Predicting Customer Churn for a Telecom Company

Data Science Classification Project

Introduction

Customer churn is a critical issue for telecom companies, as losing subscribers directly impacts revenue. This project aims to **predict which customers are likely to churn** using demographic, account, and usage data. The dataset contains **3,333 customers and 21 features**, including account length, call usage, and customer service interactions. By accurately identifying potential churners, the company can implement **proactive retention strategies**, reduce revenue loss, and improve customer satisfaction.

Objectives

- **Primary Goal:** Predict which telecom customers are likely to churn.
- **Why It Matters:** Early identification of potential churners allows the company to implement **targeted retention strategies**, reducing revenue loss.
- **Approach:** Use customer demographic, account, and usage data to build **classification models** that accurately identify churners.
- **Business Impact:** Improve customer retention, increase revenue, and enhance customer satisfaction.

Data Understanding

Dataset Size: 3,333 customers with 21 features.

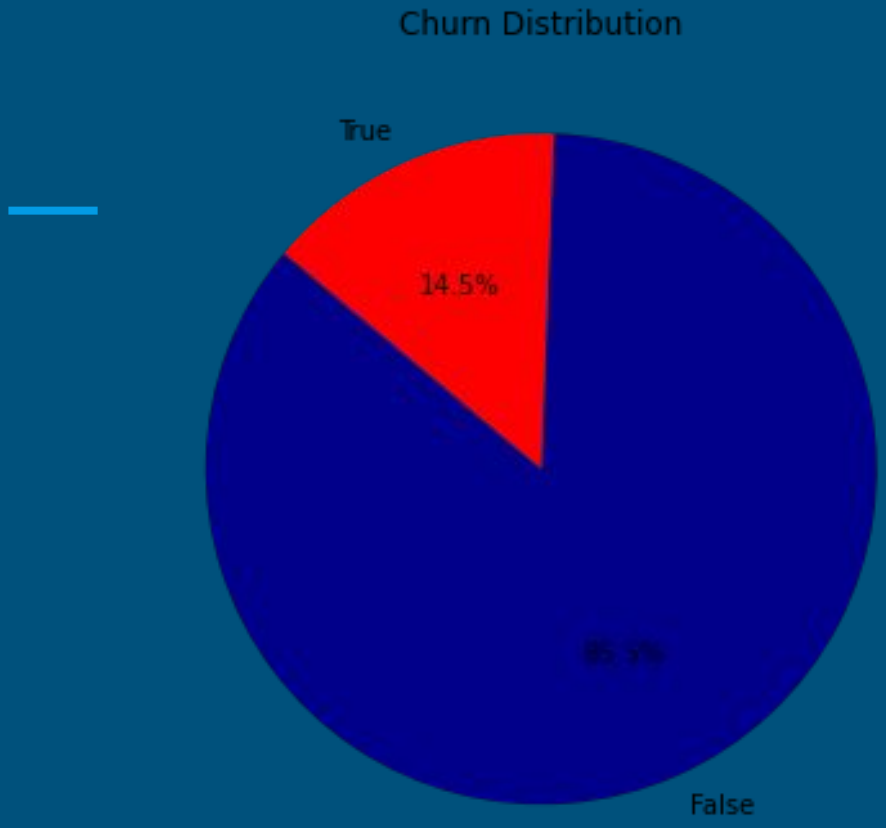
Target Variable: churn (1 = churned, 0 = stayed).

The dataset is **imbalanced**, with many churners and few non-churners.

Numerical Features:

- Account-related: account length, total day/eve/night/intl charges
- Usage-related: total day/eve/night/intl minutes and calls, number vmail messages
- Customer service interactions: customer service calls

Categorical Features: state, area code, international plan, voice mail plan


Churn Distribution
True
14.5%
False

Data Exploration and Cleaning

Churn Insights:

- Customers with **higher customer service calls** are more likely to churn.
- Those with **international plans** or **high day minutes** show higher churn rates.

Feature Relationships:

- Certain numeric features (e.g., total day minutes, total eve minutes) are correlated with churn.
- Outliers exist in usage metrics like total night calls and total intl minutes.

Model Preprocessing

Data Cleaning: Checked for missing values and handled inconsistencies.

Feature Encoding: Categorical variables (state, area code, international plan, voice mail plan) encoded for modeling.

Feature Scaling: Numerical features scaled for models sensitive to magnitude (e.g., Logistic Regression).

Feature Selection: Top predictive features identified using correlation analysis and model-based importance.

Train-Test Split: Dataset split into training and testing sets (stratified by churn to maintain class distribution).

Model Training

Models Trained:

- Logistic Regression (baseline)
- Decision Tree
- Random Forest

Approach:

- Iterative modeling: feature selection → scaling → training → evaluation → tuning.
- Top features used to improve model efficiency and interpretability.

Goal: Identify the model that best predicts churn while balancing precision and recall.

Model Evaluation

Evaluation Metrics: Focused on **accuracy, precision, recall, and F1-score**, especially for churners.

Key Findings:

- Logistic Regression: Poor precision and F1 for churners (many false positives).
- Decision Tree: Good balance, interpretable.
- Random Forest: Very high precision but lower recall.
- Random Forest (Threshold 0.3): Best balance between recall (0.78) and precision (0.70), highest F1 (0.74).

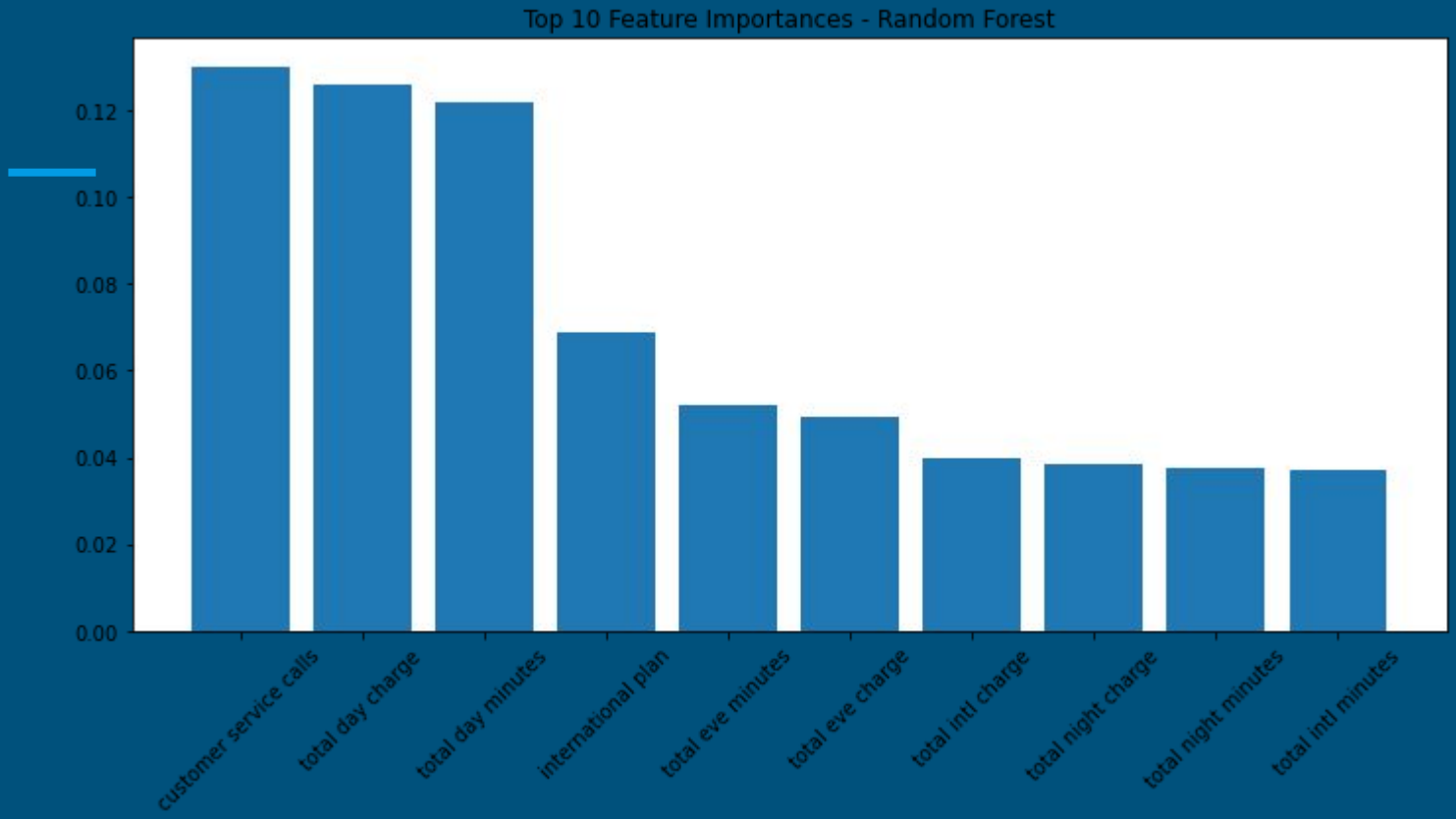
Insights

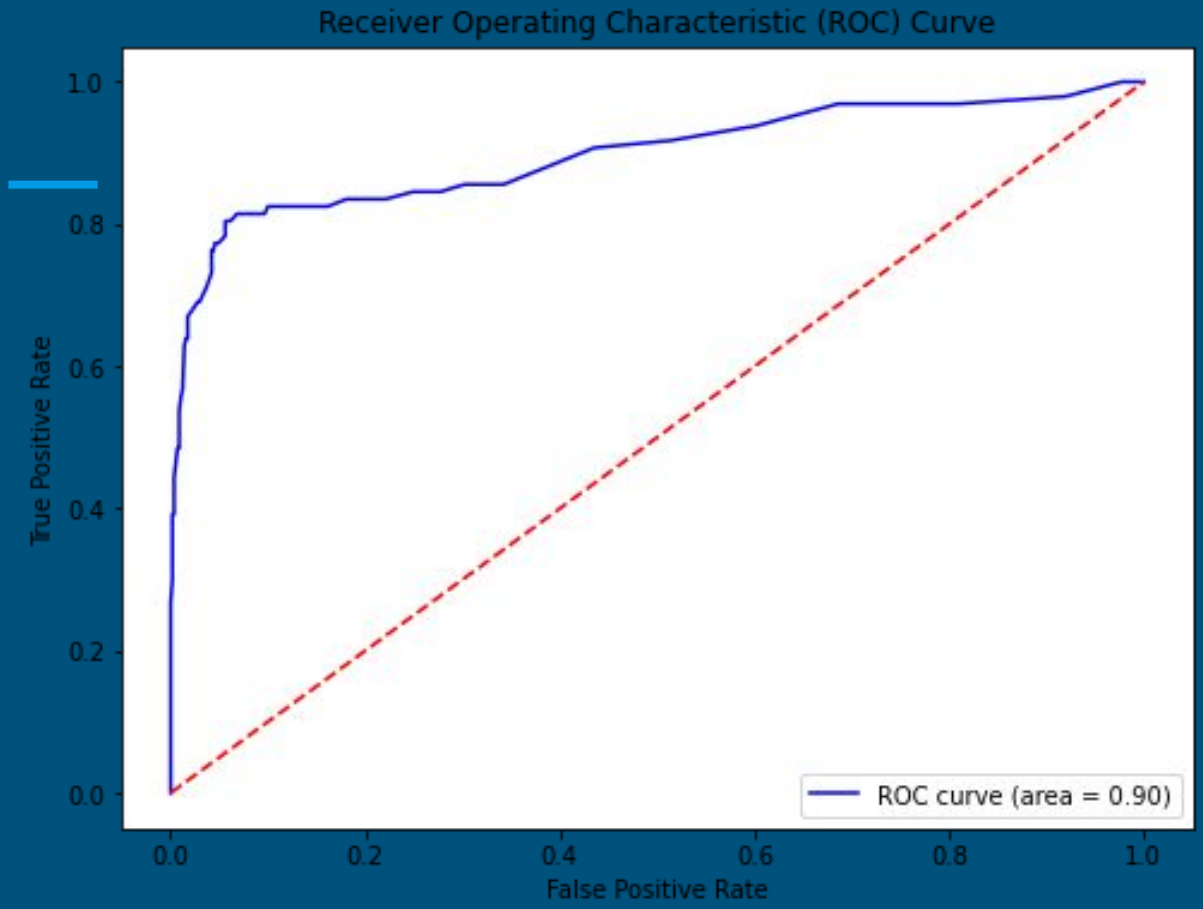
Best Model: Random Forest (Threshold 0.3) balances recall (0.78) and precision (0.70), achieving the highest F1 (0.74) for churners.

Key Observations:

- Customers with **higher customer service calls** are more likely to churn.
- **International plan subscribers** and heavy **daytime users** show higher churn rates.
- Decision Tree is **interpretable** but slightly lower F1 (0.69).
- Logistic Regression is weak; standard Random Forest is conservative (high precision, low recall).

Business Implication: Focus on **high-risk churners** for proactive retention strategies.


Top 10 Feature Importances - Random Forest
0.12
0.10
0.08
0.06
0.04
0.02
0.00
customer service calls
total day charge
total day minutes
international plan
total eve minutes
total eve charge
total intl charge
total night charge
total night minutes
total intl minutes


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.90)

Recommendations

- **Deploy the Random Forest (Threshold 0.3)** model for churn prediction.
- **Target high-risk customers** identified by the model with proactive retention campaigns.
- **Monitor model performance** regularly and retrain with updated data to maintain accuracy.
- **Adjust thresholds** if business priorities change (e.g., higher recall vs higher precision).
- **Leverage key features** like `customer service calls`, `international plan`, and `total day minutes` to guide marketing and retention strategies.
  Address class imbalance in future modeling using oversampling, undersampling, or class weighting to improve churn detection.