SyriaTel Customer Churn

A Data-Driven Analysis for Strategic Intervention

Table of contents

1

INTRODUCTION

Project Overview

Objective: Develop a machine learning classification model that predicts whether a customer will churn, enabling SyriaTel to take proactive actions that reduce churn and improve profitability.

Data: SyriaTel Customer Churn Dataset on Kaggle

Audience:

SyriaTel Management
Customer Retention Team

Business Understanding

- **Stakeholder Need:** Understand and reduce customer churn by identifying high-risk customers and implementing targeted retention strategies.
- **Business Questions:**
  - Which customer segments are most likely to churn?
  - What behavioral, service, or account factors contribute most to churn?
  - How can SyriaTel effectively intervene to retain high-risk customers?

- **Success Outcome:** An evidence-based customer retention strategy that reduces churn and improves customer lifetime value.

2

METHODOLOGY

Data Source

- The dataset used in this analysis is sourced from the **SyriaTel Customer Churn Dataset** on Kaggle.
- It contains information on **3,333** customers of SyriaTel, a telecommunications company.
- The dataset is structured for a **binary classification problem**, with the target variable churn indicating whether a customer left the company (True/False).
- The dataset captures customer account details, usage behavior (calls, minutes, charges), and interactions with customer service.

Data Summary

- **3,333** customers with **21** attributes: demographics, account info, usage, and churn.
- **Average usage:** 101-day accounts, ~100 calls/day period, 180–200 minutes/day.
- **International usage**: low (10 minutes, 4–5 calls).
- **Customer service calls:** 1.6 on average, up to 9.
- **Plans:** ~90% no international plan, ~72% no voicemail plan.
- **Coverage:** 51 U.S. states; phone numbers unique (not predictive).
- **Churn:** 14.5% churned, 85.5% active.
- **Data quality:** no missing values or duplicates – ready for modeling.

Data Cleaning

- **Data Cleaning**: Dropped non-predictive columns, standardized categories, and checked data types.
- **Key Analysis Fields:** Customer usage, subscription plans, service calls, and churn; created categorical labels for churn and area code.
- **Outlier Handling:** Retained extreme values to preserve meaningful customer behavior patterns.

Modeling

- Encoded categorical variables:
  - Binary: churn, international plan, voicemail plan
  - Non-binary: State, Area code (one-hot encoded)
- Addressed multicollinearity by creating two datasets:
- Removed engineered/multicollinear features
- Models trained: **Logistic Regression**, **Decision Tree**, **Random Forest**, **XGBoost**
- Key metrics: **F1 Score** & **Precision** prioritized for operational relevance

Tools

Python

Core Programming Language

Pandas & Numpy

Numerical Analysis & Data processing

Scikit Learn

Modeling

Seaborn & Matplotlib

Visualization

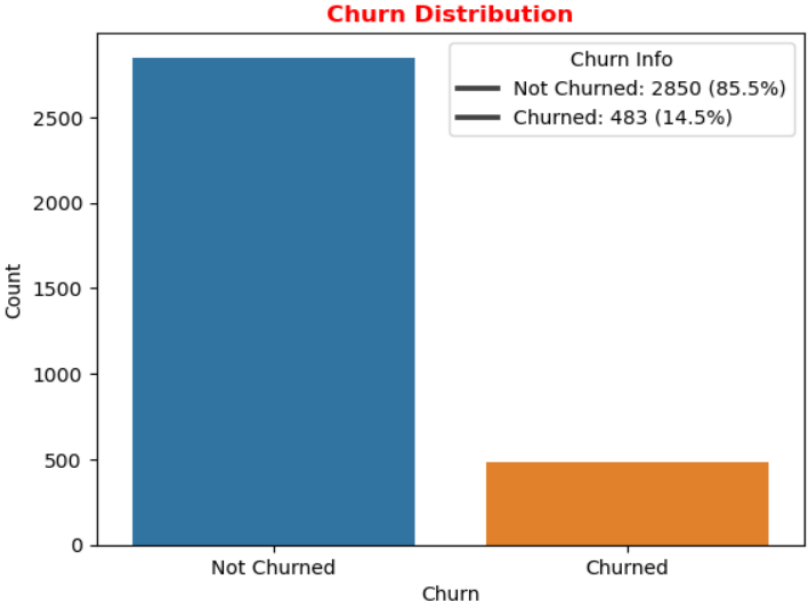
Jupyter Notebook

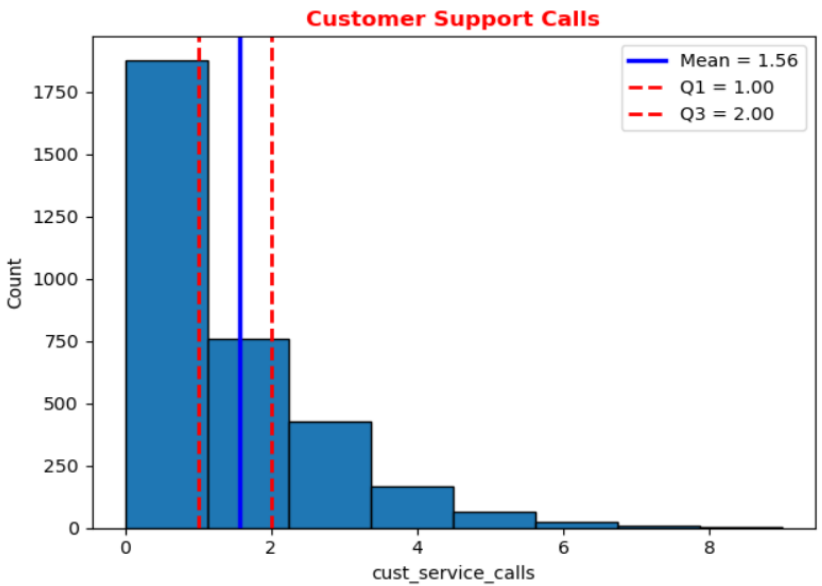
Interactive analysis and Documentation

3

RESULTS

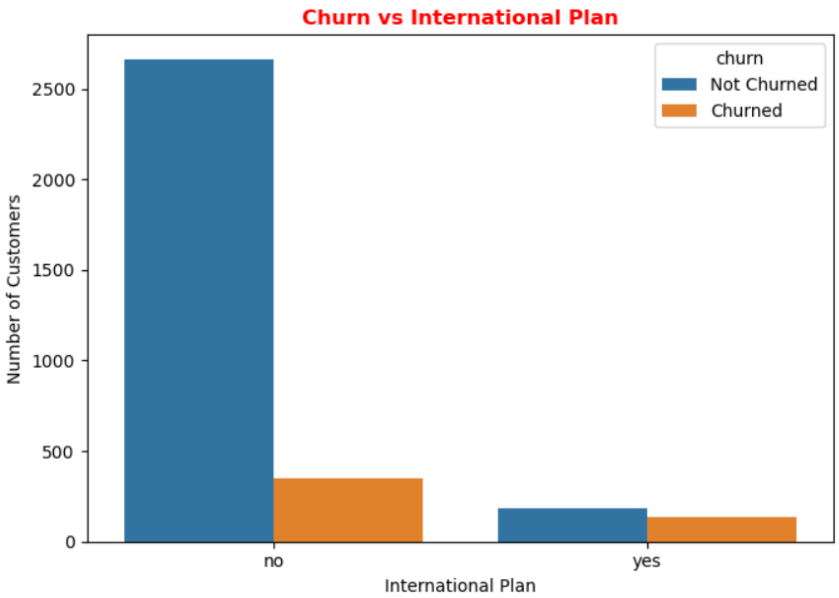
Churn and Service Calls

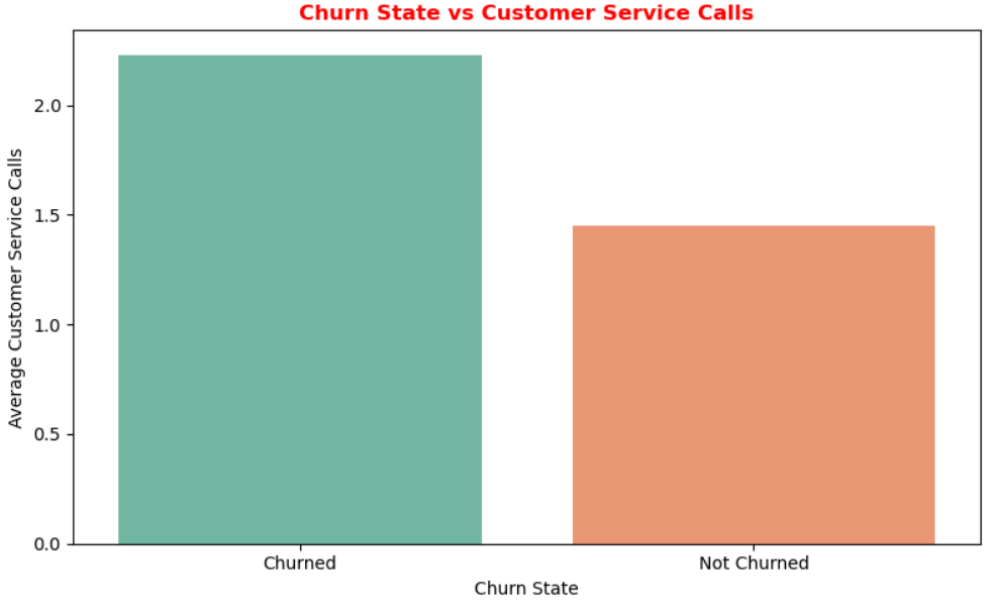




- Target variable imbalance: 85% did not churn, 14.5% churned.
- Majority of customers (over two-thirds) did not make any customer support calls.
- Customers who contacted support averaged 1 call, with a maximum of 8 calls; distribution is left-skewed.

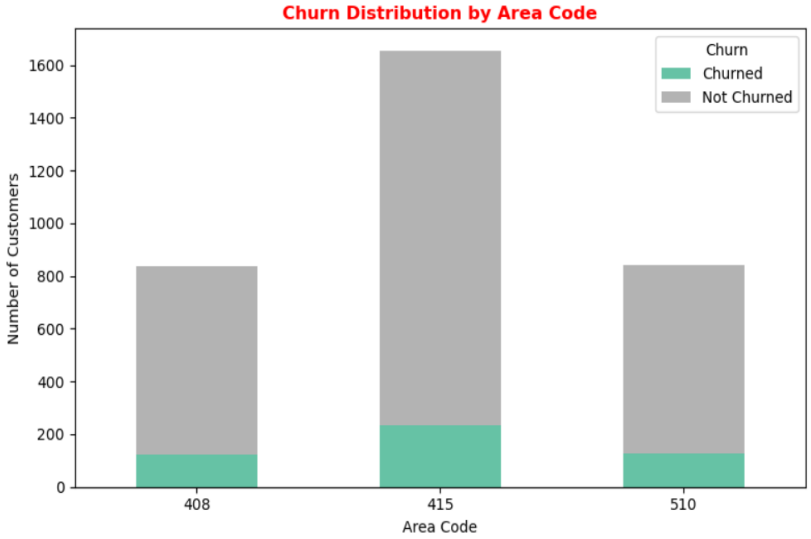
International Plan and Service Calls

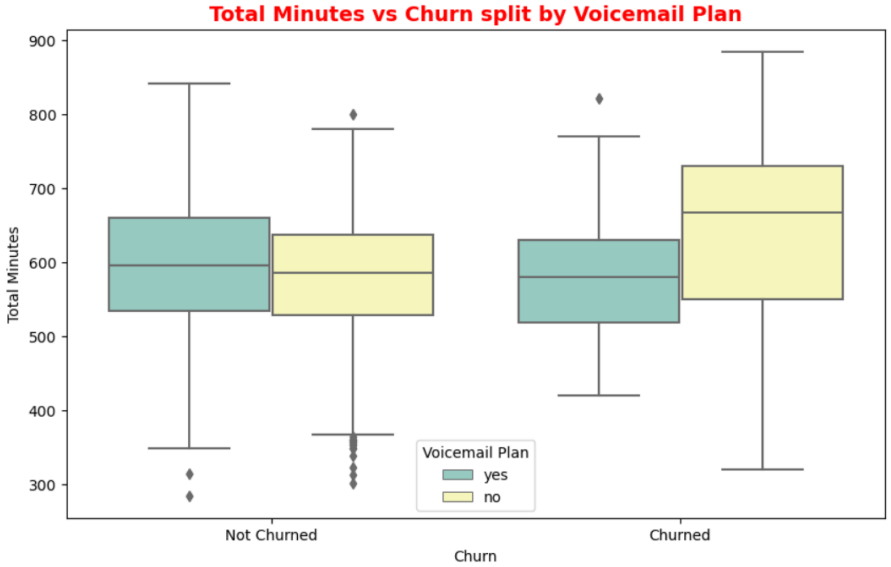




- Customers with an international plan show higher churn, indicating a greater service expectation.
- Customers who churned made more customer service calls on average than those who did not. Frequent support interactions may indicate unresolved issues or dissatisfaction, potentially driving churn

Area Code and Voice Mail Plan





- Area code 415 has the highest customer count and a notably larger churn rate, indicating potential service quality or satisfaction issues in this region.
- Churned customers without a voicemail plan show higher variability and more outliers in usage, suggesting dissatisfaction among heavy users. Non-churned customers with a voicemail plan have more consistent usage, indicating that voicemail access may support retention.

Model Performance

- **Without handling multicollinearity:**
  - XGBoost best F1: 88.99%, Precision: 100%
  - Tree-based models show strong performance; Logistic Regression weaker (F1: 36.46%)
- **After removing multicollinearity:**
  - XGBoost F1: 80.18%, Precision: 88.12%
  - Random Forest Precision: 90.14%
- **Class imbalance addressed with SMOTE:**
  - XGBoost F1: 78.76%, Precision: 84.76%
  - Random Forest F1: 59.30%, Precision: 75.64%
- Cross-validation confirms **XGBoost** as best trade-off for capturing churners while minimizing false positives

Final XGBoost Model

- **Final XGBoost performance:**
  - Accuracy: 93.88%, Churn F1-score: 0.77, Churn Precision: 0.82
- **Top features driving churn:**
  - Area Code
  - Voice Mail Plan –
  - Customer Service Calls
  - International Plan

4

RECOMMENDATIONS

Focus Retention on At-Risk Segments

- Identify customers with International Plans and frequent customer support calls.
- Offer proactive outreach, loyalty rewards, or usage-based incentives to these high-risk groups.

Improve Quality of Service and Billing

- Optimize customer service to reduce repeat or unresolved support calls.
- Implement flatrate billing for heavy users to prevent unexpected high charges.

Leverage Plan Features to Reduce Churn

- Encourage adoption of voicemail plans for high-usage customers.
- Offer customized international or bundled plans for heavy international callers.
- Educate customers on plan benefits to highlight convenience, predictability, and cost savings.

Regional Retention Strategies

- Investigate and resolve service or infrastructure issues in area code 415 with high churn.
- Deploy localized promotions and faster issue resolution to strengthen loyalty in at-risk states.

THANK YOU

Prepared by: KIPROP BRIAN
Email: brian.kibor@student.moringaschool.com
Questions? I'm available for walkthroughs or further data analysis sessions.