MORINGA SCHOOL

DATA SCIENCE

SYRIATEL CUSTOMER CHURN PROJECT REPORT

AUGUSTINE KOMEN

DECEMBER 2024

Contents

1. PROJECT OVERVIEW

This project utilizes machine-learning algorithms to build a model that effectively predicts customers at risk of churning. The dataset, sourced from Kaggle, includes 20 features mainly related to customer usage patterns, and consists of 3,333 records. Of these, 483 customers are identified as churners, while 2,850 are non-churners. The goal of the model is to classify the "churn" target variable using classification algorithms. The model's performance is evaluated using recall as the metric. Ultimately, the decision tree model, fine-tuned with hyperparameters, proves to be the most effective.

2. BUSINESS UNDERSTANDING

2.1 Business Problem

For telecommunications companies, growing their revenue base relies on both gaining new customers and enhancing customer retention. A major challenge for large enterprises is customer churn, which occurs when a subscriber or regular customer ends their subscription or stops doing business with the company. Churn can be driven by various factors, such as switching to a competitor with better prices, leaving due to poor customer support, or disconnecting from a brand because of limited engagement opportunities.

Syriatel, a mobile telecommunications and data services provider based in Damascus, Syria, offers a variety of services, including calls, messaging, GSM, and internet. The company has earned a strong reputation by focusing on customer satisfaction and social responsibility. Syriatel understands that maintaining long-term customer relationships is more cost-effective than constantly acquiring new customers. Therefore, predicting customer churn has become a crucial part of the company's strategy. This project aims to create a model that accurately identifies customers who are likely to churn and pinpoints the key features leading to this prediction. With this information, Syriatel can take proactive steps to prevent customer churn.

2.2 Objectives

1. To develop a machine-learning model that can effectively predict customer churn using the information provided in the dataset.

2. To pinpoint the key features that play a significant role in predicting customer churn.
3. To provide recommendations on specific actions the Syriatel company can take to reduce churn based on model insights.

2.3 Metrics of success

- Recall Score: ≥ 80%

3 DATA UNDERSTANDING

Source of Data: Kaggle

Data Description: The dataset contains 3333 rows and 21 columns, including

demographic data, usage statistics, service plans, and the churn target variable.

3.1 Numeric Columns:

1. **account length**: The number of days or months a customer has been subscribed to the service.
2. **area code**: A numeric code representing the geographical area where the customer's phone is registered.
3. **number vmail messages**: The count of voice mail messages stored in the customer's account.
4. **total day minutes**: The total number of minutes the customer used during the day.
5. **total day calls**: The total number of calls made during the day.
6. **total day charge**: The total cost of calls made during the day.
7. **total eve minutes**: The total number of minutes the customer used during the evening.
8. **total eve calls**: The total number of calls made during the evening.
9. **total eve charge**: The total cost of calls made during the evening.
10. **total night minutes**: The total number of minutes the customer used during the night.
11. **total night calls**: The total number of calls made during the night.
12. **total night charge**: The total cost of calls made during the night.
13. **total intl minutes**: The total number of international minutes used by the customer.
14. **total intl calls**: The total number of international calls made by the customer.

15. **total intl charge**: The total cost of international calls.
16. **customer service calls**: The total number of times the customer called customer service.

3.2 Categorical Columns:

1. **state**: The state where the customer resides.
2. **phone number**: The customer's unique phone number.
3. **international plan**: Indicates whether the customer has subscribed to an international call plan (e.g., "Yes" or "No").
4. **voice mail plan**: Indicates whether the customer has subscribed to a voicemail plan (e.g., "Yes" or "No").

4 DATA PREPARATION & ANALYSIS

4.1 Data Preparation

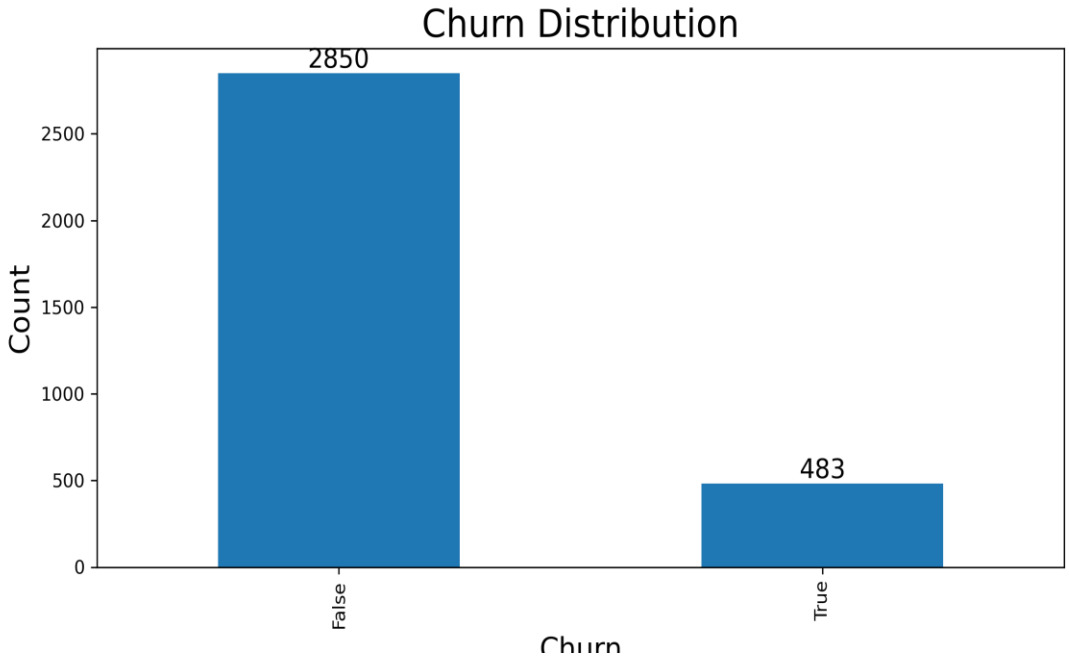
- Checks Performed:
- No missing or null values.
- No duplicate rows detected.
- Outlier analysis on numeric columns

Actions Taken:

- Dropped irrelevant columns such as phone number column.
- Encoded categorical features (state, international_plan, etc.) using one-hot encoding.

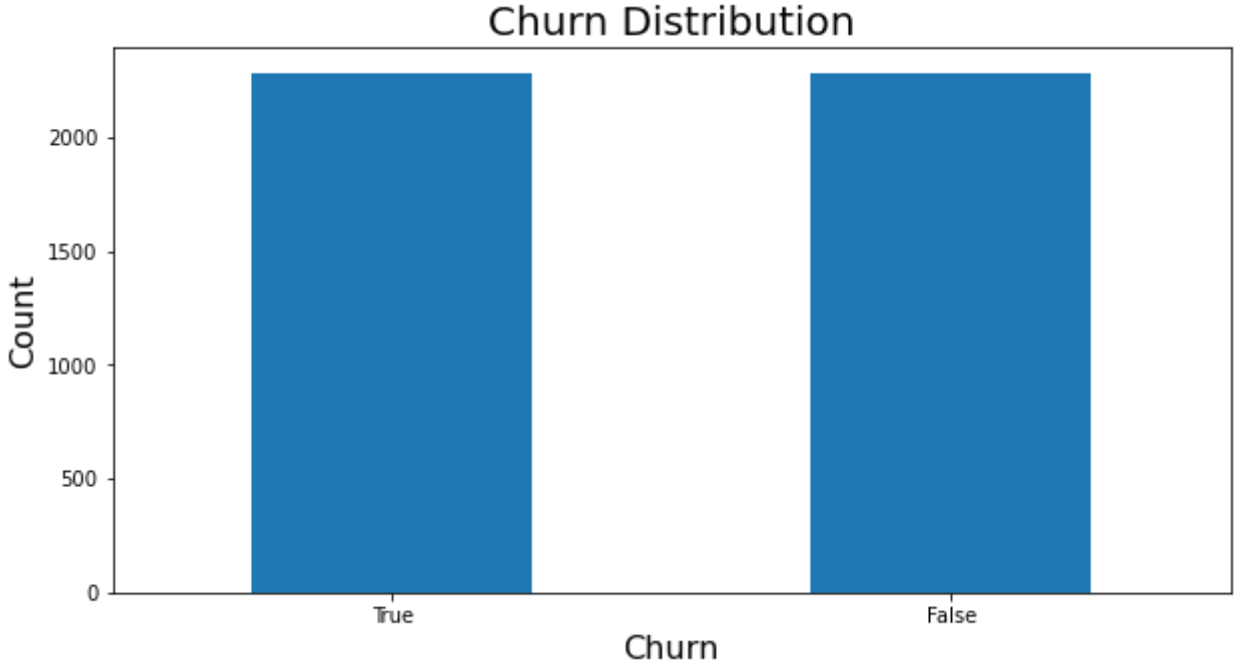
4.2 Data Analysis

- **Univariate Analysis:**
    - Distribution of churn: Approximately 85% non-churners and 15% churners (class imbalance).



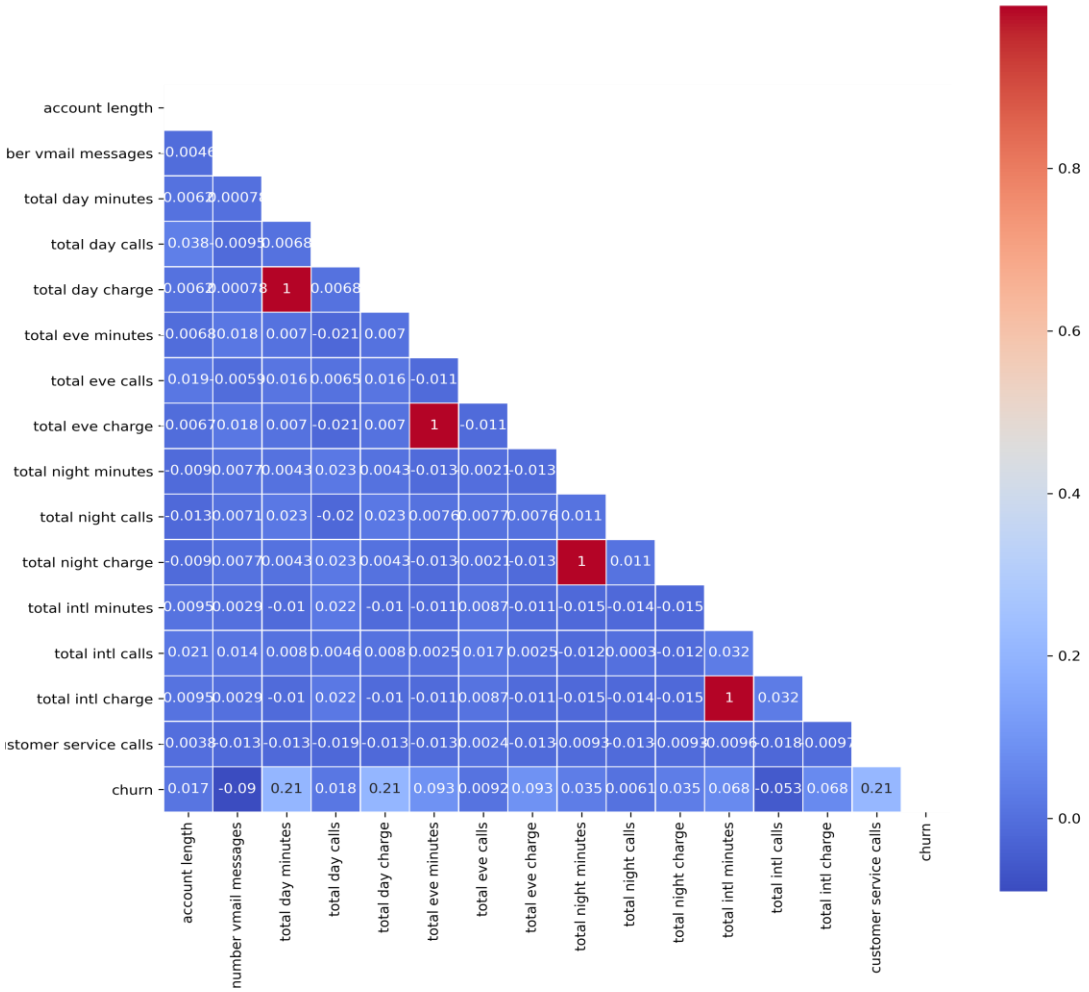
- Usage patterns of services (for example; total day minutes distribution).

SMOTE was used to address class imbalance issues by oversampling the minority class with replacement.



- **Bivariate & Multivariate Analysis**:

- Churn rates are higher for customers with the international_plan.
- High correlation between total day charges and total day minutes (dropped multicollinearity before modeling).



- Most features exhibit a notably weak correlation with each other.
- Nonetheless, a perfect positive correlation is observed between specific pairs of variables: 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 correlation is anticipated since the charge of a call is inherently influenced

by the call's duration in minutes. To address multicollinearity, it will be necessary to eliminate one variable from each correlated pair.

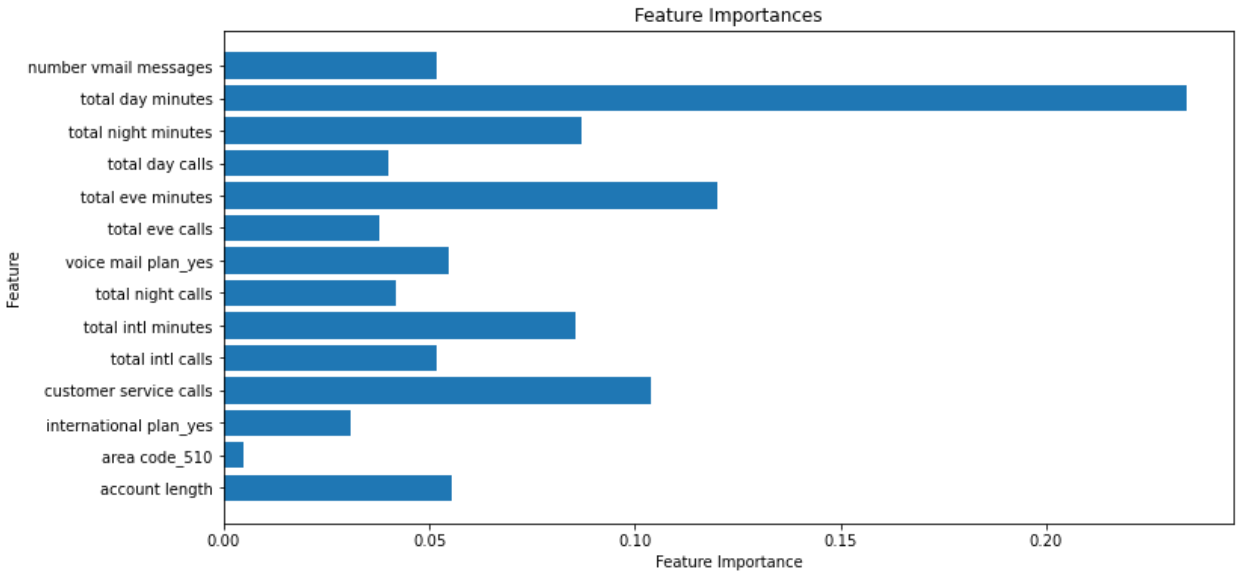
5 MODELLING

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.

6 MODEL EVALUATION

- The logistic regression model, which served as our baseline model exhibited overfitting as it excels on the training data but underperforms on the test data. Despite achieving high accuracy on the test dataset, it displays a notably low recall score. This suboptimal performance is primarily attributed to the substantial count of false negatives.
- The most influential factors for predicting customer churn are, in descending order of importance, total day minutes, total evening minutes, and customer service calls. On the other hand, the state variable has minimal significance in predicting customer churn.



- The decision tree model with tuned hyperparameters emerged as the top-performing model. It is optimized with the following parameters: {'criterion': 'entropy', 'max_depth': 24, 'max_features': 11, 'min_samples_leaf': 2, 'min_samples_split': 2}. This model excels in achieving the highest recall score, and its accuracy and precision scores are well above average. However, it's worth noting that the recall score falls slightly short of the target of at least 80%.

7 CONCLUSIONS

- The chosen model for predicting customer churn is the decision tree with fine-tuned hyperparameters, boasting the lowest count of false negatives.
- Key features crucial for forecasting customer churn are:
  - Total Day Minutes: Reflects the cumulative minutes spent by the customer on daytime call.
  - Total Evening Minutes: Sums up the minutes spent on evening calls by the customer.
  - Customer Service Calls: Indicates the number of calls the customer has initiated to reach customer service.

8 RECOMMENDATION

1. **Customer Service Strategy**: Syriatel should ensure a robust customer service strategy to meet customer expectations effectively and assess customer interactions. This includes tracking and addressing both positive and negative feedback from customers.
2. **Resolve Customer Issues**: Since customer service calls have the highest correlation with churn, implying that the more times a customer initiates a call to reach customer service, the more likely they are to churn. Syriatel should ensure that issues raised by clients who call are resolved promptly and efficiently.
3. **Call Charge Rates**: Given that total day and night minutes are key factors for predicting churn, Syriatel should assess its call charge rates in comparison to competitors. Lowering the charges per minute for calls could prove instrumental in retaining customers and preventing churn.

9 NEXT STEPS

- Deploy the model for real-time customer churn prediction
- To further refine the model and overcome challenges like overfitting, additional data collection is essential. This could include gathering data on customer satisfaction, competitor pricing, and service quality metrics. Deploying the model into a live environment will also provide valuable feedback for continuous improvement.