Predicting Customer Churn for a Telecom Company

Team Insight

Nisarg Patel, Parth Patel

Department of Information Systems, University of Maryland, Baltimore County

IS 733: Data Mining

Dr. Karen Chen

October 07, 2024

Table of Contents

1. Project Proposal: Predicting Customer Churn for a Telecom Company

1.1 Use Case Development

1.1.1 Context:

Telecom companies face the challenge of customer churn, where users discontinue their services. Understanding and predicting which customers are likely to churn can help the company take proactive steps to retain them. By utilizing machine learning models, this project aims to predict customer churn and provide insights into the factors influencing customer decisions.

1.1.2 Importance:

Reducing churn directly impacts profitability. Retaining customers is more cost-effective than acquiring new ones, making this prediction model valuable for business strategy. With this system, telecom companies can take targeted actions (e.g., offering discounts, improving customer service) to prevent at-risk customers from leaving.

1.1.3 Beneficiaries:

Marketing and customer service departments will benefit from the insights generated, as they will be able to develop more effective retention strategies. The company's bottom line will also benefit from reduced churn rates and increased customer satisfaction.

1.1.4 Use Case:

A telecom service manager uses the churn prediction tool to identify customers who are most at risk of leaving. Based on the predictions, the team launches personalized retention campaigns. These actions could involve targeted offers, improved service packages, or timely customer support interventions.

1.2 Datasource Description

1.2.1 Data Source:

The dataset for this project comes from the Telco Customer Churn Dataset on Kaggle. It contains detailed records of over 7,000 telecom customers, including demographic information, contract details, payment methods, service usage patterns, and whether or not they churned.

Definition of Churn: In the Telco Customer Churn dataset, a customer is marked as "churned" if they discontinued service within the period that the data was collected. The churn status is typically indicated at the end of the billing cycle or after the customer's contract was terminated.

Probing into Churn Timing:

- **Analysis of Contract Types**: Determine if churn events are more common at certain stages of a customer's contract (e.g., end of month-to-month contracts versus long-term contracts).

- **Feature for Last Interaction**: Identify if there are features or derived metrics that can show the last interaction date before churn or periods of inactivity leading to churn.

1.2.2 Existing Work:

Several models like logistic regression, decision trees, and random forests have been applied to churn prediction. However, this project will aim to use advanced models such as gradient boosting or ensemble methods to achieve higher accuracy and interpretability.

1.3 Model Development

1.3.1 Machine Learning Problem:

This is a **supervised learning** problem, focusing on **classification**. The goal is to predict whether a customer will churn (yes/no) based on various features.

1.3.2 Training Data:

Each row in the dataset represents one customer. The columns (features) include demographic data, service usage, and billing information.

1.3.3 Input Variables:

- Customer demographics (gender, age, etc.)
- Contract type (monthly, yearly)
- Payment method (electronic check, mailed check)
- Service usage (internet, phone)
- Monthly charges and total charges

1.3.4 Output/Label:

The target variable is a binary outcome indicating whether the customer has churned or not (1 for churn, 0 for no churn).

1.4 Model Evaluation

1.4.1 Performance Metrics:

The model performance will be evaluated using the following metrics:

- **Accuracy**: To measure the overall success rate of predictions.
- **Precision/Recall**: To understand how well the model identifies actual churners.
- **F1-score**: To balance precision and recall.
- **AUC-ROC Curve**: To evaluate how well the model distinguishes between churners and non-churners.

These metrics will be gathered using cross-validation and testing on a holdout dataset.

1.5 Risk & Support

1.5.1 Risks:

- The dataset may have missing values or imbalances (e.g., more non-churners than churners), which could affect model performance.
- Model interpretability may be challenging with more complex algorithms like gradient boosting.

1.5.2 Support

Will discuss as we work on the project.