Classification Model Project for Customer Churn Prediction

A Comparative Analysis of Logistic Regression and Decision Tree Models

Presenter's Name: ODHIAMBO APIYO
Date: 1ST September 2024

Overview

- This analysis is based on churning of customers based Syria Telecommunications company.

- We will use machine learning techniques to analyze telecommunications data and customer behavior patterns, with the goal of deriving actionable insights, enhancing service quality and majorly customer retention.

Introduction

- Business understanding
- In the telecommunications industry, churn is a significant concern as it directly impacts revenue and market share.
- The primary goal of churn analysis is to understand the factors that lead customers to switch to competitors or discontinue service altogether.
- By identifying patterns and behaviors that precede customer churn, companies can develop proactive strategies to enhance customer retention, improve customer satisfaction, and maintain a stable revenue stream

- **Objective**:

The primary goal for this project was to develop a model that predicts customer churn and identifies key features contributing to churn. Understanding these features helps the company proactively address potential churn and improve customer retention strategies.

- **Data Source**:

The data used was from telecom customer churn dataset containing customer demographic and usage data, including details like call minutes, charges, and service calls.

Data Preparation

- **Data Encoding**:

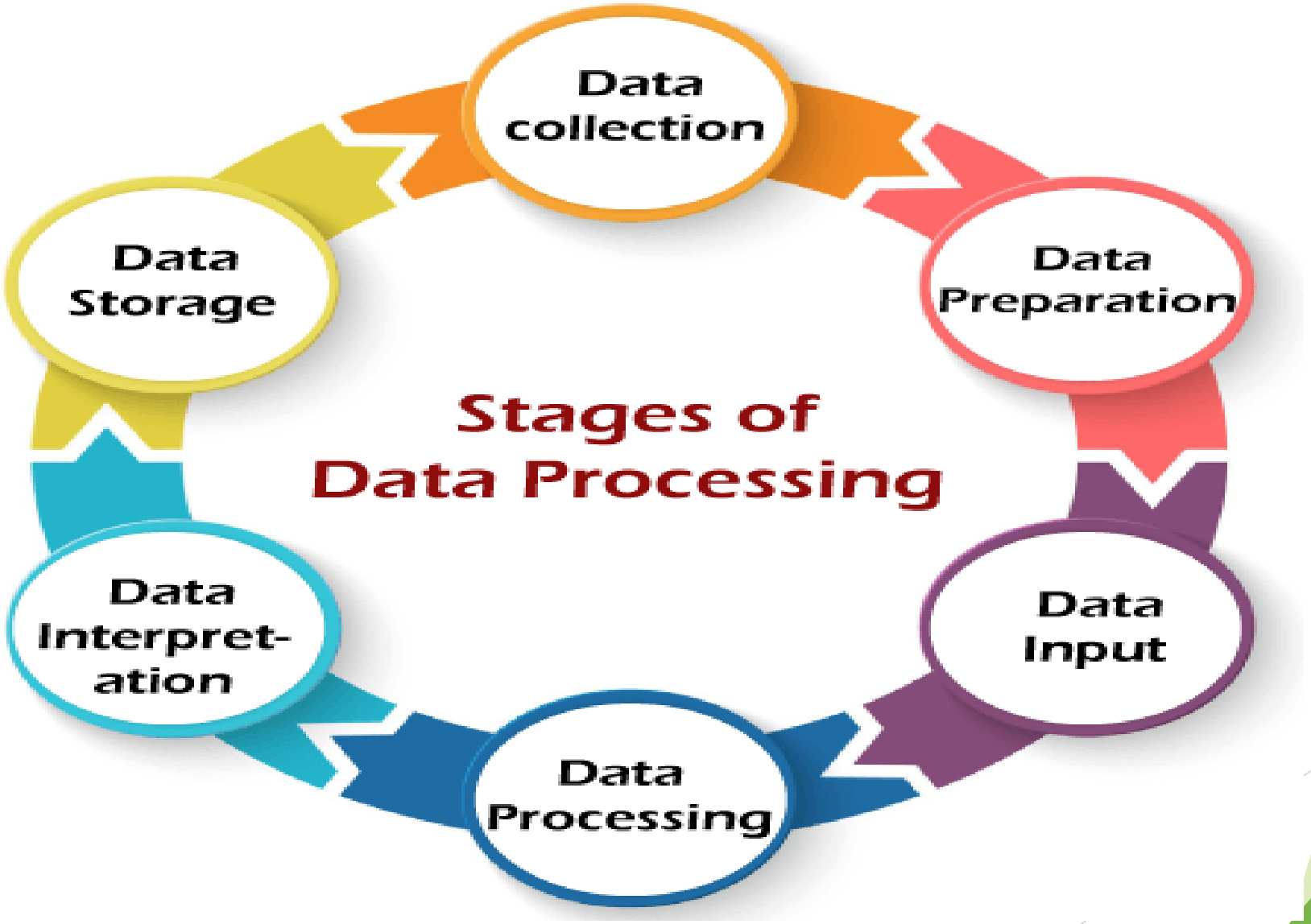
Categorical features, such as the 'International Plan', were encoded to numerical values to ensure compatibility with the models.

- **Feature Scaling**:

I applied feature scaling using StandardScaler, which standardizes features by removing the mean and scaling to unit variance. This step is crucial for algorithms like Logistic Regression that are sensitive to feature scales.

- **Train-Test Split**:

To evaluate our models effectively, we split the dataset into 80% for training and 20% for testing. This allows us to train the model on one subset and evaluate its performance on unseen data.


Data collection
Data Preparation
Data Input
Data Processing
Data Interpret-ation
Data Storage
Stages of Data Processing

Data Exploration

- The columns in the data set had no missing columns and this indicated that we were dealing with a clean dataset.
- The columns were divided into two;
    1. *Numerical columns*,
    2. *Categorical columns*.

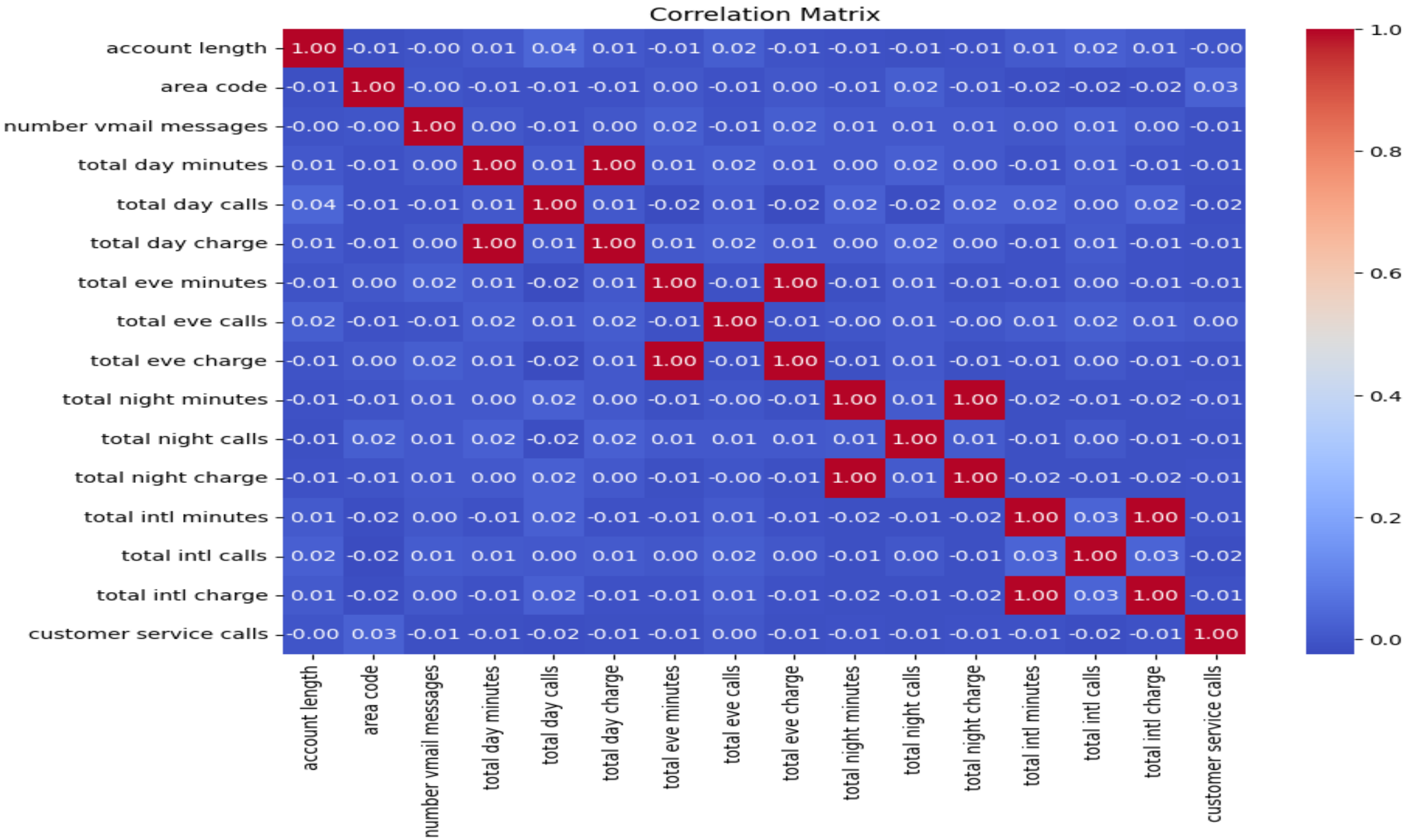
Feature Distribution

Numerical features;

The majority of the feature distribution of the numerical columns had a normal distribution. This showed that the data set had an even distribution with a small percentage of outliers

Categorical Features

More customers without a voicemail plan have churned: There is a higher frequency of churn among customers without a voicemail plan (labeled as "no" on the x-axis) compared to those with a voicemail plan (labeled as "yes" on the x-axis).


Correlation Matrix
account length 1.00 -0.01 -0.00 0.01 0.04 0.01 -0.01 0.02 -0.01 -0.01 -0.01 -0.01 0.01 0.02 0.01 -0.00
area code -0.01 1.00 -0.00 -0.01 -0.01 -0.01 0.00 -0.01 0.00 -0.01 0.02 -0.01 -0.02 -0.02 -0.02 0.03
number vmail messages -0.00 -0.00 1.00 0.00 -0.01 0.00 0.02 -0.01 0.02 0.01 0.01 0.01 0.00 0.01 0.00 -0.01
total day minutes 0.01 -0.01 0.00 1.00 0.01 1.00 0.01 0.02 0.01 0.00 0.02 0.00 -0.01 0.01 -0.01 -0.01
total day calls 0.04 -0.01 -0.01 0.01 1.00 0.01 -0.02 0.01 -0.02 0.02 -0.02 0.02 0.02 0.00 0.02 -0.02
total day charge 0.01 -0.01 0.00 1.00 0.01 1.00 0.01 0.02 0.01 0.00 0.02 0.00 -0.01 0.01 -0.01 -0.01
total eve minutes -0.01 0.00 0.02 0.01 -0.02 0.01 1.00 -0.01 1.00 -0.01 0.01 -0.01 -0.01 0.00 -0.01 -0.01
total eve calls 0.02 -0.01 -0.01 0.02 0.01 0.02 -0.01 1.00 -0.01 -0.00 0.01 -0.00 0.01 0.02 0.01 0.00
total eve charge -0.01 0.00 0.02 0.01 -0.02 0.01 1.00 -0.01 1.00 -0.01 0.01 -0.01 -0.01 0.00 -0.01 -0.01
total night minutes -0.01 -0.01 0.01 0.00 0.02 0.00 -0.01 -0.00 -0.01 1.00 0.01 1.00 -0.02 -0.01 -0.02 -0.01
total night calls -0.01 0.02 0.01 0.02 -0.02 0.02 0.01 0.01 0.01 0.01 1.00 0.01 -0.01 0.00 -0.01 -0.01
total night charge -0.01 -0.01 0.01 0.00 0.02 0.00 -0.01 -0.00 -0.01 1.00 0.01 1.00 -0.02 -0.01 -0.02 -0.01
total intl minutes 0.01 -0.02 0.00 -0.01 0.02 -0.01 -0.01 0.01 -0.01 -0.02 -0.01 -0.02 1.00 0.03 1.00 -0.01
total intl calls 0.02 -0.02 0.01 0.01 0.00 0.01 0.00 0.02 0.00 -0.01 0.00 -0.01 0.03 1.00 0.03 -0.02
total intl charge 0.01 -0.02 0.00 -0.01 0.02 -0.01 -0.01 0.01 -0.01 -0.02 -0.01 -0.02 1.00 0.03 1.00 -0.01
customer service calls -0.00 0.03 -0.01 -0.01 -0.02 -0.01 -0.01 0.00 -0.01 -0.01 -0.01 -0.01 -0.01 -0.02 -0.01 1.00
account length
area code
number vmail messages
total day minutes
total day calls
total day charge
total eve minutes
total eve calls
total eve charge
total night minutes
total night calls
total night charge
total intl minutes
total intl calls
total intl charge
customer service calls
1.0
0.8
0.6
0.4
0.2
0.0

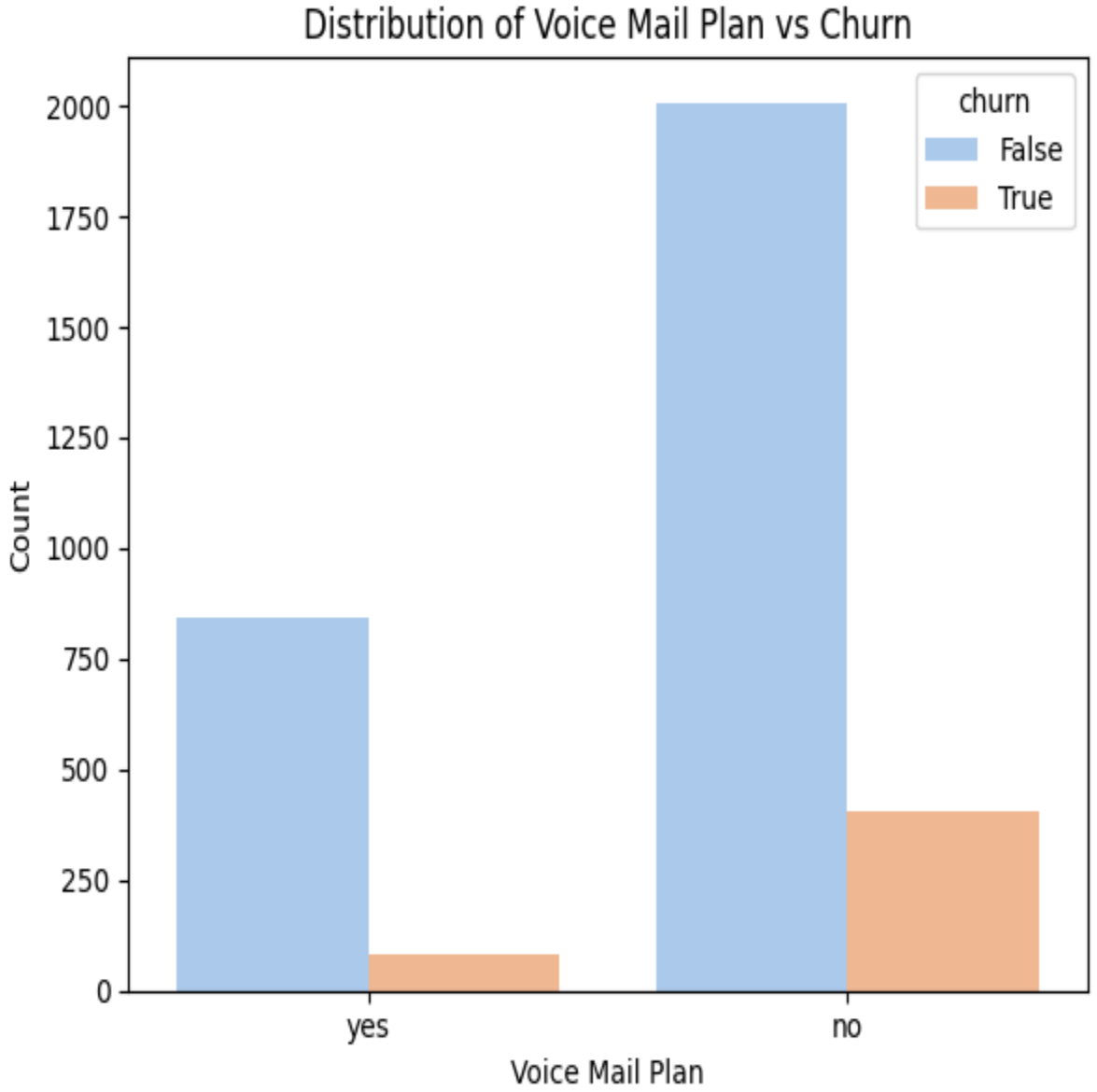
Correlation matrix for Numeric features

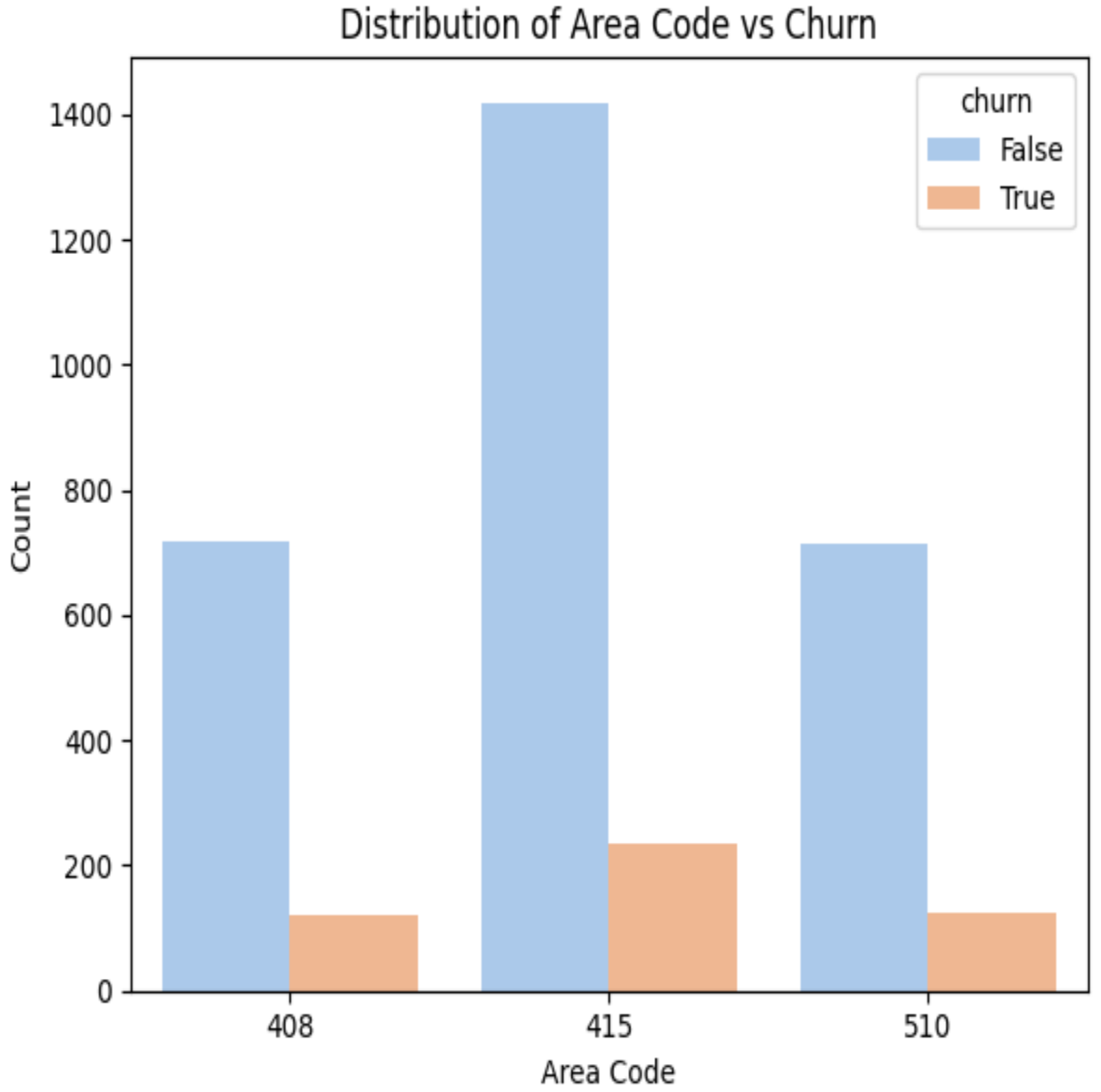
- The correlation matrix provides insights into how each feature relates to one another, which can help us identify patterns and dependencies in our data.

- Features with high correlation.

  - **total day minutes and total day charge have a correlation of 0.9999999521904007**
  - **total eve minutes and total eve charge have a correlation of 0.9999997760198491**
  - **total night minutes and total night charge have a correlation of 0.9999992148758795**
  - **total intl minutes and total intl charge have a correlation of 0.9999927417510314**

-


Distribution of Voice Mail Plan vs Churn
churn
False
True
Count
2000
1750
1500
1250
1000
750
500
250
0
yes
no
Voice Mail Plan


Distribution of Area Code vs Churn
churn
False
True
Count
1400
1200
1000
800
600
400
200
0
408
415
510
Area Code

Detecting Outliers

- **Impacts of Outliers in Classification Models**
  1. Model Performance: Outliers can distort the decision boundary in classification models, leading to less accurate predictions. Models might become too sensitive to these outliers, affecting the overall performance
  2. Bias and Variance: Outliers can increase the variance of the model, leading to overfitting. This can make the model too complex and less generalizable to new data.
  3. Feature Importance: Outliers can affect the importance of features, potentially leading to misleading insights about which features are most relevant for classification.
  4. Number of outliers removed: 164

MODELLING

- The two main models used were:
  1. Logistics model
  2. Decision trees.

Logistic Regression Model

Model Overview:

Logistic Regression is a simple linear model used for binary classification.

It predicts the probability of a customer churning based on the weighted combination of input features.

Performance of the logistic regression model.

- **Training Accuracy**: The model achieved an accuracy of 86.59% on the training set.
- **Test Accuracy**: The model's accuracy on the test set was 84.85%, indicating decent generalization to new data.
- **AUC (Test)**: An Area Under the Curve (AUC) score of 0.79 suggests moderate capability in distinguishing between churn and non-churn cases.

- **Key Observations**:

The model has a low recall for predicting churn, which means it may miss identifying some customers who are likely to churn.

Decision Tree model.

- The Decision Tree model is a non-linear model that splits the data into branches based on feature importance, providing a set of decision rules for predicting outcomes.

- **Hyperparameter Tuning**:
  We optimized parameters like max_depth (maximum depth of the tree), min_samples_split (minimum number of samples required to split an internal node), and min_samples_leaf (minimum number of samples required to be at a leaf node) to enhance the model's performace.

Performance:

- **Training Accuracy**: The model achieved 100% accuracy on the training set, suggesting overfitting.
- **Test Accuracy**: The model performed well on the test set with a 91% accuracy.
- **AUC (Test)**: The AUC score of 0.85 indicates a good ability to distinguish between churn and non-churn customers.

- **Key Observations**:
  While the model shows excellent performance, the perfect accuracy on the training set hints at potential overfitting. However, the test results suggest it generalizes reasonably well to unseen data

Feature Importance (Decision Tree)

- **Top Features Influencing Churn**:

The top five features that most significantly impact customer churn are:

1. **Total Day Charge**: This is the most significant predictor of churn.
2. **Total Day Minutes**: Closely related to the total day charge, indicating that high daytime usage is a churn indicator.
3. **Total Evening Charge**: High evening charges also correlate with churn risk.
4. **Total International Charge**: Charges for international calls are a key factor in customer decisions.
5. **Total International Calls**: The number of international calls made also impacts churn likelihood.

- **Implication**:
  These features suggest that customer usage patterns, particularly related to call charges and international usage, play a critical role in determining churn.

ROC Curve Analysis

- **Logistic Regression vs. Decision Tree**:
- **Comparison**: The ROC curves for both models demonstrate their performance in distinguishing between churn and non-churn customers.
- **Findings**: The Decision Tree model has a higher AUC score on the test data, indicating better recall and ability to capture churn cases compared to Logistic Regression.

- **Conclusion**:

The Decision Tree model slightly outperforms Logistic Regression in both recall and overall predictive ability.

Conclusion

- **Model Comparison**:

Decision Trees provide better interpretability and performance compared to Logistic Regression, especially in identifying key features driving customer churn.

- **Feature Insights**:

Key usage patterns, like daytime and international charges, are significant churn predictors, which can guide targeted customer retention strategies.

- **Overall Model Choice**:

Given the better performance and interpretability, Decision Tree is the preferred model for this problem.

Recommendations

- **Reduce Churn**:
  - **Monitor High-Usage Customers**: Target customers with high day and evening charges with tailored retention offers.
  - **Improve Customer Service**: Address customer complaints effectively, as frequent service calls are linked to higher churn rates.
  - **Reevaluate International Plans**: Review the pricing and structure of international call plans to ensure they are competitive and aligned with customer needs.

Future Steps:

- Further refine the decision tree model to prevent overfitting and enhance generalization.
- Explore advanced models like Random Forests or Gradient Boosting Machines for potentially better performance.

Acknowledgments:

Thank You:

I would like to extend my gratitude to the entire team especially the technical mentors from moringa school for their support and contributions throughout this project. Your input has been invaluable in achieving these insights.