Customer Churn Prediction Based on Big Data and Machine Learning Approaches

Ziyu Zhu(✉)

Eli Broad College of Business, Michigan State University, Okemos, East Lansing, USA
zhuziyu3@msu.edu

Abstract. Telecom companies are facing fierce competition in the market. For telecom operators, customers are living. Due to the high upfront investment in acquiring new customers, they prefer to retain existing customers rather than acquire new ones. The loss of old customers means that telecom operators are losing their share in the telecom market. To prevent customer churn, business analysts and customer relationship management (CRM) analysts need to understand and analyze the behavioral patterns of existing customer churn data. The study used three models (i.e., LGBM, Logistic Regression, and Random Forest) to construct a valid and accurate churn prediction model for the telecom industry. Furthermore, the empirical evaluation results suggest that Logistic Regression selected by the AUC metric is the most suitable model. These results provide an understanding of customer features and preferences to predict customers that might to be churn and the reasons for churn and to take preventive measures in advance.

Keywords: Customer Churn Prediction Model · Logistic Regression model · Prediction Algorithms · telecom

1 Introduction

Companies in the telecom industry struggle to acquire new customers, making it very costly for them to lose one. It is always cheaper to maintain an existing customer than acquire a new one in telecom companies. In other words, a company's existing customers are more valuable and more straightforward to market its products than a company that struggles to have new customers. One of the benefits of existing customers is that they can easily use word-of-mouth for marketing their products, which makes them brand ambassadors. There are also benefits to retaining customers, including the fact that customer reach determines a company's success in new markets, helps improve ROI, and serves as an indicator of the quality of service an organization provides. The most critical aspect of customer retention is their satisfaction, trust, and commitment to an organization. Despite the importance of customer retention, it is difficult for companies to maintain 100 percent customer retention, especially telecommunication companies, due to intense competition. The rate at which a company loses customers is known as the churn rate, customer attrition rate, or simply the churn rate, which is also described

Y. Jiang et al. (Eds.): ICFIED 2023, AEBMR 237, pp. 4–15, 2023.
https://doi.org/10.2991/978-94-6463-142-5_2

as the percentage of customers who stop doing business with a company. Churn rates in telecommunication companies are often caused by poor customer service experiences, incomplete service, and unresolved issues. They contribute to significant issues and problems that result in lost revenue and sometimes affect brand image.

The history of this problem is that most companies in the telecom industry tends to lose customers on a daily basis, which reduces their revenues and affects their business. In addition, some of the services offered by telecom companies include cell phone services, Internet, online services, technical support, and other services that require subscriptions. Therefore, companies are eager to learn ways to enhance their customer retention programs. In this pursuit, organizations need to have models that predict customer churn in order to implement methods that can easily promote consumer retention. In other words, organizations and their leaders intend to find a model that predicts customer churn in order to combat this initiative by implementing methods that can boost customer retention.

Previous research indicates that customer churn reduction is critical in reducing organizational losses compared to chasing new customers [1]. According to statistics, the average customer rate for Telco companies in the United States is 1.9% per month across four major companies and is likely to increase to 67% per year when considering prepaid services. Research also indicates that customer churn is a severe issue in the industry. The annual churn rate ranges from 32% to 35%, which means that companies must also have higher retention rates. Data indicates that the customer retention rate in an industry with up to 35% customer churn should be 68% to maintain a higher customer market share [2]. These values clearly show that higher customer churn in a company reduces its revenue and overall performance in the industry. Previous research examining reliable predictive customer churn models shows that the simple value model helps understand the problem to help companies design better customer retention plans that also boost customer satisfaction [3]. The use of the value model is based on determining the customers with the most significant value in an organization, which in turn helps create retention plans that help companies retain more customers. In further research, the authors indicated that there is a need for telecom companies to understand factors contributing to customer churn to take necessary steps to reduce this problem [4]. Using machine learning procedures on big data helps develop a model that predicts customer churn in these organizations. Previous research applying machine language in developing a model that predicts customer churn proposed using the Area under Curve model resulted in a 93.3% value. It applied features of Social Networking Analysis (SNA) to boost the performance of the model's performance.

There are other machine learning and visualization models useful in customer churn prediction. A good example is HEAVY.AI, an accelerated analytic model that can process big data sets from customers [5]. Studies indicate that this model is trusted by big Telco companies, which enables them to conduct customer churn analysis with analytics and uses the results to spot irregularities, manage a fleet, and improve reliability.

According to research, the churn prediction problem involved models that only predicted customer churn for the current or next month [6]. Models that predict churn for a short period are ineffective due to a lack of sufficient time for organizations to develop and implement strategies to retain those customers. Therefore, the study indicated using

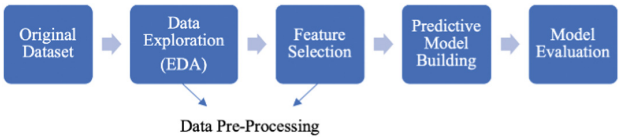

Fig. 1. Step model for predicting the customer churn framework.

a new T + 2 churn customer prediction model [7]. The beauty of the new T + 2 model is that it predicts customer churn for two months, leaving the Telco companies enough time to develop and implement customer retention strategies. The focus of this paper is to examine a possible customer churn prediction model for Telco companies that is useful in depicting the results of the customer churn prediction model. The research achieves the desired objective through the data and method section, results and discussion, explanation of imitations, and a conclusion. The data and method section contains information and explanation of various models including LightGBM (LGBM), followed by logistic regression and random forest models.

Figure 1 shows the pictorial representation of research architecture. It consists of various stages, i.e., data exploration, feature selection, model building, and evaluation. This includes the various stages of the proposed model. It comprises five stages, i.e., Stage 1: Identifying the most pertinent data. (Using EDA) Analysis of variance and correlation matrix and dealing with missing values by estimation and deletion. Stage 2: Feature selection. Stage 3: Development of prediction models (LGBM, logistic regression, and Random Forest). Finally, the prediction models are evaluated (using accuracy and AUC curves). The rest part of the paper is organized as follows. The Sect. 2 will introduce the data, algorithms, and models used in the research. The Sect. 3 will explain the results obtained from the analysis. Eventually, a brief summary will be given in Sect. 4.

2 Methodology

In this study, this paper uses several models to predict the churn behavior of Telco operators' retail customers. This section presents the sources and analysis of the sample data. Additionally, this section described algorithms, the methodology basis of the model, and the evaluation criteria as well as used them to analyze the method's performance.

2.1 Data

The data obtained for the study was hypothetical and is based on information from the Kaggle website [8], which was collected from IBM data set collection. The data analysis and collection begin with importing using the build-up basic libraries. Some of the unique features of the data are that it contains features that make it easy for this paper to make inferences. The raw data in this study contains 7043 unique values representing customers and features represented in 21 columns. The sample data has 19 columns of independent variables describing clients' features. It also contains a churn column, a response variable indicating whether customers departed the company's services in the past month or stayed.

There are two types of variables considered in this study: dependent and independent variables. The independent variables include customer location, lifetime value of a customer, services, and monthly charges. The dependent variables are internet devices, online security, product dissatisfaction, competitor offering superior devices, and month-to-month contracts (IBM Community).

The dependent and independent variables used in this study provide an overview of the situation at the telco company. Data included in the data set include the number of customers that have already left the company, those who remained, and those joining the firm. In addition, there will be elementary scores to be used in the analysis, including the Customer Lifetime Value (CLTV) index, churn score, and satisfaction score. These scores depend on the reasons why customers left the company. Reasons for leaving the company further depend on factors, e.g., the ability of the company to offer better devices and superior services than the competition. Online security is another critical variable that dictates the churning of customers. When customers feel that their online safety is not taken seriously by the company, they begin to churn. Online security as a dependent variable further affects customers' perception of other products and services offered by the company. It may lead to another variable known as product dissatisfaction. Product dissatisfaction is a variable that depends on the quality of products offered by the telco company.

2.2 Methods

2.2.1 Exploratory Data Analysis

Exploratory data analysis (EDA) explores the underlying features present in the data by performing visualization, summarization, and analysis the data to maximize the understanding of the data set and thus discover missing values. EDA aims to discover possible variables from the database that describe or distinguish customer behavior. EDA can also perform bivariate analysis to find the relationship between two variables. In addition, it can find correlations between different features and thus remove redundant features.

2.2.2 Light Gradient Boosting Machine

This model, known as a light gradient boosting machine (LGBM), has different features, including the high accuracy and fastest training speed compared to others. The model uses a machine-learning algorithm that makes it faster data processing [9]. Additionally, the LGBM algorithm is capable of horizontal growth and helps reduce depicting areas of growth where customers are likely to leave. The step involved in LGBM is that it relies on the application of discrete-continuous feature values that it utilizes a histogram algorithm where the discrete values appear in the bins. One finds the optimal cut-point through analysis of all presented features and data.

2.2.3 Logistic Regression

Logistic regression is a widely used statistical modeling technique commonly used to solve binary classification problems [10]. Logistic regression is used to predict a dependent data variable through the analysis of the relationship between one or more existent

independent variables. The model can be written as.

$$P(Y = 1|X) = \frac{1}{1+e^{-\sum_{i=1}^{n} \beta_i X_i - \beta_0}} \quad (1)$$

where $P(Y = 1|X)$ is the probability function predicting a positive outcome, y is a binary dependent variable representing whether things happen or not (if Y = 1, it means things happen; otherwise, Y = 0), $\beta_0 \sim \beta_n$ describe the different regression coefficients, $X_1 \sim X_n$ are the independent value that affects the dependent variable. In addition, logistic regression is essentially sigmoid. The logistic regression results appear in binary form, i.e., "yes" or "no," rather than as numbers, making it easy to plot similar curves. The first step in logistic regression analysis is to pre-process the data. Subsequently, logistic regression donates existing customer data, which allows the company to find or predict customers who may stop being served. Lastly, a confusion matrix and visualization of the test set results, and analysis are provided.

2.2.4 Random Forest

Random forests, consisting of many individual decision trees, operate as an ensemble. By "bagging" or "bootstrap aggregation," the random forest algorithm produces a "forest" that can be trained [11]. By forecasting the decision trees, the random forest can determine the outcomes by extracting and predicting the mean or average of the various trees. The accuracy of the results can be raised by increasing the number of the trees. Therefore, the accuracy of the outcomes can be enhanced by raising the number of trees.

2.2.5 Metrics

This part introduces the evaluation metrics used in this paper, precision, recall, f1 score, accuracy, confusion matrix, and AUC. These evaluation metrics can help quantify the performance of a predictive model.

First, the results of the empirical study are assessed through the most advanced evaluation metrics (i.e., precision, recall, and f1 score,). Precision indicates how many of the predicted churned customers will churn. The recall is the probability of finding an authentic churn from the customer base. The F1 score can be used as a composite metric to evaluate classification performance. Accuracy is the ratio of all correctly predicted quantities. For binary classification, the accuracy can also be calculated positively and negatively.

The mathematical formulas of these evaluation metrics are presented below. More details can be obtained from the study mentioned in [12].

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1score = \frac{2(Precision \times Recall)}{(Precision+Recall)} \quad (4)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Here, TP means True Positives, TN means True Negatives, FP means False Positives, and FN means False Negatives.

The four confusion terms are performance parameters, and the results are usually displayed through the confusion matrix. The confusion matrix is binary classification where the actual values and the predicted values do not lie on the same axis:

- True Positive (TP): Both predicted and actual are positive
- True Negative (TN): Both predicted and actual negative
- False Positive (FP): Predicted positive, actual negative
- False Negative (FN): Predicted negative, actual positive

After knowing these four values, the true-positive rate (TPR), false-negative rate (FPR), true-negative rate (TNR), and false-negative rate (FNR) can be calculated. For the confusion matrix, the data values of TPR and TNR should be as high as possible, and FPR and FNR should be as low.

$$TPR = \frac{TP}{ActualPositive} = \frac{TP}{TP+FN} \tag{6}$$

$$FNR = \frac{FN}{ActualPositive} = \frac{FN}{TP+FN} \tag{7}$$

$$TNR = \frac{TN}{ActualNegative} = \frac{TN}{TN+FP} \tag{8}$$

$$FPR = \frac{FP}{ActualNegative} = \frac{FP}{TN+FP} \tag{9}$$

AUC was used to evaluate the predictive models, which is the key to analyzing model capabilities to determine the best high-performance model for the dataset. AUC is the area under the ROC curve, which is often used to measure the quality of a probabilistic classifier [13]. The value of AUC represents the performance of the model. The perfect classifier will have a solid, valid positive rate and a low false-positive rate (AUC = 1). The formula for calculating AUC is below:

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1} \tag{10}$$

where n_0 means the number of positive examples and n_1 means the number of negative examples. Also, $S_0 = \sum r_i$, where r_i is the rank of the i th positive example in the ranked list [14].

3 Results and Discussion

Table 1, Table 2, and Table 3 illustrate the evaluation measures for the three models. The tables clearly shows the Precision, Recall, F1 score, and accuracy for each model; In the table, the data are divided into "True" and "False," with "True" representing churn and "False" being the opposite.

For the "False" part, Logistic Regression has the highest precision with 0.93. While Random Forest has the lowest precision, the recall has the highest with 0.92. The F1

Table 1. LGBM Classifier

	Precision	Recall	F1-Score	Support
false	0.88	0.79	0.83	1697
true	0.56	0.72	0.63	628
accuracy			0.77	2325
macro avg	0.72	0.75	0.73	2325
weighted avg	0.80	0.77	0.78	2325

Table 2. Logistic Regression

	Precision	Recall	F1-Score	Support
false	0.93	0.72	0.81	1697
true	0.53	0.84	0.65	628
accuracy			0.75	2325
macro avg	0.73	0.78	0.73	2325
weighted avg	0.82	0.75	0.77	2325

Table 3. LGBM Classifier

	Precision	Recall	F1-Score	Support
false	0.82	0.92	0.87	1697
true	0.67	0.46	0.54	628
accuracy			0.79	2325
macro avg	0.75	0.69	0.70	2325
weighted avg	0.78	0.79	0.78	2325

scores of the three models are not significantly different, 0.83, 0.81, and 0.87, respectively. For the "True" part, Random Forest has the highest precision but the lowest recall. This result is the opposite of the "False" part, where the LGBMClassifier has the highest precision, and the Logistic Regression has the highest recall. These two models have similar F1 scores, but the F1 score of the Random Forest is 0.54. Accuracy is the most intuitive measure of performance. In Tables 1, 2 and 3, Random Forest has 79% accuracy, the highest among the three models. The accuracies of the other two mods are 77% and 75%, respectively.

With the help of the confusion matrix, it can not only analyze whether the model is working well even if the data are unbalanced but also calculate other performance metrics. From Figs. 2, 3 and 4, it is easy to see that the True Negative Rate of Random Forest is higher than the other two models, 0.92. Nevertheless, the True Positive Rate is

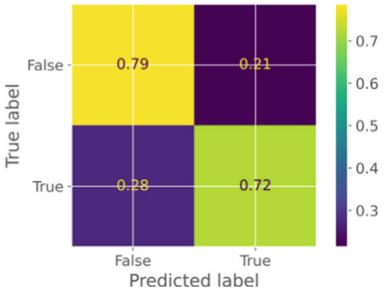

Fig. 2. Confusion Matrix of LGBMClassifier.

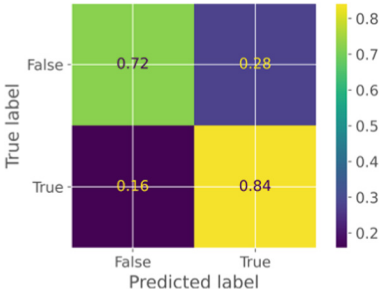

Fig. 3. Confusion Matrix of Logistic Regression.

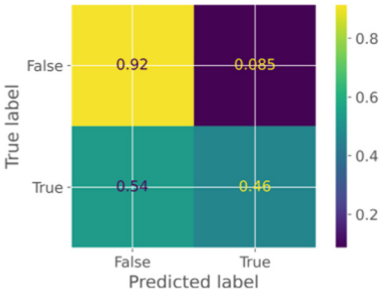

Fig. 4. Confusion Matrix of Random Forest

lower than all the other models, and the TPR of Logistic regression is the highest. For example, the forest has a TPR of 0.46, and Logistic regression has a TPR of 0.84. For the Negative part, the Random Forest has the lowest FNR at 0.083 in comparison, Logistic Regression has an FNR of 0.28. However, the FPR results are the exact opposite of the FNR. The random forest has the highest FPR of the three models at 0.54. In comparison, The FPR of Logistic Regression is only 0.16.

Based on Figs. 5, 6 and 7, it can be determined that the AUC values of LGBMClassifier, Logistic Regression, and Random Forest are 75%, 78%, and 69%, respectively. The AUC value of Logistic Regression is better than the other models and is therefore

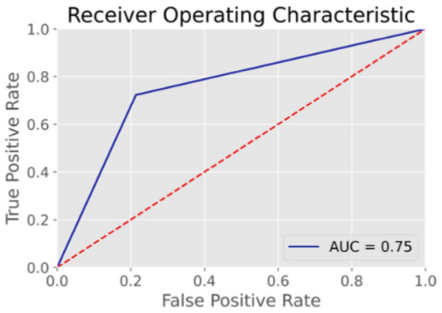

Fig. 5. Area under the ROC curve (AUC) of LGBM

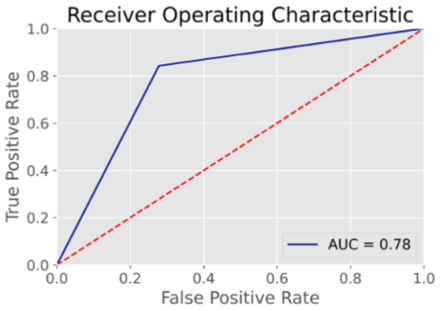

Fig. 6. Area under the ROC curve (AUC) of Logistic Regression

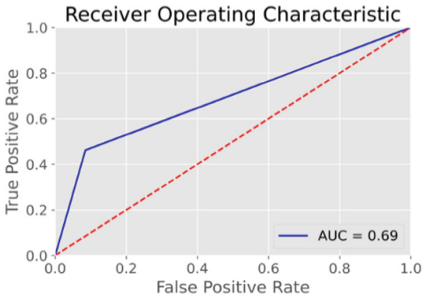

Fig. 7. Area under the ROC curve (AUC) of Random Forest

considered the best classification algorithm for this prediction model. The AUC value of LGBMClassifier occupies the second place. In contrast, the value of Random Forest is ranked last due to the unbalance of the data.

This paper obtained the two best models, which are random forest based on accuracy and logistic regression based on AUC. However, by using arbitrary customer features to judge the performance of the two models, the result of the Random Forest was found to be unbalanced (seen from Fig. 8). On this basis, it indicates that Random Forest does

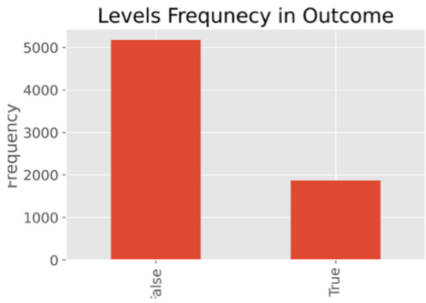

Fig. 8. Random Forest Outcome with arbitrary customer feature.

not apply to all customer features, and this model can only determine the characteristics of churned customers.

After experimenting with the models, this paper looked at the results of each model and determined that the best predictive model based on accuracy and recall was Random Forest. By reconfirming, the use of random forest leads to some data imbalance. Therefore, it can be concluded that Random Forest is more suitable for finding and analyzing the causes of customer churn. Although Random Forest did not identify churned customers well, the model has successfully identified more customers. It is one of the goals of customer churn management. Furthermore, Logistic Regression is the best prediction model based on AUC Curves. This can measure how good a classifier is compared to a random guess. Therefore, it is a suitable model for predicting and identifying the customers about to churn.

Although widely accepted, this model suffers from limitations due to the customer features. Customer features change as customers change. Since the customer features in the dataset only contain features that are present in most customers, it is hard to predict other customers that do not include their features. In the future, other customer features (e.g., dial types, complaint information, and location) should be added to the dataset to increase the accuracy of customer churn prediction. Furthermore, there is an unbalanced classification in this research. To be specific, the unbalanced distribution of the number of people may affect the selection of the subsequent model. Future research needs to do more work to pay attention to the unbalanced classification.

4 Conclusion

In summary, this paper investigates several predictive classifiers based on customer churn. As technology advances, telecom industry services are increasing along with it. Therefore, it is advantageous for telecommunication operators to predict the churn customers and develop retention strategies. This research paper provides the usage of various machine learning techniques to predict customer churn in the telecom industry, i.e., Light Gradient Boosting Machine, Logistic Regression, and Random Forest. According to the analysis, Logistic Regression is the most suitable model based on AUC, and this model is considered ideal for identifying the customers who are about to

churn. Random Forest is the most appropriate model based on accuracy; this model is deemed to be suitable for analyzing the causes of customer churn.

Customer churn prediction is a very complex task. In the future, customer characteristics can be further added to the database to improve the accuracy and precision of prediction. In conclusion, this model will be very effective in helping telecom operators to predict customer churn and improve related services to meet customer demand accurately and efficiently. Overall, these results offer a guideline for customer churn prediction.

References

1. T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory, vol. 55, 2015, pp. 1-9.
2. A. Moreno, "End-to-end machine learning project: Telco customer churn. Medium", 2019. Retrieved 6 May 2022, from https://towardsdatascience.com/end-to-end-machine-learning-project-telco-customer-churn-90744a8df97d.
3. V. V. Saradhi, G. K. Palshikar, "Employee churn prediction. Expert Systems with Applications," vol. 38(3), 2011, pp. 1999–2006.
4. P. Lalwani, M. K. Mishra, J. S. Chadha, P. Sethi, "Customer churn prediction system: a machine learning approach. Computing, vol. 104(2), 2022, pp. 271-294.
5. K. A. Saran Kumar, D. Chandrakala, "A survey on customer churn prediction using machine learning techniques." International Journal of Computer Applications, vol. 975, 2016, 8887.
6. A. Amin, et al., "Customer churn prediction in the telecommunication industry using data certainty," Journal of Business Research, vol. 94, 2019, pp. 290-301.
7. K. Dahiya, S. Bhatia, "Customer churn analysis in the telecom industry." In 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions, vol. 4, pp. 1–6. IEEE, 2015, September
8. BlastChar,"Telco Customer Churn" 2018, Retrieved 18 April 2022, from https://www.kaggle.com/datasets/blastchar/telco-customer-churn/metadata
9. G. Ke, et al. "LightGBM: a highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems, vol. 1, 2017, pp. 3146–3154
10. D. W. Hosmer, S. Lemeshow, Applied logistic regression, Second Edition. New York, John Wiley & Sons, Inc, 2018.
11. M. Onesmus, M, "Introduction to Random Forest in Machine Learning. Section," 2020. Retrieved 6 May 2022, from www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.
12. A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, K. Huang, "Customer churn prediction in the telecommunication sector using a rough set approach," Neurocomputing, 2017, 237
13. D. J. H, R. J. Till, "A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems, Machine Learning," vol. 45, 2001, pp. 171–186
14. H. Jin, C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms", IEEE Transactions on Knowledge and Data Enginerring, vol 17, 2005, pp.299-310