Customer Loyalty Prediction in Telecommunications Industry: A Comparative Analysis of Machine Learning Models

Ikramuddin Ahmed
Indiana University

Deep Himmat Gori
Indiana University

Shivani Milind Latkar
Indiana University

project-ikahmed-deepgori-shlatkar

Abstract

The increasingly competitive telecommunications industry in the present day world has caused each network provider in the market to spend millions of dollars in research for predicting churn rates and devising targeted methods to boost their customer retention. Retaining loyal customers has become one of the prime objectives of these companies in order to survive, grow and increase their profits. However, predicting customer churn rate is a complex task which involves analyzing the effect of multiple factors. Our project aims to address this challenge by leveraging a comprehensive range of machine learning classification models to predict customer loyalty. We evaluate the following machine learning models: K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVM), Random Forest, Soft and Hard Voting Classifier (Ensemble techniques), and Gradient Boost. A Telecom customer dataset has been obtained from Kaggle platform and each model will be trained, tested and its performance will be analyzed thoroughly. The results from this project would give the best optimized machine learning model that could be scaled and utilized by companies for maximising their customer retention.

Keywords

Customer Churn Rate, KNN, Logistic Regression, SVM, Random Forest, Soft Voting Classifier, Hard Voting Classifier, Gradient Boost, Ensemble Techniques.

1 Introduction

With high speed 5G network becoming widespread, customers have become very prudent in selecting a network carrier with the best range and bigger data-packs at cheaper costs. The Telecommunication Industry today has multiple big players in every country and they have been competing with each other to provide the best service plans. Customer Churn is defined as the percentage of customers that have stopped using a company's service. The Telecom industry is known to have a high average churn rate of about 25 to 30 percent.

There can be multiple causes for the increased churn rates such as poor customer service, bad products, high prices or improperly catering the needs of different demographics. Telecom companies with less loyal customers have to allocate huge proportions of their budget to new customer acquisition methods such as advertisements which often drives them into loss. Predicting what factors increase customer loyalty and retention is always a better and cheaper alternative for these companies. Machine Learning model is an extremely beneficial tool in analyzing big telecom datasets and making customer churn predictions on them. Customer Churn Prediction falls under the classification category in machine learning and we would be evaluating different classification models and selecting the best one in this study.

Previous work

In [1], the authors analyzed the use of machine learning models in hospitality venues and described the advantages. The authors in [2] obtained a logistic regression model which predicted customer retention with 95.5 percent accuracy. The authors in [3] worked on a Syrian telecom dataset and introduced a novel approach for customer segmentation called Time-frequency-monetary (TFM) and defined loyalty for each segment. In [4], the authors developed a customer online behaviour analysis tool which integrates pricing data and customer segmentation to analyze purchase behaviour and perform predictions.

2 Methods

The first step is to collect a comprehensive database of customer history for a telecom company. It should contain multiple customer attributes such as demographics, subscription details, service usage, billing information, customer interactions etc. The next step is to preprocess the data by handling missing values and outliers, perform feature selection or creation to get features that best capture the important trends related to customer loyalty, and also normalize and scale these features . We next select the following models: KNN, Logistic Regression, SVM, Random Forest. The data is then split into training and testing datasets to asses the model's performance accurately, and then each model is trained on the training data. Hyperparameter tuning is done in order to get the most optimal model for each classifier. Techniques such as grid search or random search are utilized to identify the best hyperparameters for improved model performance.

The models' performance is evaluated on the testing data using various classification metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve. We also employ ensemble techniques such as Soft and Hard Voting Classifier to combine the predictions of the individual models. The ensemble model's performance is evaluated and compared to the individual models to determine if there is an improvement in prediction accuracy. Finally, we draw conclusions from the results about which models or combinations of models is the most effective for the dataset.

References

[1] McIntyre NH. Aluri A, Price BS. Using machine learning to cocreate value through dynamic customer engagement in a brand loyalty program. *J Hosp Tour Res*, 43(1), 2019.

[2] Adeduro O. Oladapo K, Omotosho O. Predictive analytics for increased loyalty and customer retention in telecommunication industry. *Int J Comput Appl*, 975(8887), 2018.

[3] Salloum K. Wassouf W.N, Alkhatib R et al. Predictive analytics using big data for increased customer loyalty: Syriatel telecom company case study. *J Big Data*, 7(29), 2020.

[4] Wei Y. Wong E. Customer online shopping experience data analytics: integrated customer segmentation and customised services prediction model. *Int J Retail Distrib Manag.*, 56(4), 2018.