PROJECT OVERVIEW

This project uses machine learning algorithms to develop a model for accurately predicting customers who churn using the dataset provided. The dataset contains 20 predictor variables primarily reflecting customer usage patterns, with a total of 3,333 records. Among these, 483 represent customers who have churned, while the remaining 2,850 are non-churners.

The target variable, churn, is categorical, making classification algorithms suitable for building the predictive model. Model performance is assessed using recall as the primary evaluation metric.

BUSINESS UNDERSTANDING

Business Problem

For telecommunications companies, growing their revenue base depends on attracting new customers while simultaneously increasing customer retention. Customer churn is a critical concern for large businesses, representing the rate at which subscribers or regular customers cancel their subscriptions or stop engaging with the company.

Churn can occur for various reasons, including switching to competitors offering better prices, dissatisfaction with customer service, or disengagement due to a lack of meaningful touchpoints.

Syriatel, a leading mobile telecommunications and data services provider based in Damascus, Syria, offers a range of services including calls, messaging, news, GSM, and internet services. The company has built a strong reputation by prioritizing customer satisfaction and social responsibility. Syriatel recognizes that fostering long-term customer relationships is more cost-effective than acquiring new customers. Retaining existing customers is central to their strategy, making churn prediction a critical business priority.

This project aims to develop a machine learning model that accurately predicts customers most likely to churn and identifies key factors driving churn. With this insight, Syriatel can proactively address customer issues and take actions to prevent churn, ultimately enhancing customer retention.

Objectives

- To build a machine learning model that can accurately predict customers who will churn based on the information available in the dataset.
- To identify the features that are important for predicting customer churn.
- To deliver insights that enable the company to implement cost-effective and targeted retention strategies, improving customer satisfaction and reducing the cost associated with acquiring new customers.

Success Criteria

- Predictive Model Performance: Develop a model capable of accurately predicting customer churn with a recall score of at least 85%, ensuring the model identifies a significant portion of churned customers achieve a balance between precision and recall to minimize false positives and false negatives.
- Insights on Churn Drivers: Identify the most important features influencing churn, providing actionable insights into customer behavior and dissatisfaction.
- Reduction of False Negatives: Ensure the model minimizes false negatives, as failing to identify churned customers is costlier for the business.
- Handling of Imbalanced Data: Successfully address the class imbalance in the dataset so that the model performs well for both churned and non-churned customers.
- Model Interpretability: Ensure the model and its predictions are interpretable, allowing stakeholders to trust and understand the results for decision-making.
- Business Relevance: Translate model outcomes into actionable business strategies, such as identifying at-risk customers and offering targeted interventions to retain them.
- Scalability: Create a solution that can be scaled and adapted to future datasets, enabling ongoing churn prediction as customer behaviors and patterns evolve.

DATA UNDERSTANDING AND PREPARATION

Source of the Data

Kaggle.

Overview

The dataset consists of customer data from Syriatel, a telecommunications company. The data is used to predict customer churn and contains various features describing customer behavior and demographics.

Structure of the Dataset

- Number of Records: 3,333
- Number of Features: 21
- Target Variable: churn (binary categorical variable)
- Predictor Variables: 20, including numeric and categorical features.

Key Features

1. **Numerical Features:**
    - Includes customer usage statistics such as total day minutes, total eve minutes, total night minutes, and total intl minutes.
    - Financial metrics like total day charge, total eve charge, and total night charge.
2. **Categorical Features:**
    - international plan and voice mail plan: Whether a customer subscribes to these plans.
    - churn: Target variable indicating whether a customer has churned (yes or no).

Initial Observations

1. **Imbalanced Target Variable:**
    - **Churn distribution:**
        - Non-churners: 85.5% (2,850 customers)
        - Churners: 14.5% (483 customers)
    - The imbalance highlights the need for techniques like oversampling or class weighting.
2. **Numerical Data Statistics:**
    - Features like total day minutes, total eve minutes, and total night minutes have continuous distributions.
    - Correlations suggest multicollinearity among variables related to minutes and charges.
3. **Categorical Data Insights:**

- Binary variables like international plan and voice mail plan need encoding for machine learning models.
- Certain categories may have a significant influence on churn behavior.

Data Quality Assessment

1. **Missing Values:**
   - No missing values were detected.
2. **Duplicate Records:**
   - No duplicate entries were found in the dataset.
3. **Outliers:**
   - Some numerical features exhibit outliers, particularly in usage-based metrics (e.g., total day minutes). Further investigation may be needed to assess their impact on model performance.
4. **Multicollinearity:**
   - Features like total minutes and total charge for day, evening, and night are highly correlated. Dimensionality reduction (e.g., feature selection may be necessary.

Summary

The dataset provides a rich set of features to explore customer behavior and predict churn. Key points to note include:

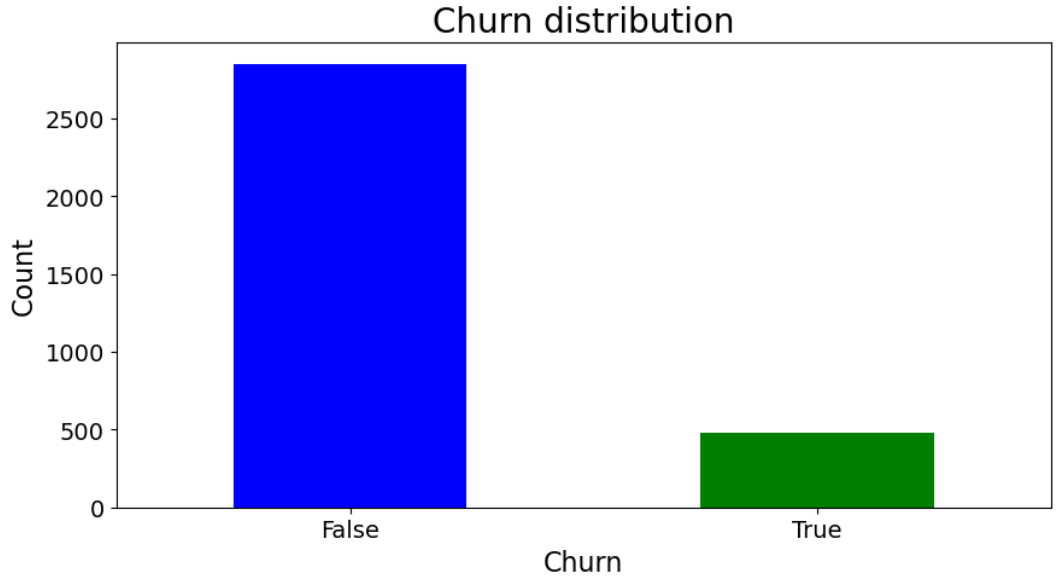
- Dropped irrelevant columns such as phone number column.
- Balancing the target variable using techniques like SMOTE.
- Encoding categorical variables into numerical format.
- Addressing multicollinearity and potential outliers in numerical features.
- Leveraging feature engineering to enhance model performance.

DATA ANALYSIS

Univariate Analysis

- Churn Distribution: Check the proportion of churned vs. non-churned customers. There is a class imbalance problem since the target class has an uneven distribution

of observations. 85.51% of the data belongs to the False class while 14.49% belongs to the true class.



Multivariate analysis

- The heatmap correlation visually represents the relationships between features in the dataset, showing how strongly they are linearly related. Correlation values range from -1 to 1.
- There is a very low correlation between most features.
- However, there is a perfect positive correlation between total evening charge and total evening minutes, total day charge and total day minutes, total night charge and total night minutes, and total international charge and total international minutes. This is expected since the charge of a call depends on the length of the call in minutes. One correlated variable will have to be dropped from each pair to handle multicollinearity.
- total day minutes, total day charge and customer service calls have a weak positive correlation with churn.
- The other features have a negligible correlation with churn, approximately 0. 0. 0.

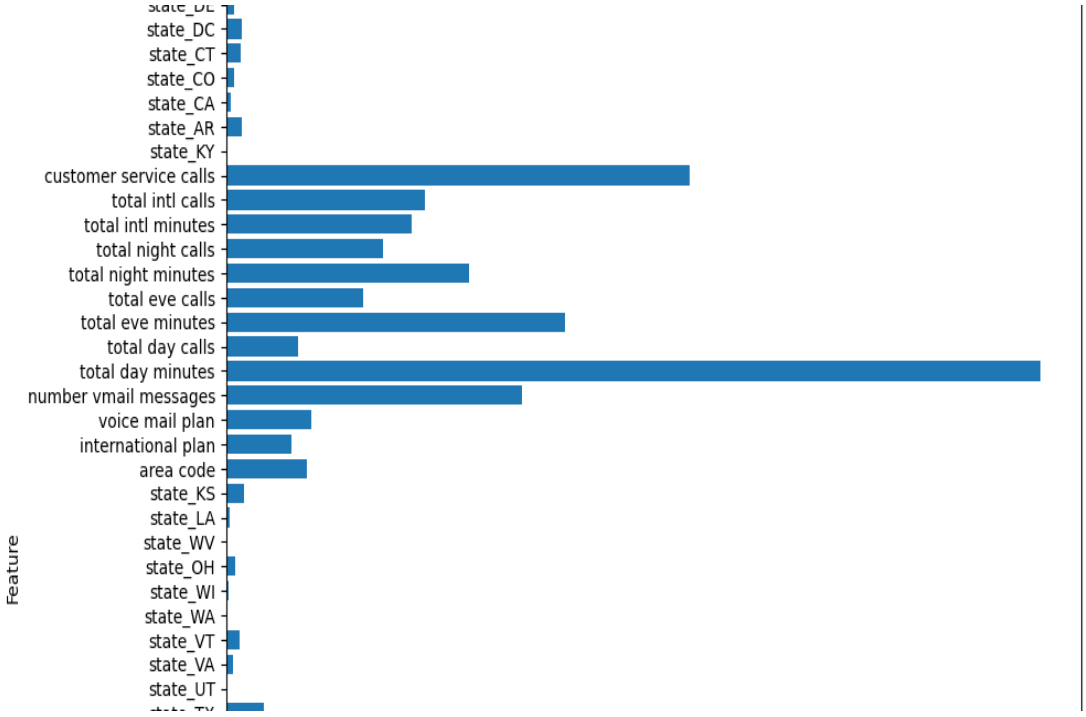
MODELLING

Model Selection

- **Baseline Model**
  - Logistic Regression for simplicity and interpretability.
- **Second Model**
  - Decision Tree (Default parameters) for feature importance analysis.
- **Third Model**
  - Hyperparameter-tuned Decision Tree for performance optimization.

MODEL EVALUATION

- The decision tree model with optimized hyperparameters demonstrates the best performance. The optimal parameters identified are:
  - 'clif_criterion': 'entropy'
  - 'clf__max_depth': 28
  - 'clf__max_features': 15
  - 'clf__min_samples_leaf': 2
  - 'clf__min_samples_split': 2
- This model achieves the highest recall score among all tested models, with accuracy and precision scores exceeding average benchmarks. However, the recall score remains below the target threshold of 85%.



CONCLUSION

The final model selected for predicting customer churn is the decision tree with optimized hyperparameters. This model minimizes the number of false negatives, making it highly effective at identifying customers likely to churn.

The key features contributing the most to predicting customer churn are:

- total day minutes: total number of minutes the customer has been in calls during the day
- total evening minutes: total number of minutes the customer has been in calls during the evening
- customer service calls: number of calls the customer has made to customer service
- total international minutes: total number of minutes the user has been in international calls

RECOMMENDATION

Syriatel should prioritize delivering exceptional customer service to meet customer expectations and carefully analyze customer interactions. Following up on both positive and negative feedback is essential for maintaining customer satisfaction.

Additionally, data indicates that customers who spend more time on calls are less likely to churn. The company should evaluate its call charge rates relative to competitors and consider reducing per-minute rates if necessary to retain more customers and reduce churn.

NEXT STEPS

The best-performing model falls short of achieving the targeted recall score of 85%. Despite hyperparameter tuning, some overfitting persists.

Increasing the size of the training dataset could help reduce overfitting and enhance the model's overall performance.