Project Title: Syriatel Customer Churn Analysis

Author: Mercy Chebet

Project Overview

- Used machine learning classification to predict customer churn.
- Dataset: 3,333 records, 20 features.
- Churners: 14.49%, Non-churners: 85.51%.
- Evaluation metric: **Recall**.

Business Understanding

- Customer churn affects telecom revenue.
- Retaining customers is cheaper than acquiring new ones.
- Goal: Predict churn and understand key drivers.

Research Objectives

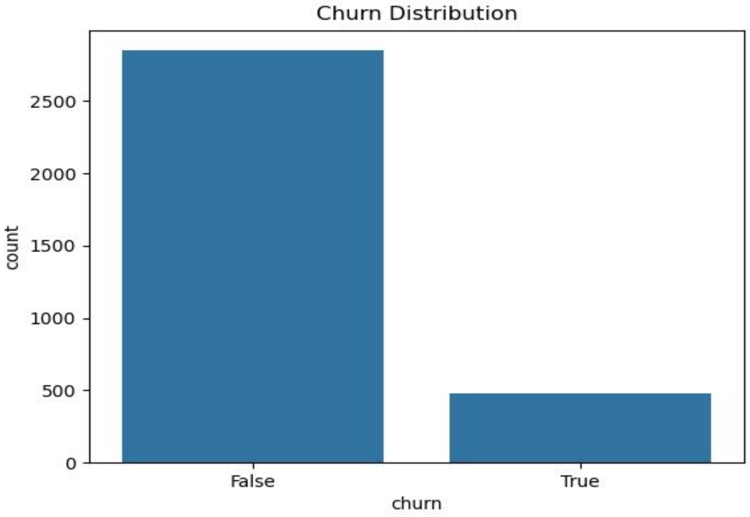
- Build an accurate model to predict churn.
- Identify important features linked to churn.

Data Understanding

- Data types: Categorical & continuous.
- No missing or duplicate records.
- Converted area code to categorical.

Target Variable Distribution

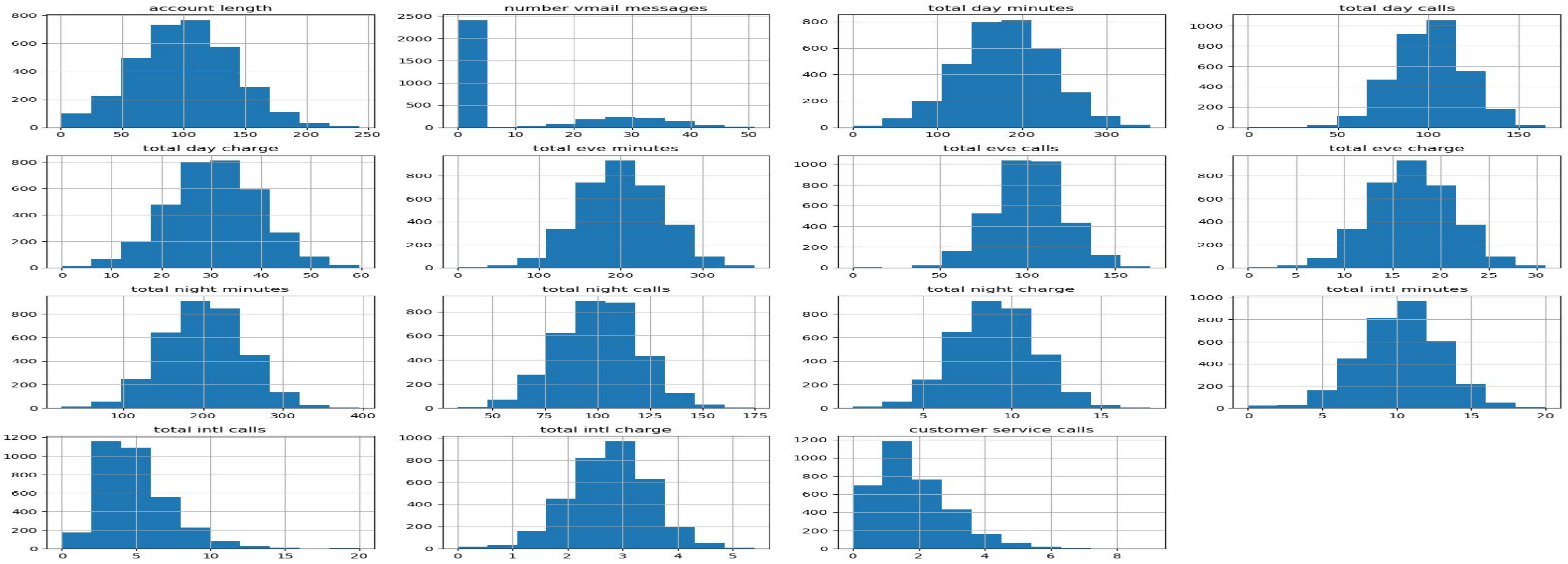
Bar chart of churn distribution



- Churn = 14.49%, Non-churn = 85.51%.
- Imbalanced dataset.

Feature Distributions

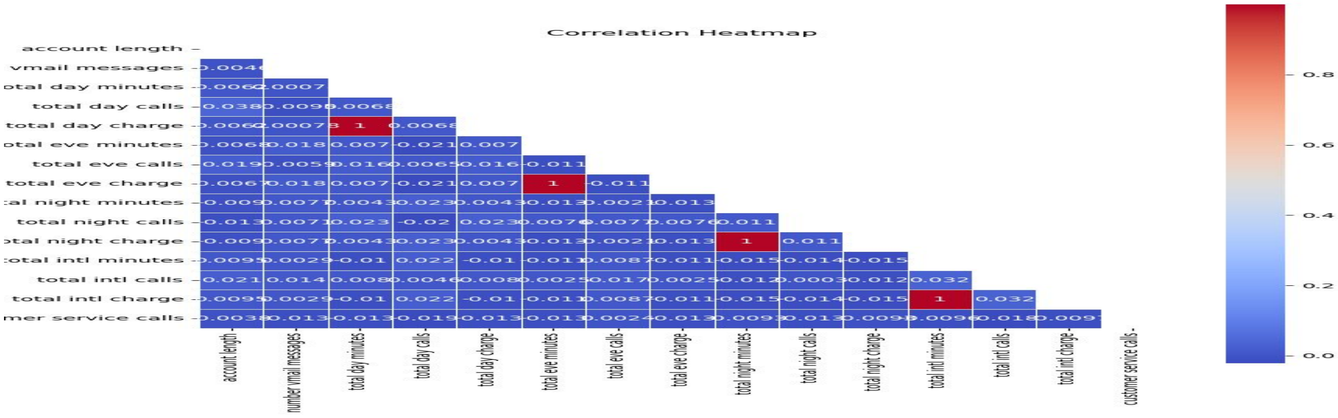
Histogram grid of numeric features



- Numeric features have varied scales.
- Some are skewed; require scaling.

Correlation Analysis

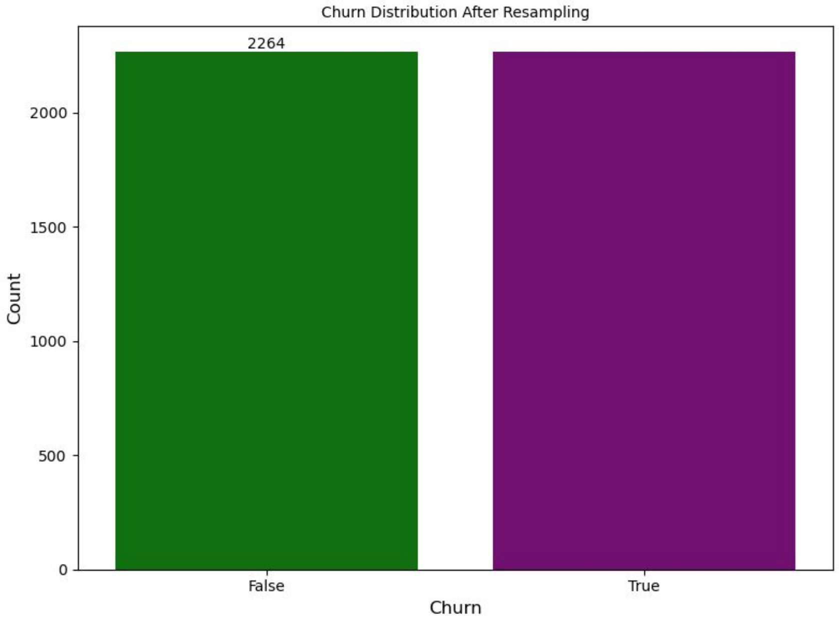
Heatmap of correlation matrix



- Strong correlations between duration and charge features.
- Weak correlation between features and churn.

Data Preparation

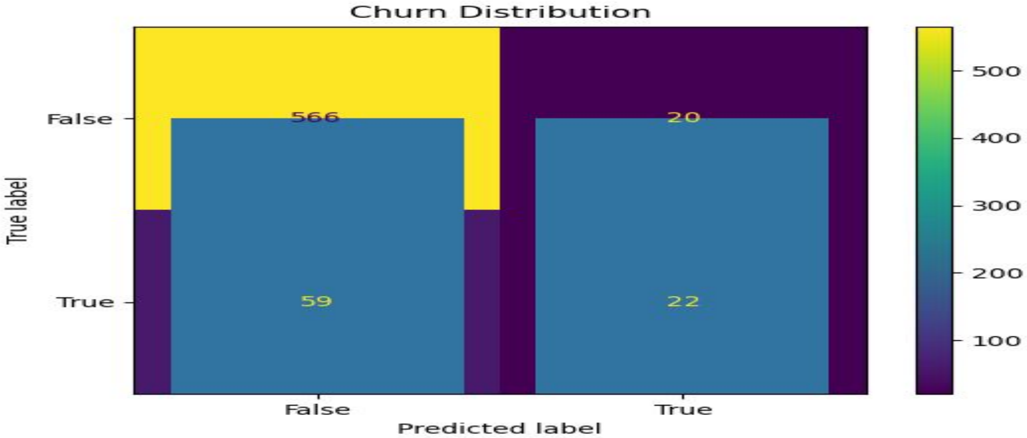
SMOTE chart showing class distribution



- Dropped highly correlated variables.
- Train-test split: 80/20.
- Applied SMOTE to balance training data

Baseline Model: Logistic Regression

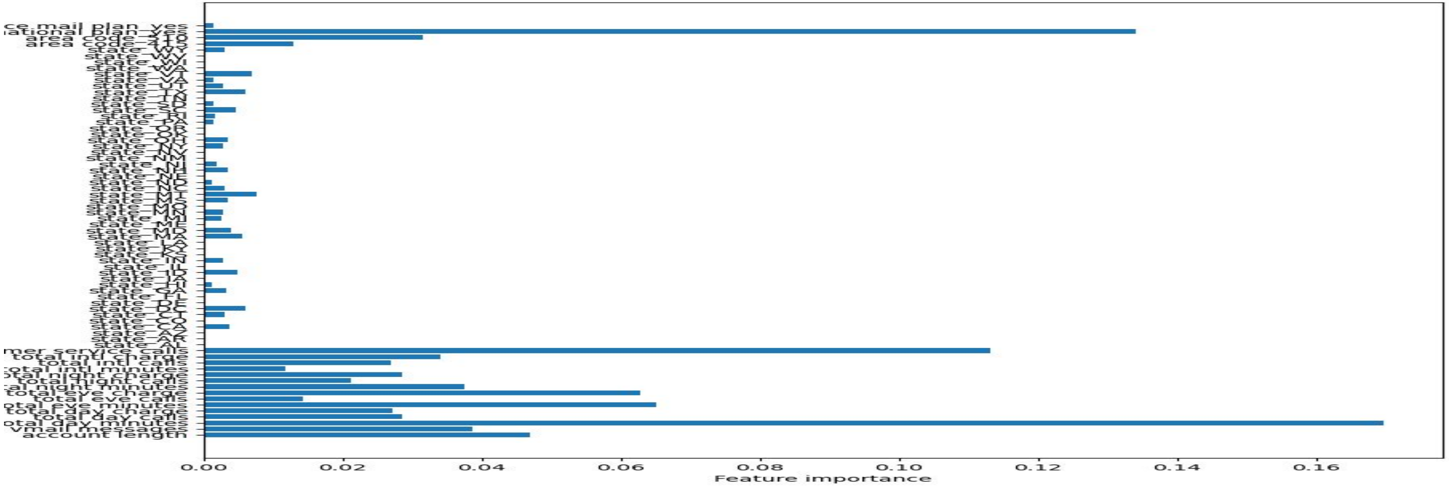
Confusion matrix plot



- Applied scaling.
- Performance:
  - High train recall, low test recall.
  - Overfitting observed.

Decision Tree Model

Feature importance plot



- Initial model had improved recall.
- Still overfitted slightly.

Feature Selection with RFECV

- Recursive Feature Elimination with Cross-Validation.
- Selected top 15 features out of 64.

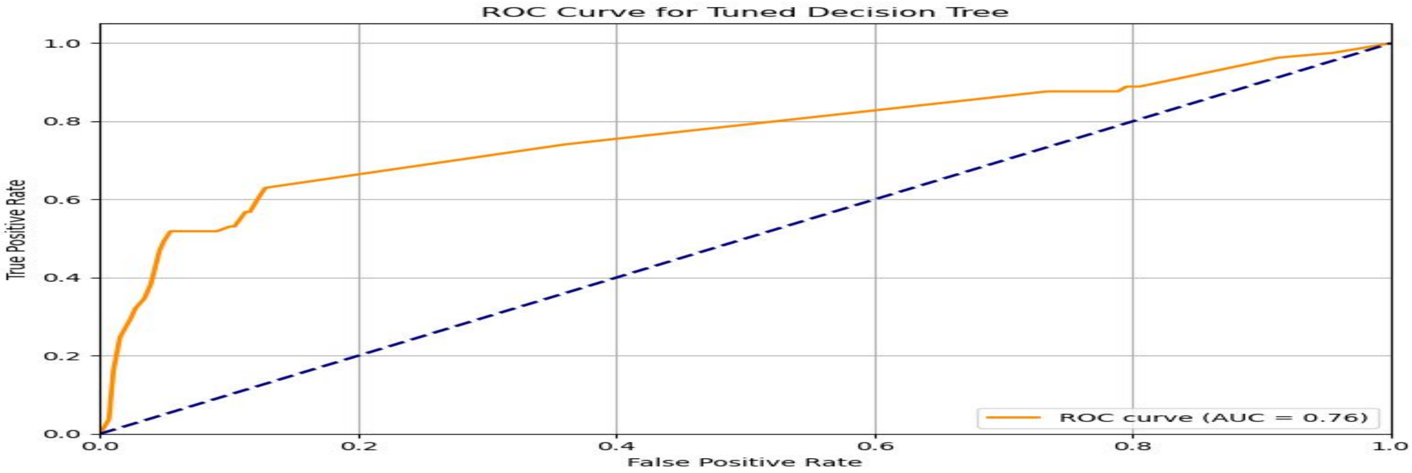
Hyperparameter Tuning

- Tuned using GridSearchCV.
- Best parameters:
  - Criterion: entropy
  - Max depth: 28
  - Min samples leaf: 2
- Performance improved, reduced overfitting.

- **Recall:** 0.65
- **Precision:** ~0.47
- **Accuracy:** ~0.86

ROC Curve

ROC Curve plot



- ROC AUC: **0.65**
- Model has moderate discriminative power.

Key Predictive Features

- Total day minutes
- Total evening minutes
- Customer service calls
- Total international minutes

Recommendations:

- Improve customer service monitoring.
- Analyze and adjust call pricing.
- Focus retention on high-duration users.

Next Steps

- Increase training data.
- Try ensemble models (e.g., XGBoost).
- Investigate external features like demographics.

Thank You!

- Contact: Mercy Chebet