CHURN PREDICTION IN TELECOMMUNICATION : a logistic regression approach

SAMREEN NAQVI
20MIY0035
School of advanced Sciences
Vellore institute of technology
Vellore India
samreen.naqvi2020@vitstudent.ac.in

GUIDE:
Prof. Neelabja Chatterjee
Assistant professor Grade 2
School of Advanced Sciences
neelabja.c@vit.ac.in

Abstract— **Churn prediction in the telecommunications industry is crucial for retaining customers and maximizing revenue. This project employs a logistic regression approach to forecast customer churn based on historical data and relevant features such as usage patterns, demographics, and service interactions. By analyzing a large dataset of customer information, including call records, service subscriptions, and customer demographics, the logistic regression model aims to identify key factors contributing to churn and predict which customers are likely to churn in the future. The findings of this study can inform telecom companies' strategies in implementing targeted retention efforts and improving customer satisfaction.**

Keywords— ***Churn prediction, Logistic regression, Historical data, Usage patterns, Demographics, Service interactions, Customer satisfaction***

I. LITERATURE REVIEW

Customer churn prediction is a critical task within the telecommunications industry, aimed at retaining customers and sustaining profitability. Various studies have investigated churn prediction using diverse machine learning and statistical techniques. For instance, Wang and Yao (2017) employed algorithms such as logistic regression, decision trees, random forests to predict customer churn in a Chinese telecom company, highlighting the importance of algorithm selection. Similarly, Verbeke et al. (2014) conducted a comparative analysis of logistic regression, neural networks, and decision trees for churn prediction in a European telecom company, with neural networks demonstrating superior performance. Feature selection and engineering are fundamental aspects of churn prediction models, often involving customer demographics, usage patterns (e.g., calls made, data usage), and account information. Li et al. (2018) proposed a feature selection method based on mutual information and decision trees to identify relevant features for churn prediction, emphasizing the significance of feature engineering. Moreover, addressing class imbalance is crucial in churn prediction tasks, with techniques like oversampling (e.g., SMOTE) and undersampling being commonly employed (Chawla et al., 2002). Finally, model evaluation metrics such as accuracy, precision, recall, and ROC-AUC play a vital role in assessing the effectiveness of churn prediction models, as emphasized by Provost and Fawcett (2013).

II. INTRODUCTION

Customer churn remains an enduring obstacle encountered by companies in the telecommunications industry, significantly impacting business sustainability and profitability. In response to this issue, the development of effective churn prediction models has become imperative for anticipating customer attrition accurately. This paper focuses on employing logistic regression as an approach to churn prediction, aiming to identify key factors contributing to churn and construct a predictive model that aids telecom companies in proactively retaining customers. By leveraging historical data and relevant features such as usage patterns, demographics, and service interactions, the logistic regression model endeavors to forecast customer churn, thereby enabling companies to implement targeted retention efforts and enhance overall customer satisfaction. Through an in-depth exploration of the methodology and algorithm used, as well as insights derived from the literature review, this paper aims to contribute to the advancement of churn prediction strategies within the telecommunications domain.

III. METHODOLOGY

1. Data Collection:

The dataset utilized in this study was obtained from Kaggle. The dataset contains 243,553 rows of Customer data from four prominent telecom partners of India: Airtel, Reliance Jio, Vodafone, and BSNL. The dataset includes various demographic, location, and usage pattern variables for each customer, as well as a binary variable indicating whether the customer has churned or not.

2. Data Preprocessing:

- Initially, the dataset underwent thorough preprocessing to ensure data quality and integrity. This process involved:
- Performing data cleaning to rectify missing values and outliers, and inconsistencies.
- Standardization or normalization of numerical features to ensure uniform scale.
- Encoding categorical variables using techniques such as one-hot encoding or label encoding.

3. Feature Engineering:

- Feature engineering was conducted to extract relevant information and create new features that could enhance the predictive performance of the model. This step included:
- Selection of pertinent features based on domain knowledge, literature review, and correlation analysis.
- Creation of new features through techniques such as binning, interaction terms, and polynomial features.

4. Model Development:

- Logistic Regression, a widely-used binary classification algorithm, was chosen as the predictive modeling technique for churn prediction due to its interpretability and effectiveness in handling categorical features. The following steps were undertaken for model development:
- Splitting the dataset into training and testing sets with a ratio of [e.g., 80:20].
- Training the logistic regression model on the training set using iterative optimization algorithms such as gradient descent.
- Tuning hyperparameters, such as regularization strength, using techniques like cross-validation to prevent overfitting.

5. Model Evaluation:

- The performance of the logistic regression model was assessed using various evaluation metrics, including:
- Accuracy: the proportion of correctly predicted instances.
- Precision: The proportion of correctly predicted positive instances out of all predicted positive instances.
- Recall: The proportion of correctly predicted positive instances out of all actual positive instances.
- F1-score: the harmonic mean of precision and recall, providing a balanced measure of the model's performance.
- Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) to visualize the trade-off between true positive rate and false positive rate across different thresholds.
- Confusion matrix to analyze the classification results in detail, including true positives, true negatives, false positives, and false negatives.

6. Interpretation and Insights:

- The coefficients of the logistic regression model were interpreted to understand the relative importance of features in predicting churn.
- Key insights regarding customer behavior and characteristics influencing churn were derived from the analysis of model coefficients and feature importance.

7. Software and Tools:

- All data preprocessing, feature engineering, model development, and evaluation tasks were performed using Python programming language, along with libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib.

8. Ethical Considerations:

- This research adheres to ethical standards regarding data privacy and confidentiality. No personally identifiable information was used, and all data were anonymized before analysis.

This methodology outlines the systematic approach adopted to develop and evaluate the logistic regression model for churn prediction.

IV. ALGORITHM

In this study, logistic regression, a widely-used statistical method for binary classification, was employed as the primary algorithm for predicting customer churn in the telecommunication industry. Logistic regression is particularly well-suited for this task due to its ability to model the probability of a binary outcome (churn or non-churn) based on a set of predictor variables. The logistic regression algorithm works by fitting a logistic function to the linear combination of predictor variables, transforming the output into a probability score bounded between 0 and 1. The logistic function, also known as the sigmoid function, maps any real-valued input into a value between 0 and 1, representing the probability of the positive class (churn in our case).

Mathematically, the logistic regression equation can be expressed as:

$$P(Y=1 \mid X) = 1 / (1+e^{-(\beta0+\beta1X1+\beta2X2+...+\beta nXn)}) \qquad (1)$$

Where:

- $P(Y=1 \mid X)$ represents the probability of the positive class (churn) given the predictor variables X.
- β0,β1,...βn are the coefficients of the logistic regression model.
- X1,X2,...,Xn are the predictor variables.
- e is the base of the natural logarithm.

The logistic regression model estimates the coefficients β0,β1,...,βn during the training process using iterative optimization techniques such as gradient descent. These coefficients represent the impact of each predictor variable on the probability of churn. To make predictions, the model calculates the probability of churn for each observation in the dataset and classifies it as churn (1) if the probability exceeds a predefined threshold, otherwise as non-churn (0). The model's performance is evaluated using various metrics such as accuracy, precision, recall, F1-score, and receiver operating characteristic (ROC) curve. Overall, logistic regression provides a interpretable and effective framework for predicting customer churn, enabling telecommunication companies to identify at-risk customers and devise targeted retention strategies.

V. INFERENCES

a) Model Performance:

The logistic regression model achieved an accuracy of approximately 48.69% on the test set, indicating its capability to correctly classify customers' churn status nearly half of the time. The confusion matrix reveals insightful details regarding the model's predictions. Specifically, it accurately identified 18,565 non-churned customers (True Negatives) and 5,150 churned customers (True Positives). However, the model also misclassified 20,363 non-churned customers as churned (False Positives) and 4,633 churned customers as non-churned (False Negatives).

b) Classification Report:

Precision, recall, and F1-score metrics offer a comprehensive evaluation of the model's performance. Notably, the model demonstrates superior predictive ability for non-churned customers (class 0) compared to churned customers (class 1), as evidenced by higher precision, recall, and F1-score for class 0.

c) Significant Predictors:

Analysis of the logistic regression coefficients reveals valuable insights into the factors influencing churn prediction. Features with higher absolute coefficient values, such as 'data_used', 'city_Hyderabad', and 'sms_sent', exert a notable influence on predicting churn. However, it is imperative to note that coefficients alone do not elucidate the magnitude of the effect or causality direction, necessitating further analysis or domain expertise for accurate interpretation.

VI. CONCLUSION:

While the logistic regression model exhibits some predictive capability for churn prediction based on the provided features, its performance remains modest, leaving considerable room for enhancement. Notably, features such as data usage, city (particularly Hyderabad), and SMS usage emerge as significant predictors of churn, suggesting that customers with higher data utilization and engagement in specific locations may be predisposed to churn. To augment model performance, avenues for improvement include feature engineering, exploration of alternative algorithms, and addressing factors such as data quality and model hyperparameters. Furthermore, leveraging domain-specific insights could facilitate a deeper understanding of the results and aid in devising actionable strategies for effective churn prevention.

ACKNOWLEDGMENT

I extend my heartfelt gratitude to Professor Neelabja Chatterjee, Assistant professor Grade 2, for his invaluable guidance and support throughout this research project. His expertise and insightful feedback have been instrumental in shaping the direction of this study. I also thank the department of School of Advanced Sciences, VIT Vellore, for providing resources.

REFERENCES

[1] Wang, Q., & Yao, H. (2017). Churn prediction in telecom using machine learning in big data platform. 2017 IEEE International Conference on Big Data (Big Data), 4251-4258.

[2] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2014). New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach. European Journal of Operational Research, 231(2), 356-369.

[3] Li, L., Jiang, G., Zhou, X., Zhao, Z., & Wu, H. (2018). Feature selection methods for customer churn prediction: A case study in the telecom industry. 2018 IEEE International Conference on Big Knowledge (ICBK), 84-91.

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[5] Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc.