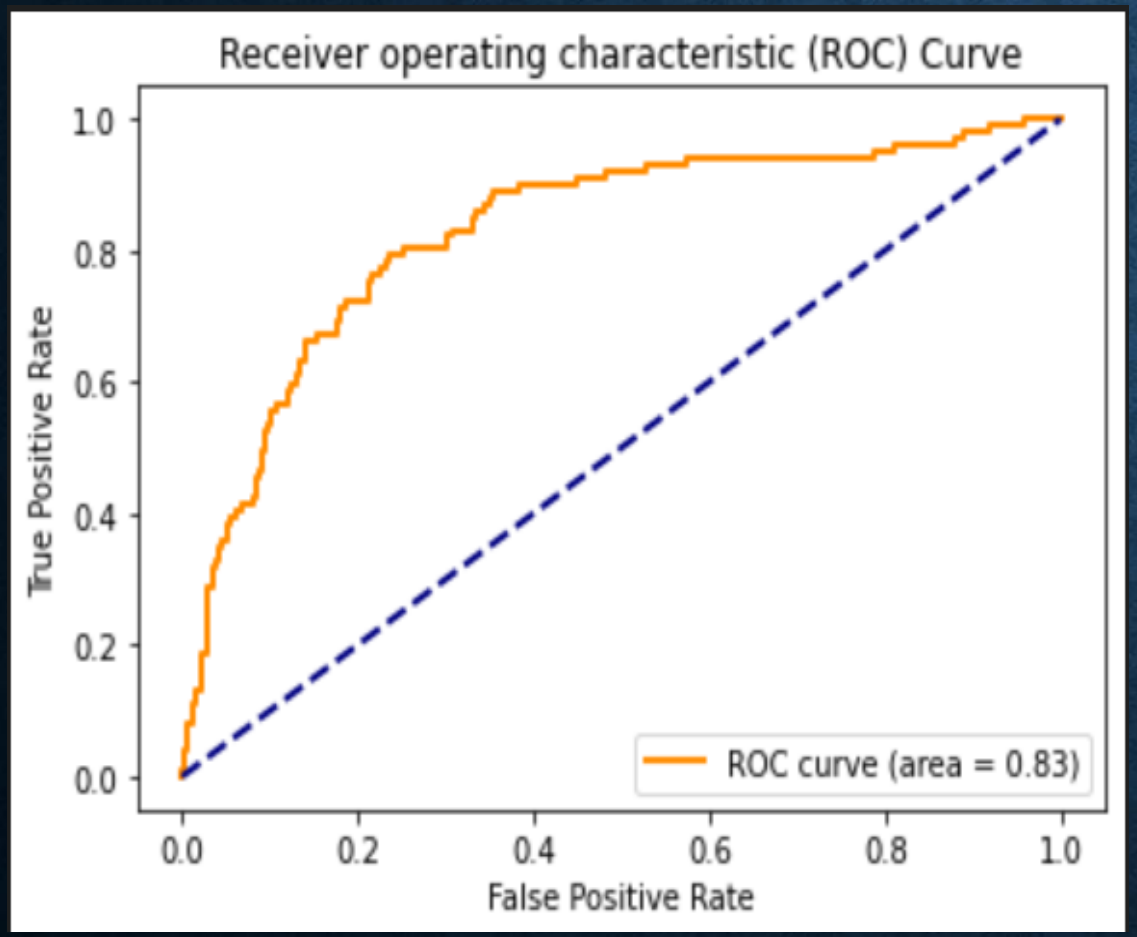
LOGISTIC REGRESSION CONFUSION MATRIX



CONFUSION MATRIX:

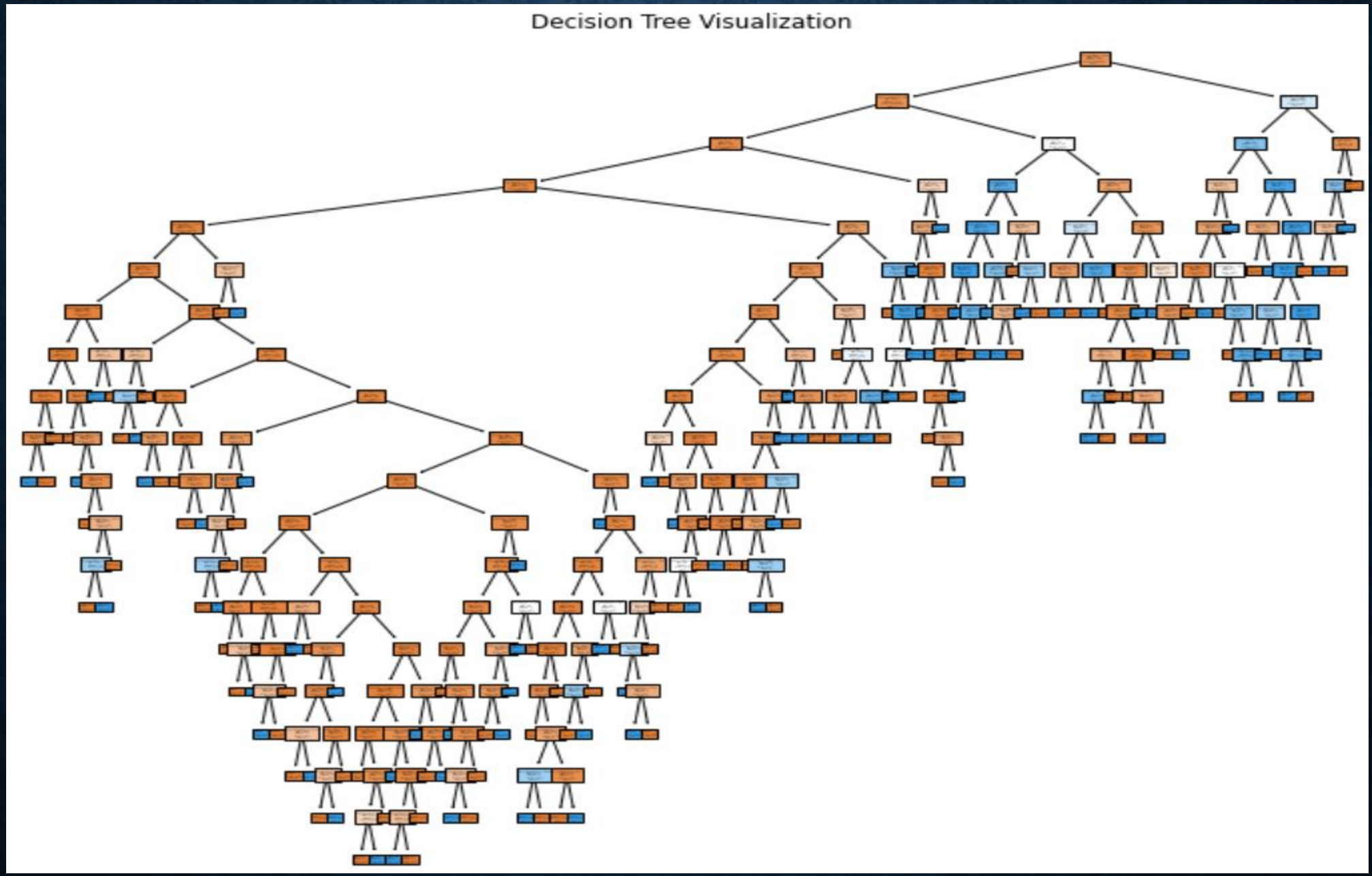
- The matrix shows a total of 667 samples in the test set.
- True positives (TP): The model correctly predicted 18 samples as not churned (class 0).
- True negatives (TN): The model correctly predicted 554 samples as churned (class 1).
- False positives (FP): The model incorrectly predicted 12 samples as churned when they were not churned.
- False negatives (FN): The model incorrectly predicted 83 samples as not churned when they were churned.

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

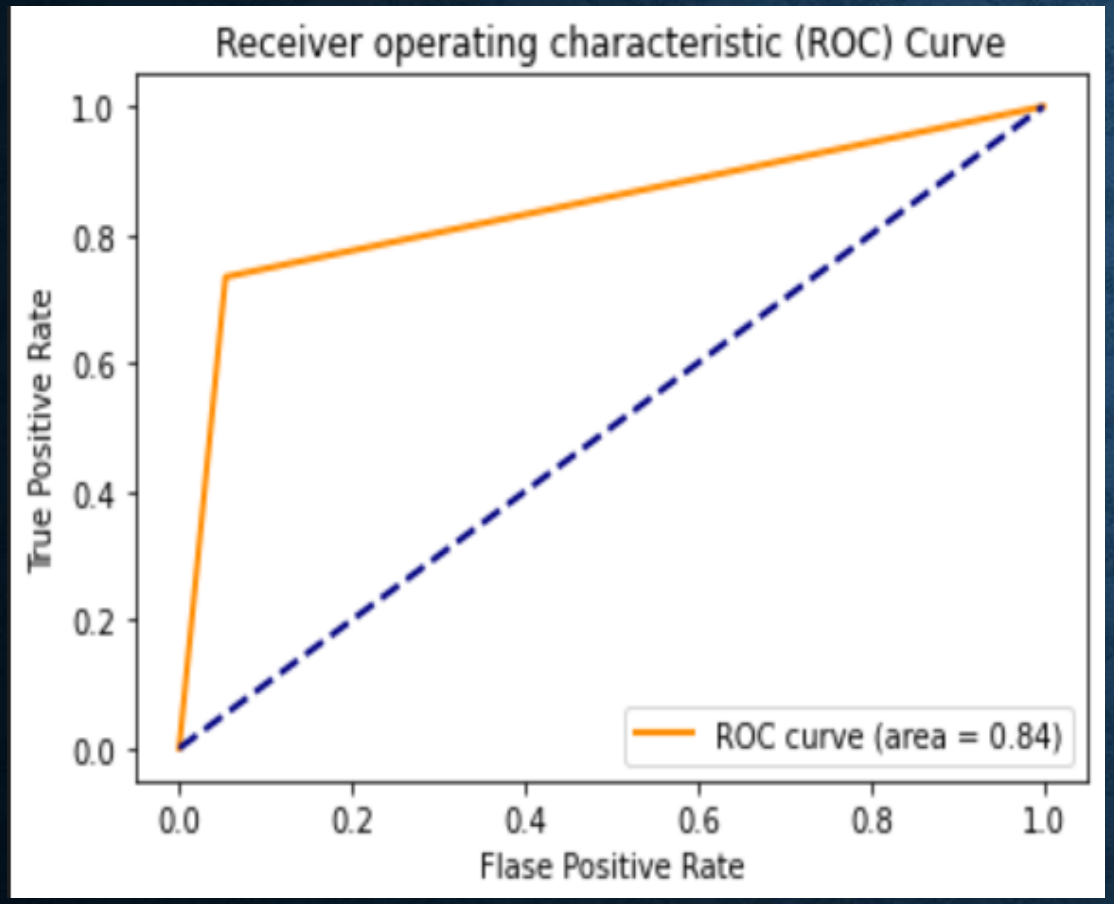


- They provide a measure of how well the model can distinguish between positive and negative samples. A model with an AUC of 1 is perfect, while an AUC of 0.5 indicates that the model is no better than random guessing.
- Auc = 0.5: the model's performance is equivalent to random guessing, and it is not useful for classification.
- Auc > 0.5: the model performs better than random guessing, and the higher the AUC, the better the model's discriminatory power.
- Auc = 1: the model perfectly distinguishes between positive and negative samples, making it an excellent classifier.
- Auc is 0.83, which is greater than 0.5 and closer to 1. An AUC of 0.83 suggests that the model has a good ability to rank the predictions, and it performs significantly better than random guessing.

DECISION TREE CLASSIFIER



DECISION TREE ROC CURVE



- The ROC curves for Logistic regression and Decision Tree were analyzed. The Decision Tree model outperformed the Logistic regression model in terms of AUC-ROC score, making it the most effective model for the given task.
- Based on the provided metrics, the Decision Trees Classifier achieved the highest accuracy of 92% and F1 score 95%. The Logistic regression model achieved an accuracy of 86% and F1 score of 92%.
- Therefore, the Decision Trees Classifier is the best model for this task.

CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the analysis suggests that we can accurately predict customer churn using a machine learning model, with the Decision Tree classifier being the best-recommended model due to its strong overall performance.

1. I would recommend that Syriatel makes use of the Decision Tree classifier as the primary model for predicting customer churn. This model has been shown to be the most accurate in predicting churn, with an accuracy of 92%. The model has a higher ROC curve and strong overall performance in terms of accuracy.
2. I would recommend that Syriatel come up with strategies to reduce customer service calls. This increases the number of customers who are likely to churn. This can be done by identifying the most important features.