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: Logistic Regression, Decision Tree, Support Vector Machines (SVM), Random Forest, AdaBoost, Bagging, and Soft and Hard Voting Classifier (Ensemble techniques). 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, Logistic Regression, SVM, Random Forest, AdaBoost, Bagging, Soft Voting Classifier, Hard Voting Classifier, 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

The authors in [1] used customer social network in the prediction model by extracting Social Network Analysis (SNA) features. This enhanced the performance of their model from 84 to 93.3 percent. In [2], the authors adopted a modified version of SMOTE algorithm called DSMOTE to handle imbalanced data which improved their model's performance. In [3], the authors analyzed the use of machine learning models in hospitality venues and described the advantages. The authors in [4] obtained a logistic regression model which predicted customer retention with 95.5 percent accuracy. The authors in [5] 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 [6], 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 was to collect a comprehensive database of customer history for a telecom company. It contains multiple customer attributes such as demographics, subscription details, service usage, billing information, customer interactions etc. The next step was 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 selected the following models: Logistic Regression, SVM, Random Forest. The data was then split into training, testing and validation datasets to asses the model's performance accurately, and then each model is trained on the training data. Hyperparameter tuning was done in order to get the most optimal model for each classifier. Techniques such as grid search or random search were utilized to identify the best hyperparameters for improved model performance.

The models' performance was evaluated on the testing data using various classification metrics such as accuracy, precision, recall, F1-score, and ROC curve (AUC). We also employed ensemble techniques such as Soft and Hard Voting Classifier to combine the predictions of the individual models. The ensemble model's performance was evaluated and compared to the individual models to determine if there was 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.

2.1 Data Preprocessing

The dataset consisted of 7043 rows and 21 columns. We dropped null values present in the Total Charges feature. The proportion of label 'Churn' and different features can be seen in Figures 1 and 2 respectively. Churn's categories 'True' and 'False' were converted to numeric 1 and 0. Next, we one hot encoded all feature columns with categorical values and also scaled the numeric values. Correlation between the label and all other features was found and plotted as shown in Figure 3. Unnecessary features such as Customer ID and the features having correlation below 0.15 were dropped from the dataset.

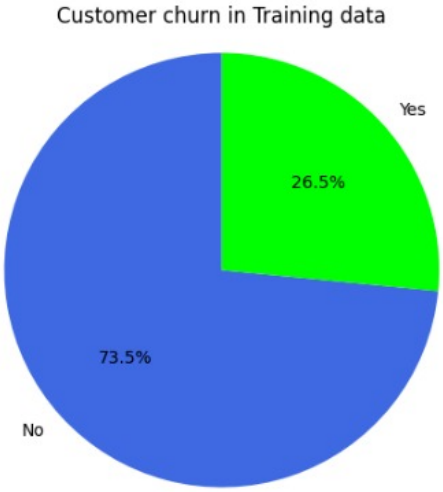

Figure 1: Proportion of Churn

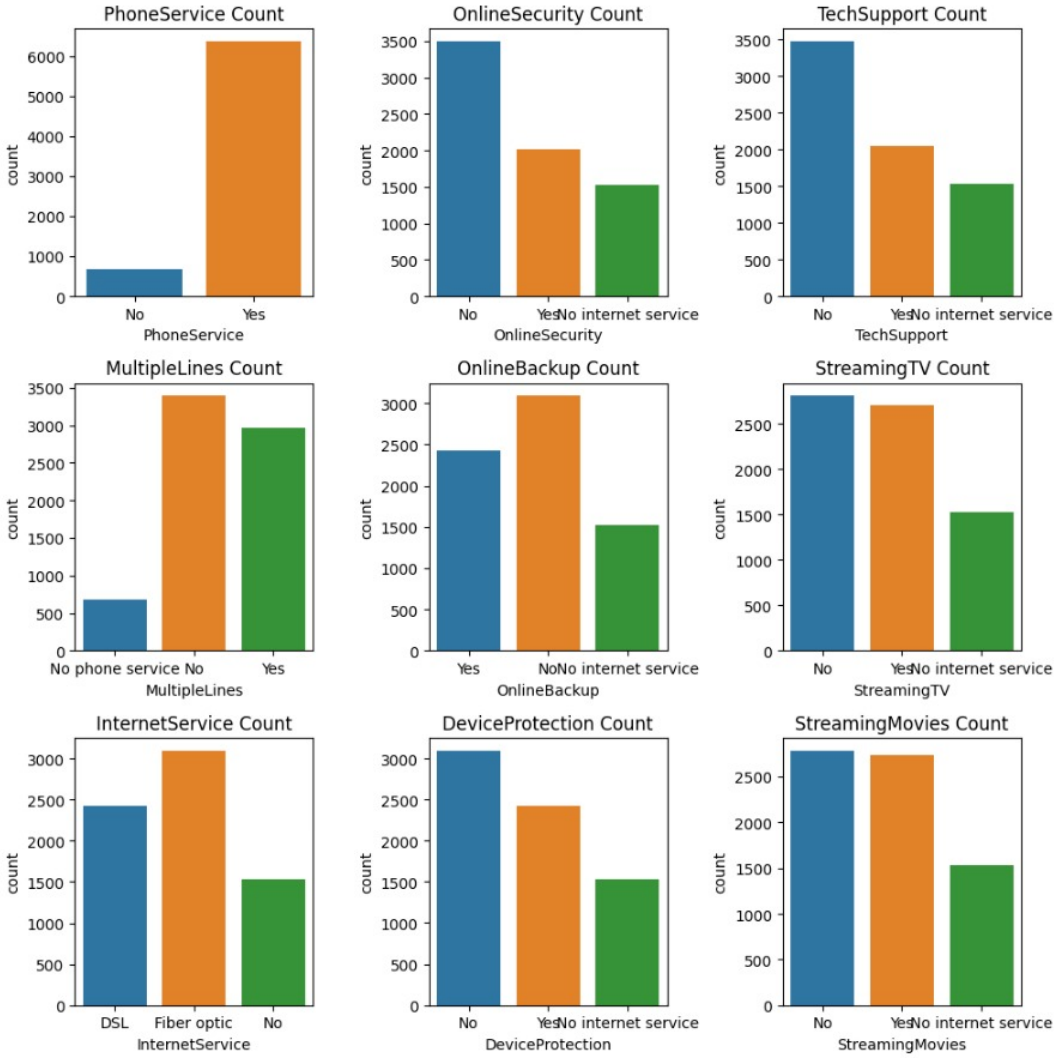

Figure 2: Distribution of different Features

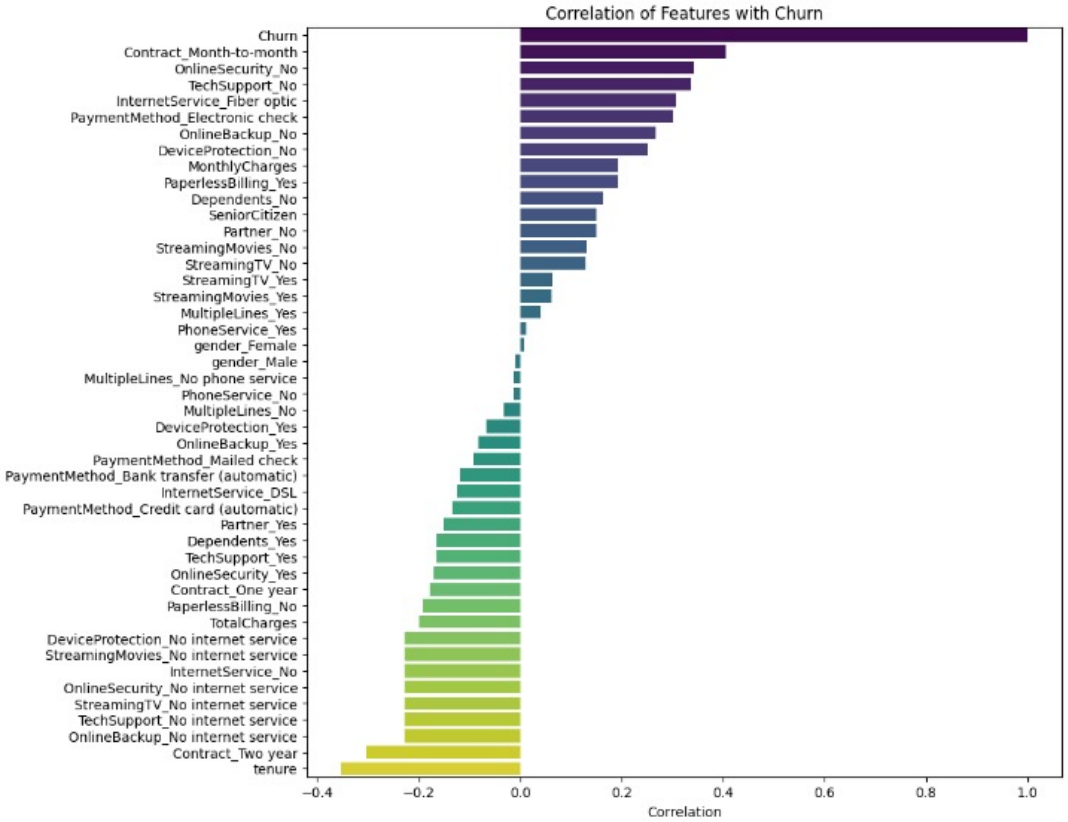

Figure 3: Correlation of Features with Churn

2.2 Model Building

SMOTE (Synthetic Minority Oversampling Technique) was applied to the features 'X' and label 'y' in order to obtain a balanced dataset. It was then split into train, test, and validation data in the ratio 60:20:20. In total, 8 different models were trained and tested on our dataset. We employed the GridSearchCV method to tune and find the best hyperparameters for each type of model. The parameters that were tuned were C, Solver, and max iter for Logistic Regression; C, kernel, degree, and gamma for SVM; and n estimators, max depth, min samples split, and min samples leaf for Random Forest. These three classifiers were fed into an Ensemble and soft and hard voting methods were utilized. The feature importance obtained from Random Forest was plotted as shown in Figure 4.

3 Results

The models' performance was analyzed based on their Accuracy, Precision, Recall, F1-score, and ROC curve. The test performance scores of each type of model can be found in Table 1. Among all models, the Soft-voting Ensemble Classifier performed the best with a test accuracy of 0.8345. Along with this, classifiers like Logistic Regression, Random Forest, and SVM also performed well. However, the Decision Tree, AdaBoost, and Bagging classifiers struggled as compared to the others. The ROC curves for Soft-voting Ensemble, Random Forest and Decision Tree can be found in Figures 5, 6, and 7 respectively.

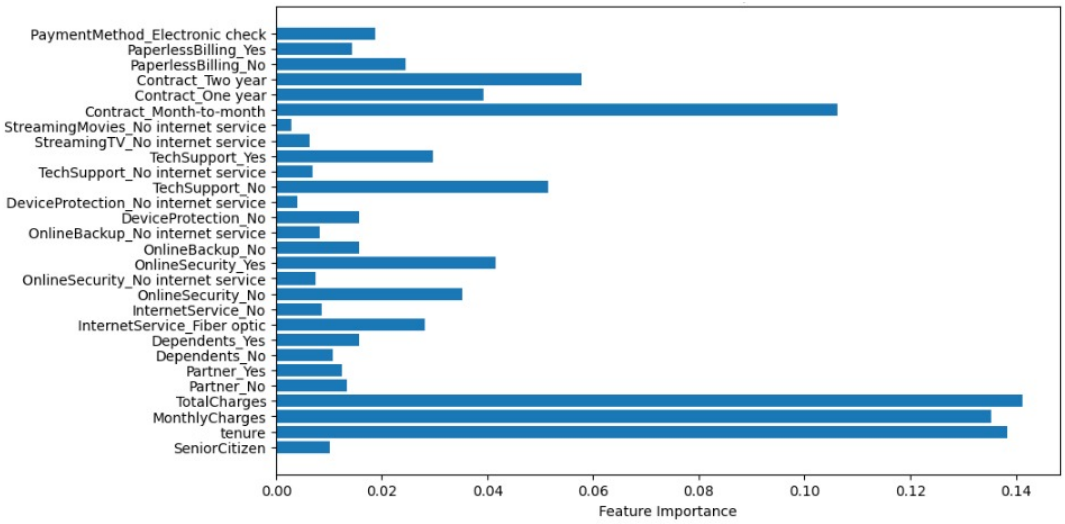

Figure 4: Feature Importance by Random Forest Classifier

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.8258	0.8277	0.8258	0.8255	0.83
Decision Tree	0.779	0.78	0.78	0.78	0.78
SVM	0.8224	0.8248	0.8224	0.8220	0.82
Random Forest	0.832	0.833	0.832	0.8319	0.83
AdaBoost	0.807	0.81	0.81	0.81	0.81
Bagging	0.81	0.81	0.81	0.81	0.81
Hard-Voting Ensemble	0.8262	0.8282	0.8262	0.826	0.83
Soft-Voting Ensemble	0.8345	0.8362	0.8345	0.8343	0.83

Table 1: Performance Scores of Models on Test set

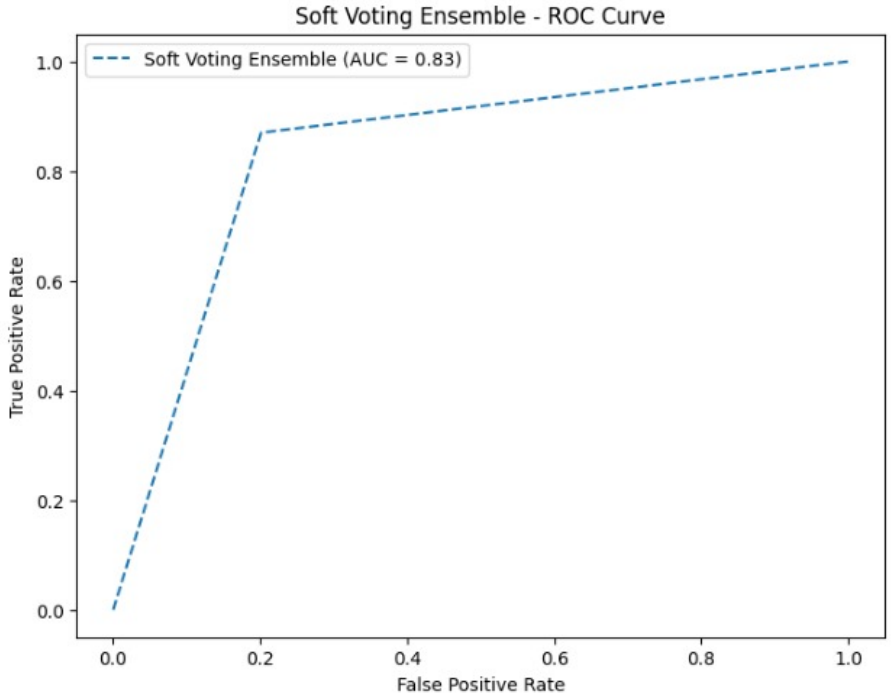

Figure 5: Soft-voting Classifier - ROC Curve

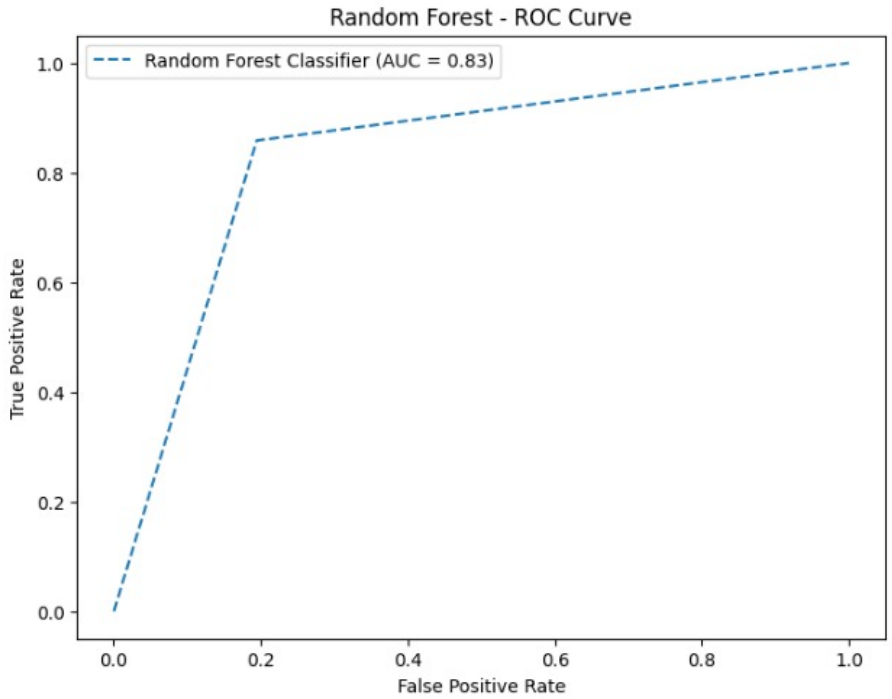

Figure 6: Random Forest Classifier - ROC Curve

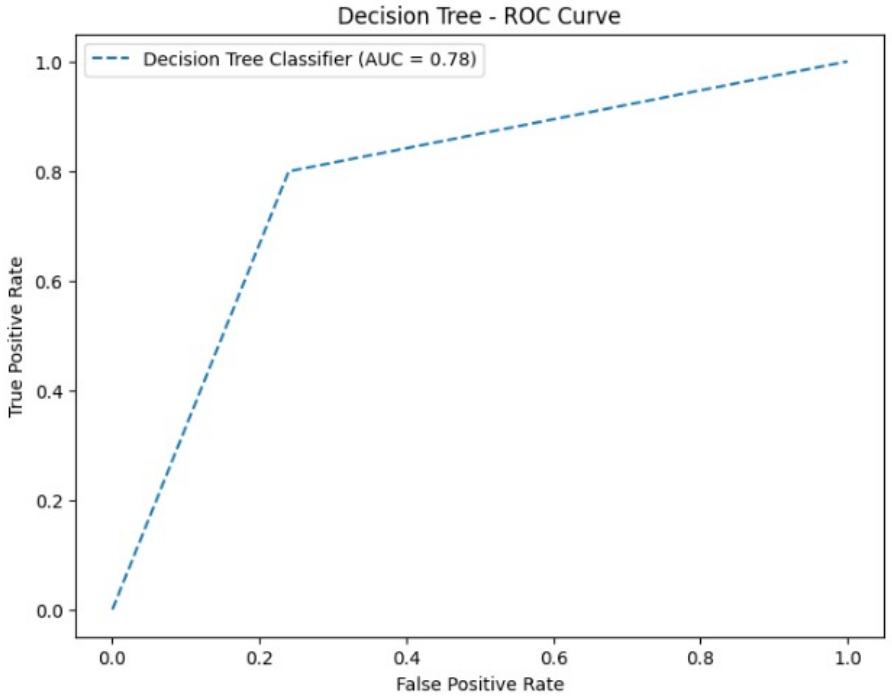

Figure 7: Decision Tree Classifier - ROC Curve

4 Discussion

Based on our results, we have found the soft voting ensemble to be the best classifier for predicting Customer Churn obtaining 83 percent accuracy. This model is not ready to be deployed into the industry, yet it has scope of further improvement in terms of feature extraction, scaling and hyperparameter tuning. A detailed analysis of the features that would impact the Churn rate most and thereby collecting more of this data would help in improving the model's performance. Other machine learning classification models such as GradientBoost, XGBoost etc. can be explored in order to obtain a higher prediction accuracy.

5 Author Contribution

Shivani Latkar undertook the task of reviewing past literature and work done on Customer Churn Predicition, and also performed the pre-processing of the Dataset. **Deep Gori** scoured the web in order to obtain the best dataset for this project, performed data visualization, helped in model building, and prepared the project presentation. Finally, **Ikramuddin Ahmed** preprocessed the dataset, built the models, tuned the hyperparameters, optimized the results, and built the project report.

References

[1] Jafar A. Aljoumaa K. Ahmad, A.K. Customer churn prediction in telecom using machine learning in big data platform. *J Big Data*, 6(28), 2019.

[2] Samaher Al Janabi and Fatma Razaq. Intelligent big data analysis to design smart predictor for customer churn in telecommunication industry. *Springer International Publishing*, pp. 246-272, 2019.

[3] 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.

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

[5] 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.

[6] 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.