Telecom Customer Churn Analysis

Akshay Abhyankar
University of Colorado Boulder
Boulder, Colorado, USA
Akshay.Abhyankar@colorado.edu

Aditi Ramesh Athreya
University of Colorado Boulder
Boulder, Colorado, USA
Aditi.RameshAthreya@colorado.edu

Gaurav Roy
University of Colorado Boulder
Boulder, Colorado, USA
Gaurav.Roy@colorado.edu

Varun Bhaskara
University of Colorado Boulder
Boulder, Colorado, USA
Varun.Bhaskara@colorado.edu

ACM Reference Format:
Akshay Abhyankar, Aditi Ramesh Athreya, Gaurav Roy, and Varun Bhaskara. 2023. Telecom Customer Churn Analysis.

1 MOTIVATION

The technique of analyzing and forecasting customer attrition in the telecom sector is known as telecom churn analysis. When users of telecom services, such as internet or mobile phone plans, cancel their subscriptions or migrate to a competitor, this is known as customer churn. Churn analysis seeks to comprehend the causes of customer turnover, identify the customers most likely to leave in the future, and create plans to lower churn rates.

For telecom providers, churn analysis holds immense significance as it contributes to safeguarding revenue, cutting costs, enhancing the overall customer experience, fine-tuning customer segmentation, enabling targeted marketing efforts, and facilitating product development. In the eyes of telecom companies, retaining existing customers is always preferable due to the fact that acquiring new customers typically incurs higher costs.

The telecom operational system is designed to cater to a specific average number of customers. Falling below this calculated threshold is considered detrimental to the company. Even a modest effort to retain an existing customer can result in a substantial increase in both revenues and profits. Hence it is essential to construct robust machine learning models that can identify the causes of churn and propose necessary enhancements for customer retention.

2 LITERATURE SURVEY

This presents a short summary of churn prediction in telecom industry aswell as related work proposed by renowned researchers

Amal et al [1], does an extensive survey on the customer churn prediction by using 5 different datasets and by applying all the preprocessing, data mining techniques that have earlier been employed and evaluating those bases on parameters which have earlier been used for evaluation. The survey explores other research papers and gathers the performance of several different models(more than 8) and compares them. This paper shows all the comparison results in a tabular format so that it gets easier to compare based on various features.

Preeti et al [2], forecast's customer churn through the utilization of statistical survival analysis methods like logistic regression and decision trees. It marks the significance of meticulously choosing attributes and features to enhance the precision of churn prediction. This system holds considerable potential for telecommunications companies as it empowers them to pinpoint and retain customers at risk of churning, consequently safeguarding revenues typically allocated to customer acquisition and retention efforts. Rather than giving too much importance to models, it majorly gives more importance to feature selection as that can result to different results in case of different models.

Prabhadevi et al [3], utilized the Kaggle dataset, which comprises 7044 instances featuring 21 distinct attributes. The study began with data collection and preprocessing, which involved tasks such as one-hot encoding and label encoding to convert categorical labels into numerical format. . Multiple classification algorithms were employed to categorize customers into churn and non-churn categories. These algorithms included Stochastic Gradient Booster, Random Forest, KNN, and Logistic Regression. The performance of these algorithms was assessed based on various metrics, including Accuracy, Precision, Recall, F1-score, support, ROC (Receiver Operating Characteristic), and AUC (Area Under the Curve). Notably, the Stochastic Gradient Booster algorithm achieved the highest accuracy among the evaluated models.

Ullah et al [4], uses two datasets. The first dataset is obtained from South Asia GSM telecom service provider for studying customer churn prediction problem. It has 64,107 instances with 29 features. The second dataset is a publicly available churn-bigml dataset http://bigml.com/user/francisco/gallery/dataset/5163ad540c0b5e5b22000383. The dataset contains 3333 instances and 16 features. The data preprocessing consists of noise removal and feature selection using Information Gain and Correlation Attributes Ranking Filter techniques. Various classification algorithms are used to classify the customers as churn and non-churn customers. Algorithms such as Random Forest, Decision Stump, J48, Random Tree, AdaboostM1 + DecisionStump and Bagging + Random Tree, Naive Bayes, Multilayer Perceptron (MLP), Logistic Regression (LR), IBK and LWL. These algorithms are evaluated using Accuracy, FP Rate, TP Rate, Precision, Recall and F-measure. The Random Forest algorithm has the highest accuracy. After classification, the churn customers are categorized into low, medium and risky customers using the

K-Means clustering algorithm.

Verbeke et al [5], explores the data mining techniques and machine learning models to predict the customer churn using real world telecom data as in many countries retention of telecommunication customers seems to be more profitable than gathering new customers. The preprocessing techniques mentioned in this paper include variable selection, which is selecting the best feature that would act as an event for the churn prediction, followed by oversampling as the datasets might have imbalanced data and then with data cleaning by removing null and unnecessary values. The paper then applies rule-based classifiers, decision trees, neural networks, ensemble methods, statistical methods, and support vector machines to predict the churns. The models are evaluated based on metrics like the area under the receiver operating curve (AUC) and top decile lift. The paper also introduces a "profit-centric" approach that evaluates models based on their ability to maximize profits by targeting the right customers for retention campaigns.

3 PROPOSED WORK

The dataset at hand, sourced from IBM Sample Data Sets, encapsulates a rich tapestry of customer information, offering insights into the factors that influence customer churn and retention. Each row within the dataset represents a unique customer, while the columns unfold a myriad of customer attributes.

In this research, we aim to predict customer churn in the telecom sector and develop targeted customer retention programs. The proposed work encompasses several key components. First, we will collect and preprocess a comprehensive dataset containing customer attributes, including churn status, subscribed services, account information, and demographic details. Data preprocessing will involve data cleaning, handling missing values, and transforming the data to ensure its quality and suitability for analysis.

Next, we will focus on feature selection and engineering, where we will identify relevant attributes, create new features, and aggregate data over different time intervals. Exploratory data analysis (EDA) will play a crucial role in uncovering insights from the dataset, using visualizations, statistical analysis, and data segmentation techniques.

For churn prediction, we will employ various machine learning algorithms, including logistic regression, decision trees, random forest and neural networks. The performance of these models will be evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC AUC. Given the potential class imbalance in the churn dataset, we will implement techniques like oversampling to address this issue.

We plan to deploy the selected churn prediction models for both real-time and batch scoring, enabling telecom companies to take timely retention actions and gain periodic insights. Customer segmentation based on churn likelihood will guide the development of tailored retention strategies, which may include personalized offers, service upgrades, and customer support enhancements.

The overarching goal of this research is to minimize churn rates while maximizing customer lifetime value and overall profitability.

4 EVALUATION

We plan on using the following strategies to evaluate our work

(1) **Confusion Matrix**: We chose confusion matrix as it is one of the most important measure while evaluating prediction analysis, it makes evaluation easy by categorizing the outcomes TP, TN, FP and FN; which in our case will be Customers correctly predicted as churners, Customers correctly predicted as non-churners, Customers incorrectly predicted as churners and Customers incorrectly predicted as non-churners. As it reports false positives and false negatives so it will be helpful for us in devising if there is anything wrong with the strategy in selecting the feature.

(2) **Accuracy**: As the most fundamental metric for evaluating ML models, accuracy will calculate the ratio of both the churners and the non churners and give us a high level view of the data.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

(3) **Precision, Recall and F1**: We will also evaluate the models predicting the churners by using Precision and Recall as they will help us in knowing the proportion of correctly predicted churners in a detailed way. And after precision and recall, we can apply F1 to check on the prediction of both churners and non churners.

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F1-score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$$

If required, we are also thinking of using AUC-ROC to evaluate the models due to its characteristics of handling discrimination in data and ways of handling imbalanced data. We plan to use

Cross-validation as our validation strategy:

Cross-Validation - Dividing the dataset into a training set and a validation/test set.
The model is trained on the training set, and its effectiveness is evaluated on the validation/test set.
Iteratively, training and validating the model while dividing the data into different subsets, cross-validation (such as k-fold cross-validation), is a strong method that makes sure all data points are taken into account. To handle outliers, include outlier data points in different folds to ensure robust model evaluation. To handle the slight imbalance in classes we will be using resampling techniques

like oversampling and undersampling. Oversampling involves increasing the number of instances in the minority class. Undersampling involves decreasing the number of instances in the majority class.

5 MILESTONES

We propose the following timeline for the Telecom Customer Churn Analysis Project. The project's first phase, encompassing 1-2 weeks focuses on comprehensive data pre-processing, a pivotal step ensuring data quality and suitability for analysis. Following this, 1-2 weeks are allocated to exploratory data analysis (EDA), offering insights into underlying patterns and trends within the data set. Subsequently, 3-4 weeks will be invested in the application of diverse machine learning models, with an emphasis on model selection and optimization. A further 2-3 weeks will be devoted to rigorous model evaluation, employing various metrics and techniques to ensure robust and accurate results. The project's conclusion is allocated 1 week for comprehensive reporting, documentation, and presentation of findings.

REFERENCES

[1] Amal M. Almana, Mehmet Sabih Aksoy, and Rasheed AlZahrani. 2014. A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry. https://api.semanticscholar.org/CorpusID:1283811

[2] Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, and V. A. Kanade. 2016. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. https://doi.org/10.1109/CDAN.2016.7570883

[3] B. Prabadevi, R. Shalini, and B.R. Kavitha. 2023. Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks* 4 (2023), 145–154. https://doi.org/10.1016/j.ijin.2023.05.005

[4] Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul Islam, and Sung Won Kim. 2019. A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access* 7 (2019), 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

[5] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research* 218 (2012), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031