

Predicting Customer Churn for a Telecom Company

Milestone Report

Team Insight

Nisarg Patel, Parth Patel

Department of Information Systems, University of Maryland, Baltimore County

IS 733: Data Mining

Dr. Karen Chen

November 04, 2024

Table of Contents

Predicting Customer Churn for a Telecom Company.....	1
Table of Contents.....	2
Milestone Report: Predicting Customer Churn Analysis.....	3
1. Project Overview:.....	3
1.1 Problem Definition.....	3
2. Exploratory Data Analysis (EDA).....	3
2.1 Initial Observations.....	3
2.2 Data Cleaning.....	3
2.3 Data Visualization.....	4
3. Initial Modeling Iteration and Model Performance.....	4
3.1 Baseline Model Results.....	4
4. Next Steps.....	5
5. References.....	5

Milestone Report: Predicting Customer Churn Analysis

1. Project Overview:

The purpose of this project is to develop a predictive model to identify customers who are likely to discontinue using services from a telecom provider. This proactive approach allows telecom companies to address potential churn by offering incentives or enhancing customer service, thus improving customer retention rates. This milestone report outlines the initial steps in the project, including exploratory data analysis (EDA), data preparation, the first iteration of modeling, and performance evaluation.

1.1 Problem Definition

Customer churn is a critical issue in the telecom industry as retaining existing customers is often more cost-effective than acquiring new ones. The primary objective of this project is to answer the question: **What characteristics or behaviors are associated with a customer's decision to churn?** By identifying these factors, we can build a model that anticipates customer churn, enabling timely interventions.

2. Exploratory Data Analysis (EDA)

The initial analysis aimed to understand the dataset's structure and identify key characteristics influencing customer churn. The Telco Customer Churn dataset includes 7,043 entries and 21 columns, with data ranging from demographic information to service usage and billing details and target attribute Churn.

2.1 Initial Observations

- **Data Types and Missing values:** 18 categorical (object) columns, 2 integer columns 1 float column This insight guided how we handled data preprocessing.
- **Variable Distributions:** A histogram analysis revealed that most customers had short tenures (below 20 months), with fewer having long-term subscriptions (above 60 months). Monthly charges ranged broadly from \$18.25 to \$118.75.
- **Churn Analysis:** The churn rate was approximately 26%, highlighting a non-trivial portion of the customer base at risk. And it also says imbalance in dataset for churned and not churned
- **Categorical Insights:** Month-to-month contracts and customers using electronic checks were more likely to churn compared to those with long-term contracts or automated payments.

2.2 Data Cleaning

The **TotalCharges** column was initially an object type, which was converted to numeric after handling non-numeric values. Missing values were replaced using the median, ensuring data consistency.

2.3 Data Visualization

Visual exploration included:

- **Histograms** for tenure and monthly charges to identify common patterns and clusters in customer behaviors.
- **Bar plots** illustrating churn distribution by Contract Type, Internet Service, Payment Method, Paperless Billing, Phone Service and by Senior Citizen revealing churn distribution and giving insight on which feature has more impact on churn distribution.
- **Distribution plots** used to analyze numerical features and Churn Distribution plots will allow us to see how tenure, MonthlyCharges, and TotalCharges distributions differ for customers who churned vs. those who didn't.
- A **Correlation Heatmap** can help us identify if any numerical features are highly related to each other or to churn. Highlighted that **tenure** and **MonthlyCharges** had notable but modest correlations with churn.
- **Bivariate Analysis** of key features explored possible interactions, particularly between MonthlyCharges by Contract Type and churn, TotalCharges by InternetService and churn using **Box plot**

These visualizations indicated that certain customer attributes and characteristics significantly impacted churn probability. And also helped to understand how feature interaction determines churn possibility.

3. Initial Modeling Iteration and Model Performance

A **Logistic Regression** model was chosen for the initial analysis due to its simplicity and interpretability. Data was split into training (80%) and testing (20%) sets. Preprocessing included scaling numerical features and encoding categorical features using a pipeline approach to maintain consistency.

3.1 Baseline Model Results

- **Accuracy:** 80.4%
- **AUC-ROC:** 0.839, indicating the model's good capability to distinguish between churned and non-churned customers.
- **Classification Report:** Precision for predicting churn was 69%, and recall was 47%, suggesting room for improvement in catching potential churn cases.
- **Cross Validation Score:** 0.78 suggesting the model generalizes reasonably well but can be optimized further.

The confusion matrix highlighted that while the model performed well in predicting non-churn cases, it struggled with identifying all instances of churn, underscoring the need to enhance recall.

4. Next Steps

The initial model provided a reasonable starting point with balanced accuracy and precision but highlighted the challenge of improving recall. The next steps involve:


1. **Address Class Imbalance:**
 - Implement techniques such as **SMOTE (Synthetic Minority Oversampling Technique)** to balance the training set.
 - Use class weights in the model to give more importance to the minority class (churn).
2. **Feature Engineering:**
 - Create new features that may better capture patterns in customer behavior (e.g., interaction between Contract type and MonthlyCharges).
3. **Model Complexity:**
 - Move to more complex models like **Random Forests, Gradient Boosting** or **Support Vector Machines** to improve predictive power.
4. **Cross-Validation:**
 - Apply k-fold cross-validation for a more robust assessment of model generalization.
5. **Hyperparameter Tuning:**
 - Optimize model parameters using grid search or randomized search for better performance.
6. **Evaluate Model**
 - Evaluate and Analyze model performance specifically the precision-recall trade-off, especially since recall for churn is low.

5. References

Github Repo:

https://github.com/niss10/Telcom-Churn-Prediction/blob/main/Telcom_Churn_Prediction.ipynb

Google Drive Project Folder:  Team Insight

Team Journal :  Team_Insight_Journal.pptx