Telecom Customer Churn Prediction

Chethan Mukkapati
AM.EN.U4EAC21023
Dept.of Electronics and Communication
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Kollam, Kerala-690525

Majidh K
AM.EN.U4EAC21035
Dept.of Electronics and Communication
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Kollam, Kerala-690525

Yeswanth Kancherla
AM.EN.U4EAC21039
Dept.of Electronics and Communication
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Kollam, Kerala-690525

Nikhil Reddy Bojja
AM.EN.U4EAC21021
Dept.of Electronics and Communication
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Kollam, Kerala-690525

Dinesh D
AM.EN.U4EAC21024
Dept.of Electronics and Communication
Amrita School of Engineering
Amrita Vishwa Vidyapeetham
Kollam, Kerala-690525

***Abstract*—This project aims to develop a robust customer churn prediction model for the telecommunications industry. Customer churn, defined as customer defection, significantly impacts telecom companies due to fierce competition and ease of switching providers. Proactive churn prediction empowers companies to implement targeted retention strategies, ultimately reducing customer acquisition costs and bolstering revenue. This study investigates the effectiveness of three machine learning algorithms for churn prediction: Logistic Regression, Decision Tree, and Random Forest. We leverage the Telecom Customer Churn dataset from IBM Sample Data Sets. The performance of each model is meticulously assessed using a battery of metrics, including accuracy, precision, recall, and the AUC-ROC curve. By identifying the most effective churn prediction model, this research seeks to contribute to the development of data-driven customer retention strategies, leading to a competitive advantage for telecom companies.**

I. Introduction

The telecommunications industry is a fiercely competitive landscape characterized by a vast array of service providers and a high degree of customer mobility. Customers can easily switch providers in search of better deals or improved service quality, leading to a phenomenon known as customer churn, or customer attrition. This churn poses a significant financial burden on telecom companies, as the cost of acquiring new customers far exceeds retaining existing ones. Studies suggest that retaining existing customers can be up to five times more cost-effective than acquiring new ones.

Predicting customer churn empowers telecom companies to implement proactive retention strategies. By identifying customers at high risk of churning, companies can tailor targeted interventions to address their specific needs and concerns. These interventions can range from offering personalized discounts and promotions to improving customer service experiences. This not only reduces customer churn but also fosters customer satisfaction and loyalty, ultimately leading to a more robust and profitable business landscape for telecom companies.

This study delves into the application of machine learning algorithms for customer churn prediction within the telecommunications industry. We leverage three well-established algorithms: Logistic Regression, Decision Tree, and Random Forest. Each algorithm offers unique strengths and weaknesses, and our investigation aims to identify the model that delivers the most effective churn prediction in this context. We utilize the Telco Customer Churn dataset, a publicly available resource from IBM Sample Data Sets, to train and evaluate our models. Furthermore, we employ a comprehensive suite of performance metrics, including accuracy, precision, recall, and the AUC-ROC curve, to rigorously assess the efficacy of each model. Through this research, we strive to contribute valuable insights into the realm of customer churn prediction in the telecom sector. By pinpointing the most effective machine learning model, we aim to empower telecom companies to develop data-driven customer retention strategies that enhance customer satisfaction, loyalty, and ultimately, financial performance.

A. Platform

R Studio : It is served as the fundamental platform for this entire customer churn prediction project. RStudio is a powerful and user-friendly environment specifically designed for statistical computing and data analysis in the R programming language. R offers a vast library of pre-built packages for machine learning algorithms, data manipulation, and visualization. In this project, we utilized R packages specific to logistic regression (e.g., glm), decision trees (e.g., rpart), and random forests (e.g., randomForest) to construct and train each churn prediction model. RStudio's integrated development environment streamlined the coding process, allowing for efficient exploration, development, and evaluation of the

various machine learning models. Furthermore, RStudio's exceptional data visualization capabilities enabled us to generate insightful plots and graphs that facilitated the interpretation of model performance and the identification of key factors influencing customer churn. Ultimately, RStudio provided a comprehensive and robust platform that empowered us to effectively conduct our customer churn prediction analysis.

B. Dataset

This project utilizes the Telco Customer Churn dataset obtained from Kaggle. Each data point represents a customer record, containing demographics, account details, service subscriptions, and crucially, churn status (whether they discontinued service). This rich dataset allows us to explore factors influencing customer churn and train machine learning models for effective churn prediction.

II. Methodology

A. Data Understanding and Exploration

In the initial phase of data understanding and exploration, we loaded the Telco customer churn dataset. To gain preliminary insights into the data structure and variable types, we employed the glimpse function. We further delved deeper using summary to obtain descriptive statistics for numerical variables. Recognizing the importance of data quality, we addressed missing values through appropriate methods (mention the spec. This initial exploration established a foundational understanding of the data, preparing it for subsequent analysis.

B. Data Preprocessing

During data preprocessing, we tackled missing values to ensure data quality. Rows containing missing values were identified and removed to create a clean dataset. We then focused on categorical variables, which are crucial for building machine learning models. To prepare them for modeling, we encoded these categorical variables as factors. This encoding process transforms character-based categories into numerical representations suitable for model algorithms, allowing them to interpret and utilize the categorical information effectively.

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to understand how customer demographics and service usage relate to churn in the Telco customer dataset. This analysis involved examining demographic factors like gender and senior citizen status to see if they correlated with a higher likelihood of churning. Additionally, churn rates were visualized across various services offered (phone, internet) and contract types (monthly, yearly) to identify if specific service combinations or billing preferences impacted churn. These visualizations, likely created using ggplot2, helped identify potential patterns and relationships between customer characteristics, service usage, and churn behavior. This initial exploration provides valuable insights for further analysis and model building to predict customer churn more effectively.

D. Model Building

To evaluate different modeling approaches, we split the preprocessed data into training and testing sets using a function like **createDataPartition**. This separation ensures the models are trained on a representative portion of the data and subsequently tested on unseen data for generalized assessment. We then built and compared the performance of three popular machine learning models: decision tree, random forest, and logistic regression. Each model offers unique strengths in handling different data complexities, and this exploration aimed to identify the model that best captures the underlying churn patterns within the Telco customer data.

1) Decision Tree: This method is one of the most commonly used tools in machine learning analysis. We will use the rpart library in order to use recursive partitioning methods for decision trees.

2) Random Forest: Random forest analysis is another machine learning classification method that is often used in customer churn analysis. The method operates by constructing multiple decision trees and constructing models based on summary statistics of these decision trees.

3) Logistic Regression: Our final statistical method will be logistic regression, a more classic method compared to the two above machine learning based methods. Logistic regression involves regressing predictor variables on a binary outcome using a binomial link function.

III. Results and Discussion

A. Decision Tree

Exploratory method will identify the most important variables related to churn in a hierarchical format.

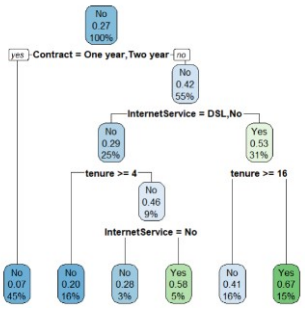

Fig. 1. Random Forest Model

Key Findings:

Contract Type: The most important factor. Customers with month-to-month contracts are more likely to leave. Internet Service: Customers with DSL internet service are less likely to leave. Tenure: Customers who have stayed for more than 15 months are less likely to leave. Accuracy: 79.79

B. Random Forest

We will begin by identifying the number of variables randomly sampled as candidates at each split of the algorithm. In the randomForest package, this is referred to as the 'mtry' parameter or argument.

Key Findings:

Contract Status: Like the decision tree, contract status is a key predictor of whether customers will leave. Tenure Length: How long a customer has been with the company is also an important factor. Total Charges: This variable is now highlighted as very important in predicting churn. Internet Service: Unlike in the decision tree, internet service is not as important in this model. Accuracy: 80.36

C. Logistic Regression

Let's fit the model using the base general linear modeling function in R, glm.

Key Findings:

Important Predictors: Tenure length, contract status, and total charges have the lowest p-values, making them the best predictors of customer churn. Accuracy: 81.36

D. Data Visualization Based On Models

Our modeling efforts pointed to several important churn predictors: contract status, internet status, tenure length, and total charges. Let's examine how these variables split by churn status.

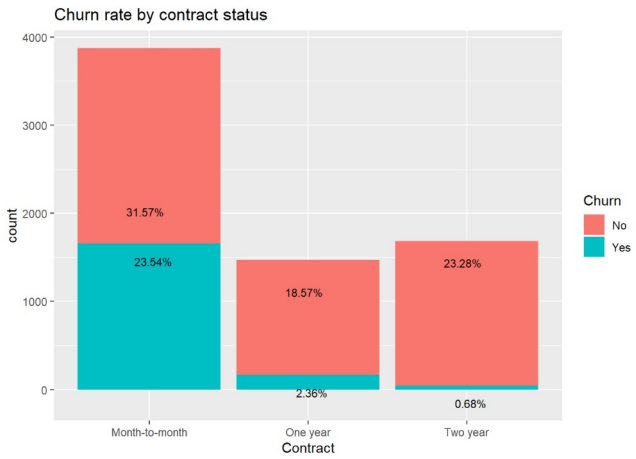

Fig. 2. Contract

1) We will begin with the contract status variable.: As would be expected, the churn rate of month-to-month contract customers is much higher than the longer contract customers. Customers who are more willing to commit to longer contracts are less likely to leave.

2) For the internet service status of the customer?: It appears as if customers with internet service are more likely to churn than those that don't. This is more pronounced for customers with fiber optic internet service, who are the most likely to churn.

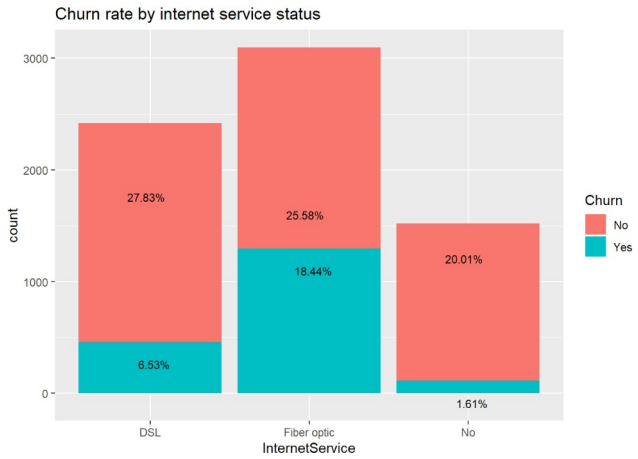

Fig. 3. Internet Service

3) churn split for the tenure length distribution.: Similar to the tenure trend, customers who have spent more with the company tend not to leave. This could just be a reflection of the tenure effect, or it could be due to financial characteristics of the customer: customers who are more financially well off are less likely to leave.

IV. Conclusion

After going through various preparatory steps including data/library loading and preprocessing, we carried out three statistical classification methods common in churn analysis. We identified several important churn predictor variables from these models and compared these models on accuracy measures.

Summary of our findings:- 1. Customers with month-to-month contracts are less likely to churn. 2. Customers with internet service, in particular fiber optic service, are more likely to churn. 3. Customers who have been with the company longer or have paid more in total are less likely to churn.

Acknowledgment

We are grateful to Electronics and Communication department of Amrita School of Engineering, Amritapuri Campus, Kollam, India for providing us all the necessary lab facilities and support towards the successful completion of the project.