Customer Churn Prediction Based on the Decision Tree and Random Forest Model

Shiyunyang Zhao*

Department of Sciences, University of North Carolina at Chapel Hill, NC, USA

*Corresponding author: zsyy0821@email.unc.edu

Abstract. The rate at which customers discontinue utilizing a company's services during a predetermined time period is known as the customer churn rate, also known as the attrition rate. Hence, developing a prediction model to predict the potential churn customers will generate an early alert for the company to provide them with better service. This study is divided into two main parts: dealing with a dataset about customer behaviors in a bank and building churn prediction models using machine learning algorithms. The data preprocessing part includes dataset description and some adjustments on original dataset to make it accessible for analysis, including deleting unimportant feature and adjusting feature names. Then the study apportions the modified dataset into train set and test set with an 80-20 split. Next, the study imports two kinds of machine learning algorithms, random forest classifier and decision tree classifier, to build churn prediction models. In each model, the study first performs feature selections and visualizes feature importance in bar graphs. Then the study tests each model on testing set and visualizes model performances using confusion matrices and accuracy scores. The results show that both models get most predictions correct while random forest model has a better performance due to its higher accuracy score of 91%.

Keywords: Machine learning; customer churn prediction; decision tree; random forest.

1. Introduction

Customer churn, or customer attrition, is the phenomena where clients of a firm stop doing business with them or making purchases from them. Customer churn can be caused by many factors, such as the success of competitors, rising prices, bad customer service. Mathematically speaking, customer churn rate is a calculation of the percentage of customers who stop using a company's products or services within a given period of time. A high customer churn rate is a poor indication for the company since it suggests that existing customers are leaving the business. A churned customer could be a dissatisfied customer who may negatively impact a company's reputation. In addition to the loss of those churned customers' spending, the company can also get unfavorable press that damages its reputation.

For many businesses, preventing client churn is a major concern. One common way to keep a low customer churning rate is to build a prediction model to look for those potential churning customers, and once those customers are located by prediction model, the company could provide better service to retain those customers [1]. Customer retention is essential to a company's success. One explanation is that acquiring new consumers is more expensive than maintaining existing ones. In fact, increasing customer retention by just 5% can lead to at least a 25% increase in profits [2]. This is due to the likelihood that returning consumers are likely to spend up to 67 percent more money on the company's products and services. As a result, the company can cut back on the operating expenses of having to acquire new customers.

An increasing number of researchers studied about customer churning in recent years and tried to build models to predict churning customers in different business areas. For instance, Li et al. developed a churn prediction model for the cable network industry by combining the customer off-grid forecasting model with customer retention [3]. They investigated and gathered data on possible customer turnover reasons in the cable network sector. This study established that customer spending, payment behaviors, customer preferences, and consumer observation intensity may all be used to assess customer attrition trends. The researchers also came to the conclusion that cable network businesses, as the industry's dominant player in traditional broadcasting and television, should place

a high value on their current clientele and adopt the bell-shaped pricing structure (where prices decrease with increasing viewing intensity). They can simultaneously bring in unique resources and draw in casual clients. In addition, it is also a marketing strategy to analyze customers' viewing preferences and provide them with more convenient payment methods. In addition, Chen et al. performed a case study to analyze churning customers' behaviors without building a prediction model [4]. According to the study, credit card churn exhibits specific tendencies. Customers did not suddenly change from being loyal to frequent customers. When identifying trends in customer behavior, time should be considered. In the past, banks simply looked at one aspect of customer behavior, but this study looks at it in conjunction with demographic data, interactions between the bank and its customers, and time trends. As a result, banks have access to more comprehensive reference data that they may utilize to develop pertinent and practical strategies for actively caring for consumers and mending frayed customer connections.

2. Method

2.1 Data description and preprocessing

2.1.1 Data description

The dataset is collected from Kaggle website [5]. The dataset consists of 23 columns representing 23 variables, including customers' age, salary, marital status, credit card limit, credit card category, etc., and 10,000 rows representing 10,000 customers including current and churned. The features are selected by binary classification, some of which number while others are categories. One thing to be noticed is that the dataset is found to be imbalanced. The data has about 84% normal (non-churn) customer and 16% churned customer. Table 1 shows the related data exploration.

2.1.2 Data preprocessing

Before building the prediction model, some features unrelated to the label are deleted, such as the feature 'Clientnum' which was not helpful for predicting churning customers. Besides, this study converted the categorical features, such as 'Gender' and 'Education' to numeric ones for the convenience of analysis. Next, using 80% of the data for the train set and the remaining 20% for the test set. After splitting the train set and test set, the 'Attrition' column in both sets is deleted to make the rest of columns be the potential features for prediction model.

Table 1. Data Exploration

Customer	0	1
Gender	1	0
Education	3	2
Marital_Status	1	2
Income	2	4
Card_Category	0	0
Attrition	1	1
Age	45	49
Dependent_count	3	5
Months_on_book	39	44
Total_Relationship_Count	5	6
Credit_Limit	12691	8256
Total_Revolving_Bal	777	864
Avg_Open_To_Buy	11914	7392
Total_Amt_Chng_Q4_Q1	1.335	1.541
Total_Trans_Amt	1144	1291
Total_Trans_Ct	42	33
Total_Ct_Chng_Q4_Q1	1.625	3.714
Avg_Utilization_Ratio	0.061	0.105
Naive_Bayes_1	0.000093	0.000057
Naive_Bayes_2	0.99991	0.99994

2.2 Machine learning algorithms

After data slicing, two kinds of machine learning algorithms were employed to obtain the churn prediction models: Random Forest Classifier and Decision Tree Classifier, and the relative performance of these two models were compared based on accuracy scores.

2.2.1 Random forest

A popular supervised machine learning technique for classification and regression issues is random forest. The ability of the random forest technique to cope with both data sets including categorical variables and data sets containing continuous variables, as in regression, is one of its most significant features [6]. Random forest is considered a highly accurate method due to the number of decision trees involved in this process. In addition, it is not affected by overfitting problems because it takes the average of all predictions, which cancels out the bias. Additionally, continuous variables can be handled by random forests by replacing them with median values or by estimating a weighted average of the missing values. The importance of features can be obtained, which helps us to select the features that contribute the most to the classifier.

In the regression modeling process, this study first imported the Random Forest Classifier Scikit-learn and then trained the model on the training set and perform predictions on the testing set. Next, the feature importance on the bar graph was visualized. From visualization, it can be observed the Random Forest Classifier ranked the features in order of feature importance from highest to lowest.

This study also visualized the confusion matrix to get know how many of random forest's predictions were correct, and when incorrect, where the classifier got confused. From the heatmap graph, the model got most predictions correct, which seemed to have impressive performance.

2.2.2 Decision tree

Just like random forest algorithm, decision-tree algorithm falls under the category of supervised learning algorithms that can perform both regression and classification tasks. Its objective is to build a model by learning straightforward decision rules inferred from data properties that predicts the value of the target variable. It is effective for output variables that are categorical as well as continuous [7]. The decision tree technique creates a node for each attribute in the data set, with the root node housing the most crucial attribute. The evaluation process begins at the root node and proceeds down the decision tree along the node that corresponds to the condition or "decision" as appropriate. This process continues up until a leaf node that contains the predictions or outcomes of the decision tree is reached.

This study did the similar regression modeling process using Decision Tree Classifier. We firstly imported the classifier from Scikit-learn and then trained the model on the training set and perform predictions on the testing set. From the feature importance visualization graph, we got the same result as Random Forest Classifier. The confusion matrix showed that the prediction model got 102 predictions incorrect.

3. Results and Discussion

3.1 Feature Importance

This study visualized feature importance in the form of bar graphs shown in Fig. 1 and Fig. 2. The bar graph from random forest classifier showed that the top three important features were total transaction amount, total transaction count and change in transaction count, which means these three features are most predictive of whether a customer will churn or not, while card category, gender and marital status were the least three important features. Decision Tree Classifier got the same result as Random Forest Classifier.

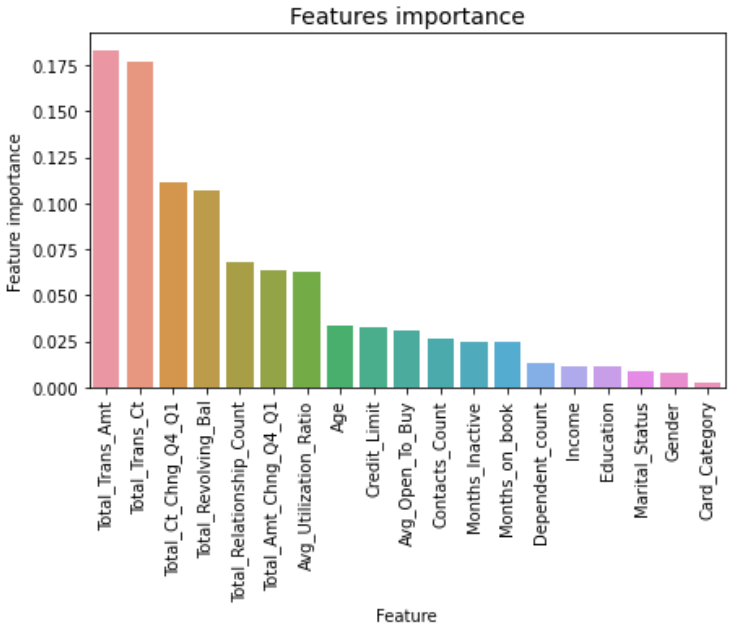

Fig. 1 Feature Importance obtained from the random forest model.

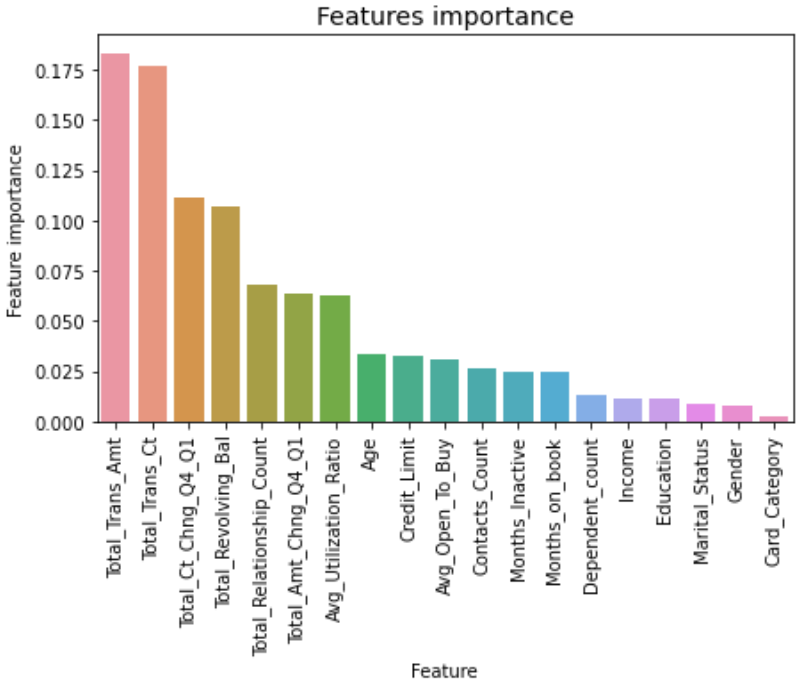

Fig. 2 Feature Importance obtained from the decision tree model

Using feature selection method allow this study to obtain fewer complex models with good performance by reducing irrelevant features. The result of feature importance indicated that those features related to total transactions were closely related to churning customers. Similar feature importance result appears in other studies. For example, AL-Najjar et al. use machine learning algorithms to develop credit card churn prediction [8]. In AL-Najjar et al. prediction model, the top six most important variables include change in transaction count, total transaction count and change in transaction amount.

3.2 Model Performance

The prediction results of the two model were visualized in two confusion matrices. As shown on the confusion matrix (Fig. 3 and Fig. 4), the random forest model got most predictions correct, which

seemed to have impressive performance. In addition, the study calculated an accuracy score of about 91% on the testing set based on the random forest model. For the decision tree model, the confusion matrix showed that the prediction model got 102 predictions incorrect. Though the incorrect predictions were a bit more than the random forest model, the decision tree model had an accuracy score of about 90.8%, which was impressive overall.

Both models adopted in this study can be used for predicting churn customers. According to the confusion matrices and accuracy scores, both models are capable of predicting the churn customers successfully. In addition, the performance of the random forest is superior to the decision tree in terms of the experimental results. The main possible reason is due to the ensembled structure of the random forest. Such a structure is more robust to the noise points and possibly produces better results. To further improve the performance of the model, other models e.g. the K-means or neural network may be considered in the future due to their excellent performance in many tasks [9, 10].

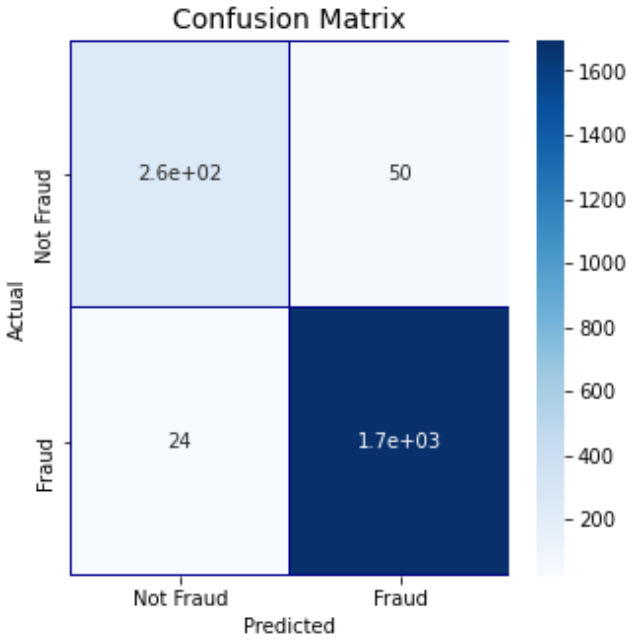

Fig. 3 Random Forest-Confusion Matrix

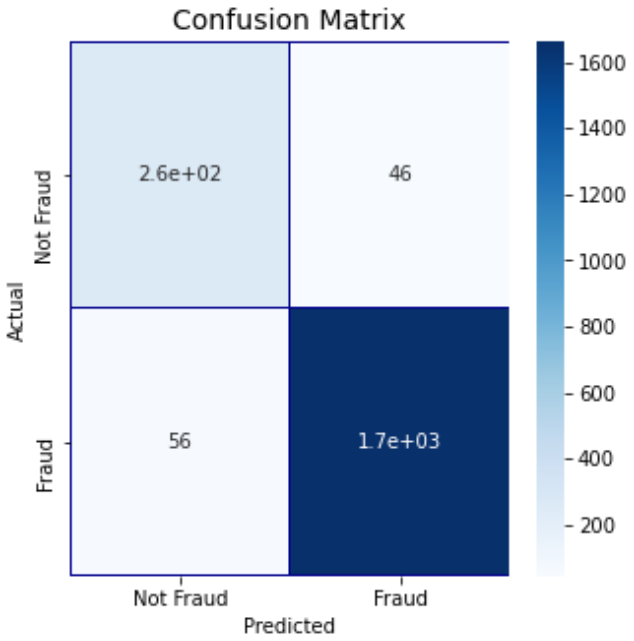

Fig. 4 Decision Tree-Confusion Matrix

4. Conclusion

This study aimed to build models to predict churn customers in a bank so that the bank could retain those customers before high churn rate negatively impact its operation. The study preprocessed the dataset firstly and split the dataset into the train set and the testing set to develop random forest and decision tree algorithms, and the study visualized the feature importance graphs and confusion matrices for both models. According to the feature importance graphs and confusion matrices, the study found that the features related to total transactions were highly correlated with churning customers and that both random forest and decision tree models had accuracy scores around 91% which were impressive overall. One thing which could be improved in this study is to take the imbalanced dataset into considerations by using SMOTE sampling. The dataset was found to be imbalanced with about 84% normal (non-churn) customer and 16% churned customer. Although such distribution is not typically imbalanced, it is still one thing to be considered in future analysis.

References

[1] Qualtrics. What is Customer Churn? Learn how to measure and prevent it, 2023. https://www.qualtrics.com/experience-management/customer/customer-churn/

[2] Hubspot. What is customer Churn? 2021. https://blog.hubspot.com/service/what-is-customer-churn

[3] Li, Yixin, et al. Giant fight: Customer churn prediction in traditional broadcast industry. Journal of Business Research 131, 2021, 630-639.

[4] Cheng, Li Chen, Chia-Chi Wu, and Chih-Yi Chen. Behavior analysis of customer churn for a customer relationship system: an empirical case study. Journal of Global Information Management (JGIM) 27.1, 2019, 111-127.

[5] Kaggle. Credit Card Customers. 2021. https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

[6] Biau, Gérard, and Erwan Scornet. A random forest guided tour. Test 25.2, 2016, 197-227.

[7] Myles, Anthony J., et al. An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society 18.6, 2004, 275-285.

[8] AL-Najjar, Dana, Nadia Al-Rousan, and Hazem AL-Najjar. Machine Learning to Develop Credit Card Customer Churn Prediction. Journal of Theoretical and Applied Electronic Commerce Research 17.4, 2022, 1529-1542.

[9] Yu, Q, et al. Clustering Analysis for Silent Telecom Customers Based on K-means++. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 1. IEEE, 2020.

[10] Bilal Zorić, A. Predicting customer churn in banking industry using neural networks. Interdisciplinary Description of Complex Systems: INDECS 14.2, 2016, 116-124.