CUSTOMER CHURN PREDICTION FOR TELECOMMUNICATION

September 2025

Project Data Report: Predicting Customer Churn Using Machine Learning

1. OVERVIEW

This project focuses on analysing whether a customer will soon stop doing business with SyriaTel, a telecommunications company, and to build models which will help to predict which customers are at risk of leaving. The goal is to identify the possible reason for customer churn and to build a predictive model that can help the company reduce customer loss. The dataset contained Key variables included tenure length, tenure period, payment method, monthly charges, and churn status.

2. INTRODUCTION

2.1 Business Problem

Customer churn is a major issue for telecommunication companies because it directly reduces the number of active subscribers. Customer churn leads to

reduction in the revenue and higher acquisition cost. Holding onto existing customers is more expensive than acquiring new ones. Understanding why the customers leave can allows better targeting at retaining efforts. The major challenge is to understand why customers churn and to predict which customer are

likely to leave. A key problem is that churn is often not evenly distributed among all customers, some groups of customers are more likely to churn than others

and this creates an imbalance. The overall problem is not just predicting churn but also understanding why customers leave; by solving this business problem

the company can reduce churn, improve satisfaction, and increase profitability.

2.2 Objectives

- Comprehend the dataset and determine the key factors impacting customer churn.
- Cleaning of the dataset.
- Perform exploratory data analysis (EDA) to identify churn pattern.
- Build and evaluate machine learning models.
- Provide actionable recommendations based on results.

2.3 Scope and Limitations

- The dataset may not capture all factors influencing churn.
- Focuses majorly on the churned and not the non-churned.
- The dataset focuses on one telecommunication company so results may not apply on the whole industry.
- Models may need updating regularly as customer behaviour changes.

3. DATA UNDERSTANDING

3.1 Data Source

The data source is Kaggle.

Contains 21 columns and 3333 rows

Customer churn is the target variable (1 = churned, 0 = active).

Qualities:

- *Plans*: Whether customers have an international plan or a voicemail plan.
- *Usage*: Total call minutes, number of calls, and associated charges across day, evening, night, and international usage.
- *Account Information*: State, account length, and area code.
- *Customer Support*: Number of customer service calls made.
- *Identifiers*: Phone number.

3.2 Relevance

The dataset provides a strong basis understanding of customer likely to churn and why, and guiding customer retention strategies.

4. DATA PREPARATION AND CLEANING

Handling Missing Values:

- Analyse missing data patterns.
- Impute numerical missing values with median; categorical with mode.
- Possibly drop irrelevant or heavily missing columns.

Encoding Categorical Variables:

- Use One-Hot Encoding or Label Encoding for categorical features (e.g., contract type).

Scaling:

- Scale numerical features.

Train-Test Split:

- Stratify split to maintain churn class ratio, e.g., 80% train / 20% test.

5. EXPLORATORY DATA ANALYSIS (EDA)

In the Exploratory Data Analysis, we examined the distribution of churn and the relationship between customer features and churn behaviour. The results showed that most customers stay but a notable proportion leaves, more customer service calls are linked with a higher likelihood of churn, customers with a month-to-month contract, higher total charges, and more frequent customer service calls had a higher likelihood of leaving the company. Correlation checks revealed no strong multicollinearity, meaning that most features contributed unique information. In conclusion, the EDA displayed service usage, billing details, and customer support interactions. of whether a customer is likely to churn.

6. MODELING AND EVALUATION

6.1 Logistic Regression

Logistic Regression was first trained without balancing the dataset. The model achieved a high accuracy since most customers did not churn. This means the model often predicted that customers would stay even if they actually left. But after balancing logistic regression performed better at detecting the churns. recall improved significantly, meaning more at-risk customers were correctly identified.

6.2 Decision Tree

The Decision Tree model without balancing captured some churn patterns but leaned heavily towards non-churn, resulting in high accuracy but weak recall for churn. It tended to overfit to specific patterns in the data, making it less effective at predicting churners. With balancing the dataset allowed the decision tree to capture more effectively on churn-related features such as service calls and contract type. Recall improved, but precision decreased as the tree predicted more false positives.

6.3 Evaluation

In all models, the unbalanced versions had higher accuracy but very poor recall for churners, making them less useful for retention strategies. While the balanced versions, while sacrificing some accuracy, greatly improved recall and F1-score.

Logistic regression works well in identifying churn vs non-churn and it is better at generalisation making it have a lower risk at overfitting, while decision trees have a slightly lower accuracy and it does not generalize as well making it to have a risk at overfitting. We recommend using Logistic Regression as the primary churn prediction model and using Decision Trees for business communication, since its decision paths are easy to understand and translate.

7. RECOMMENDATION

Based on the analysis, the company should prioritize strategies that reduce churn by focusing on customers who show early warning signs, such as frequent customer service calls, short contract types, or high day-time usage. The company should Launch targeted retention campaigns for customers with month-to-month contracts, improve customer service quality, since high service calls are linked to churn, continuously monitor churn with updated models and real-time data, management should consider introducing more flexible contract options to retain customers who prefer month-to-month plans.