10/12/2025

ADELINE MAKOKHA
DSA 8502 Predictive and Optimisation Analytics
191199

TITLE

A Predictive and Optimisation Analytics Framework for Modelling Customer Churn in the Telecommunications Industry

Problem Statement

Telecommunication companies operate in a highly competitive environment where customer expectations for affordable, reliable, and seamless connectivity continue to rise. In recent years, many subscribers have migrated from one service provider to another in search of better service quality, resulting in persistent and costly customer churn across the industry. Churn directly affects revenue stability, market share, and long-term profitability. On average, acquiring a new customer costs significantly more than retaining an existing one, making churn reduction a strategic priority for telecom operators.

Despite the importance of forecasting churn, most existing churn-prediction approaches used by telecom companies remain largely reactive. They rely on static metrics, such as usage patterns, complaints, or revenue trends, which are evaluated through manual analysis. These methods are slow, prone to human bias, and insufficient for capturing the evolving behavioral signals that precede customer departure. As a result, operators struggle to anticipate churn early enough to design targeted retention interventions. This disconnect between technical churn models and actionable business insights has led to recurring revenue leakage and limited the effectiveness of retention strategies.

The problem this project aims to solve is developing a predictive model capable of identifying high-risk churn customers proactively, using machine learning and feature-engineering techniques that better represent customer behaviors and service experiences. By integrating various indicators such as service usage fluctuations, interaction patterns, and customer value profiles, the proposed solution intends to generate more accurate churn forecasts and provide decision-makers with interpretable insights that support timely, data-driven retention actions.

This project is significant because a well-designed churn prediction system can help telecommunication operators:

- Reduce revenue loss caused by unexpected customer exits
- Improve customer experience by enabling proactive outreach
- Optimize marketing and retention budgets
- Strengthen competitive advantage in a saturated market

The goal is to build an intelligent, scalable churn prediction model that bridges the gap between technical prediction outputs and real business value, enabling telecom firms to identify at-risk customers long before they leave the network.

Objectives for the Project

1. Develop a churn prediction model capable of accurately identifying customers at risk of leaving the network, with a target predictive accuracy benchmark of at least 85%.
2. Examine and document the key methodological and conceptual limitations found in existing telecom churn studies, highlighting gaps that the proposed model aims to address.
3. Build and train multiple machine learning models, such as Logistic Regression, Random Forest, and Gradient Boosting, using telecom customer data to forecast churn events.
4. Evaluate and compare the performance of these models using metrics such as precision, recall, F1-score, and AUC to determine the most reliable algorithm.
5. Deploy the best-performing churn model within an interpretable, interactive dashboard that provides actionable insights to decision-makers, enabling proactive customer retention strategies.

Type of Data and Size of the Data

The project will utilize a structured dataset sourced from the Mendeley Data repository, containing operational records from a major Bulgarian telecommunications operator. The dataset focuses specifically on business customers, including small, medium, and large enterprises, and is extracted from the company's internal business information system.

It consists of 8,454 instances**,** each representing a unique business account, along with 14 predictor attributes and one target variable indicating whether the customer has churned. The features capture a wide range of CRM and financial information, such as customer identification codes, value segmentation, geographic location, key account manager assignment, numbers of active, inactive, and suspended subscribers, total subscriber count, average mobile and fixed-line revenues, total revenue, and ARPU. The target attribute, Churn, specifies whether a business customer has switched to another telecom provider. As a structured dataset made publicly available through Mendeley, it offers a rich foundation for developing and evaluating machine learning models for churn prediction.

The Model to be used

The churn prediction task will be approached as a supervised classification problem in which the objective is to distinguish between customers who are likely to leave the network and those who will remain. To achieve this, the project will evaluate a diverse set of predictive models, including Logistic Regression, K-Nearest Neighbors, Decision Trees, Random Forests, Support Vector Machines, Naive Bayes, and Gradient Boosting methods. These algorithms were selected because they capture different patterns in customer behavior, ranging from linear relationships to more complex nonlinear interactions, making them well-suited for telecom datasets that contain both numerical and categorical attributes. Ensemble models such as Random Forest and Gradient Boosting will be given particular attention, as they are known to perform effectively on imbalanced datasets and often deliver higher predictive accuracy. Logistic Regression and Decision Trees will serve as interpretable baseline models, enabling comparison of transparency versus performance. After training and evaluating all models using established metrics such as accuracy, precision, recall, and F1-score, the best-performing algorithm will be chosen for the final churn prediction system. Complementary optimization techniques, such as hyperparameter tuning and class imbalance handling (e.g., SMOTE), will also be applied to improve model reliability.