Ensembled Learning Based Model for Bank Churn Prediction

Jieyuan Deng[1][a]*, Junda Huang[2][b], Dongliang Liu[3][c] and Shaohan Yang[4][d]
[1]*Nanjing University of Posts and Telecommunications, 9 Wenyuan Road, Qixia District, Nanjing, Jiangsu, China*
[2]*South China University of Technology, 382 Waihuan Dong Road, Panyu District, Guangzhou, Guangdong, China*
[3] *HDU-ITMO Joint Institute, Hangzhou Dianzi University, 1158 2nd Street, Qiantang District, Hangzhou, Zhejiang, China*
[4]*Rosedale Global High School, 7030 Woodbine Ave #800, Markham, Canada*

Keywords: Machine Learning, Ensemble Learning, Neural Network, Bank Customer Churn Rate.

Abstract: Predicting customer churn rate helps banks retain customers, stabilize their market position, and improve services, providing a better customer experience for both parties, especially for developed countries. Although various scholars have conducted different studies in various locations, there has been no rigorous research on predicting bank customer churn. The purpose of this study is to develop a satisfactory predictive model to forecast the probability of customer churn for banks, providing reliable references for numerous banks. This paper implements various machine learning methods and deep learning models, including Logistic Regression, Random Forest, Neural Network, XGBoost, Decision Tree, Gradient Boosting, and Ada Boost. Among all models, the combination of Random Forest and Neural Network achieved the best results, with an adjusted recall1 of 0.6 and precision0 of 0.9. In addition, we used insights obtained from these powerful ensemble learning models to analyze factors leading to bank customer churn.

1 INTRODUCTION

In recent years, neural network models for customer churn prediction have garnered attention due to their superior nonlinear modeling capabilities. Neural network models simulate the workings of the human brain's neural system, possessing characteristics such as adaptability, nonlinearity, and parallel processing, which make them well-suited to handling the complexities of customer churn prediction. By training neural networks on bank customer data, we can more accurately predict customer churn and devise corresponding retention strategies (Tang, 2021).

This paper takes Germany, France, and Spain as examples, three regions belonging to developed countries with thriving economies and intensely competitive financial service markets. These regions offer a good case study for this research, given the diverse and mature range of services provided by banks to customers. The aim of this study is to leverage machine learning technologies to construct precise customer churn prediction models for banks.

The structure of this paper is as follows: Section 2 introduces peer-reviewed works related to customer churn prediction by showcasing various categories. Section 3 provides detailed insights into the methods chosen, reasons for their selection, and theoretical introductions to these methods. Subsequently, in Section 4, experimental results are examined and analyzed. Last but not least, the conclusions of this study are outlined in Section 5, followed by references.

2 RELATED WORK

The customer churn rate is determined by many factors such as credit score, tenure, balance and estimated salary. With the continuous advancement of new technologies, more and more machine learning technology has been applied to the forecast

[a] https://orcid.org/0009-0004-1145-4291
[b] https://orcid.org/0009-0008-6833-795X
[c] https://orcid.org/0009-0002-1145-3470
[d] https://orcid.org/0009-0004-7455-7909

Deng, J., Huang, J., Liu, D. and Yang, S.
Ensembled Learning Based Model for Bank Churn Prediction.
DOI: 10.5220/0012887100004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 10-15
ISBN: 978-989-758-713-9

model of customer churn (Li, 2019). Scholars around the world have conducted a lot of inquiry in this regard so that the models can predict more accurately (Wang 2022, Chandar 2006).

Hu proposed that the problem of customer churn of retail banks can be used for research and solution with data mining technology. Chandar et al. used three algorithms (CART, C5.0 and TreeNet) to predict the churn of bank customers. It was found ultimately that the classification prediction of the CART algorithms was the best. To deal with imbalanced data, Wangyu Liao (2012) used the improved Boosting method, proving that the improved Boosting method has enhanced the ability of model processing imbalanced data, and reduced the predictive deviation caused by the imbalance of the data set (Liao, 2012). Huang et al. proposed understandable support vector machine, and at the same time, he used simple Bayez tree to build a customer loss model, which has high accuracy in prediction (Huang, 2014). On the basis of the use of support vector machine algorithms, He et al. also explored the prediction of commercial bank customer churn. Focusing on data imbalances, the model is further improved by random samples method, and the results show that the method can significantly improve the accuracy of the model forecast (He, 2017). Huang et al. proposed an algorithms that combines Particle Swarm Optimization and Back Propagation to establish a warning model of corporate customer churn (Huang, 2018). However, the Back Propagation has a lot of disadvantages, such as unable to converge quickly, high possibility of caught in local minimum. Swetha P and Dayananda B proposed the Improvized-XGBOOST model with feature functions for the prediction of customer churn. The result illustrates that the model is more efficient and it can be suitable for complex data sets (Swetha, 2020).

It can be seen from the above that many scholars have conducted a series of related studies on customer churn, while most of the studies used a variety of single models to build the customer churn predictive model, having achieved some results.

3 METHODOLOGIES

3.1 Data Preprocessing

The complexity of data types and their internal correlations, and the disunity of data quality will have a negative impact on data interpretation and analysis. Data preprocessing is a crucial step in the process of machine learning. The quality of the data greatly affects the outcome of the machine learning model. It includes data cleaning, data integration, data conversion, data reduction and other steps. Through these steps, the accuracy, interpretability and robustness of the model can be improved. In this research, the process of data preprocessing includes five parts: deleting redundant features, performing one-hot encoding of text information, processing missing values, scaling features, and using SMOTE method to balance the data set. In addition, the original dataset is split into a training set (80%) and a test set (20%).

- **Deleting Redundant Features**

Redundant features will increase the computational complexity of model training. Deleting redundant features can reduce the computational cost and improve the efficiency of the model. In this research, Surname and CustomerID were removed from features, because they duplicate RowNumber.

- **Performing One-Hot Encoding of Text Information**

The text information in the feature quantity is one-hot encoded, aiming to convert the text information into a numerical type. In addition, one-hot encoding can prevent the model from having a preference for values of different sizes, thus affecting the accuracy of the model. In this research, one-hot encoding is applied to the geographical location and gender in the feature quantities to help the model better understand and utilize classification information.

- **Processing Missing Values**

Missing values will cause the model to lack effective information during training and prediction, thereby reducing model performance. Therefore, filling in missing values in the data set can improve the stability and interpretability of the model. In this research, missing values only appear in HasCreditCard and Is ActiveMember. Considering its actual meaning, that is, whether the users have credit cards and whether they are active users, 0 is used to fill the missing values.

- **Scaling Features**

Feature scaling can prevent some feature values from being too large, causing these features to overly affect the prediction results. In this research, StandardScaler is used to normalize numerical features to ensure they have the same scale.

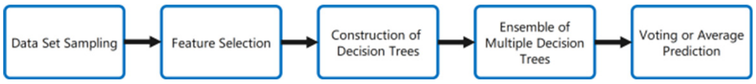
Data Set Sampling
Feature Selection
Construction of Decision Trees
Ensemble of Multiple Decision Trees
Voting or Average Prediction

Figure 1: Random forest model training process (Photo/Picture credit : Original).

- **SMOTE Method Balance the Dataset**

In unbalanced data sets, SMOTE method can be used to solve the overfitting problem of random oversampling method. SMOTE mainly balances the sample distribution of different categories by synthesizing some new minority class samples. In this research, through data analyzing, it is evident that the number of people who have not exited is much greater than the number of people who have quit. Therefore, SMOTE method is used to balance the dataset.

3.2 Model Selection and Construction

3.2.1 Logistic Regression

Logistic regression is a generalized linear regression analysis model. In practical applications, it mainly solves binary classification problems. The logistic regression model first performs a weighted summation of the input features and adds an offset term (intercept term) to obtain a linear combination.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

where z is the linear output, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the feature weights, and $x_1, x_2, \ldots, x_n$ are the input features.

This linear combination is used to represent the log odds of the dependent variable. It then transforms the results obtained from the linear regression model through a non-linear sigmoid function,

$$p = \frac{1}{1+e^{-z}} \quad (2)$$

where p is the probability of the event occurring, and e is the base of the natural logarithm. Sigmoid function produces values within the range [0, 1]. Setting the threshold to 0.5, achieving binary classification by comparing the results with the threshold.

3.2.2 Random Forest

Random forest is an algorithm that integrates multiple decision trees using the idea of ensemble learning. Relevant studies show that the random forest model has a good performance in processing large-scale data, and can handle both discrete data and continuous data (Sandeepkumar, 2020). Due to attribute perturbation, the initial performance of random forest is inferior to that of decision tree. However, when the number of base learners is large, the random forest will converge to a lower generalization error. In addition, random forest has stable performance and good noise resistance, because each tree randomly selects samples and its features. As for the dataset in this research, which contains about 10,000 pieces of data, using random forest model is an effective method. The basic flow of random forest model is shown in figure 1.

3.2.3 AdaBoost

AdaBoost is an ensemble learning algorithm used for binary classification problems. It constructs a strong classifier by combining multiple weak classifiers, each of which is relatively simple and performs slightly better than random guessing.

The basic principle of AdaBoost is to iteratively learn a series of weak classifiers and combine them based on their performance to obtain a more accurate classifier. In each iteration, AdaBoost adjusts the weights of the samples based on the classification results of the previous models, focusing more on the samples that were previously misclassified. Through multiple iterations, AdaBoost gradually learns a strong classifier with high classification accuracy.

3.2.4 XGBoost

XGBoost (eXtreme Gradient Boosting) is a powerful machine learning algorithm widely used in regression and classification problems. It is an ensemble learning model based on decision trees that improves the predictive capacity of the model through boosting.

Firstly, Since XGBoost employs parallel computing techniques and optimized algorithms, making it efficient in handling massive data and large-scale features, it has high performance in both training speed and model prediction speed. In addition, XGBoost can handle various types of feature variables, including discrete and continuous features. It also supports custom loss functions, providing flexibility to adapt to different problems and tasks. Moreover, XGBoost introduces regularization terms to control the model complexity and prevent overfitting. The regularization terms include L1 regularization and L2 regularization, allowing the model to generalize better to new data. XGBoost can also calculate feature importance or weights, helping users understand the prediction process of the model. By ranking feature importance, it can identify the most influential features in predicting the outcome.

4 EXPERIMENT

4.1 Dataset Overview

This paper utilizes the Bank Customer Churn Prediction from Kaggle, which contains information on bank customers who either left the bank or continue to be a customer. Each entry in the dataset consists of 13 attributes (shown in Table 1).

Table 1: Description of Attributes in the Dataset.

Attribute	Description
CustomerId	A unique identifier for each customer
Surname	The customer's surname or last name
CreditScore	A numerical value representing the customer's credit score
Geography	The country where the customer resides (France, Spain or Germany)
Gender	The customer's gender (Male or Female)
Age	The customer's age.
Tenure	The number of years the customer has been with the bank.
Balance	The customer's account balance.
NumOfProducts	The number of bank products the customer uses (e.g., savings account, credit card).
HasCrCard	Whether the customer has a credit card (1 = yes, 0 = no).
IsActiveMember	Whether the customer is an active member (1 = yes, 0 = no).
EstimatedSalary	The estimated salary of the customer.
Exited	Whether the customer has churned (1 = yes, 0 = no).

In the hope of selecting features that are conducive to classification, correlation between Exited and other attributes is explored.

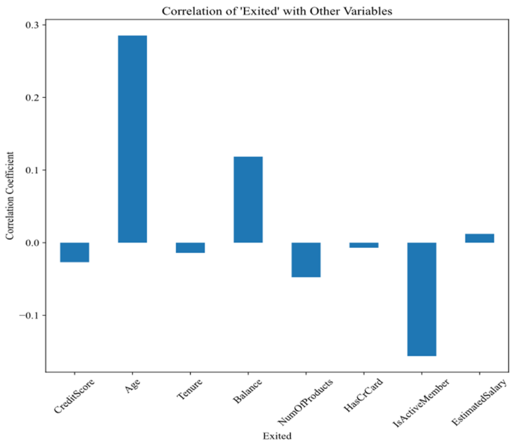

Figure 2: Correlation Between ‘Exited’ and Other Attributes (Photo/Picture credit: Original.

As is shown in figure 2, the four attributes—Age, IsActiveMember, Balance, and NumOfProducts—have a relatively high correlation with Exited.

4.2 Data Exploration

We conducted analysis of the dataset according to the correlation (Figure 3 and figure 4).

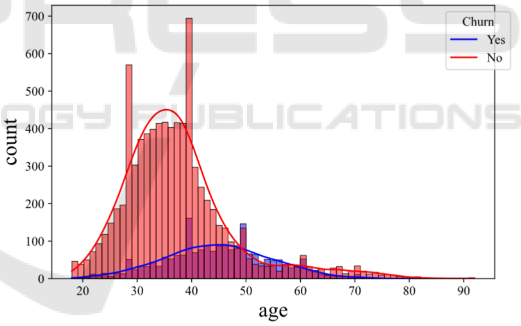

Figure 3: The correlation between age and credit (Photo/Picture credit : Original).

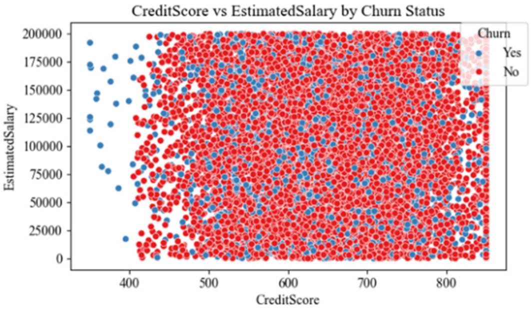

Figure 4: Credit Score vs Estimated Salary (Photo/Picture credit: Original).

According to the figure 3 and figure 4, the age of churned customers is concentrated between 40 and 50 years old. Customers with a credit score below 400 are all churned customers.

4.3 Experimental Settings

In this research, all models were implemented in Python 3.9.13 environment, with Pandas, Scikit-Learn, Tensorflow and XGBoost packages.

The parameter settings for each model we used are as follows:

- Logistic Regression

The LogisticRegression class implemented in Python's scikit-learn library is utilized, with the default parameters.

- Random Forest

The Random Forest model took advantage of a total of 100 trees, and the split criterion is gini.

$$\mathrm{Gini(D)} = 1 - \sum_{i=1}^{m} p_i^2 \qquad (3)$$

- AdaBoost

Stagewise Additive Modeling using a Multiclass Exponential loss function, Real version (SAMME.R) is utilized in the model, with 50 estimators.

- XGBoost

The booster we choose in the XGBoost model is gbtree, and a total of 100 gradient boost trees are utilized. The objective function is logistic.

- Neural Network

In our research, a 3-layer deep neural network is implemented.

4.4 Modal Evaluation

The primary objective is to improve Recall 1 since the goal is to predict customer churn.

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \qquad (4)$$

Among all the models, as expected, Logistic Regression model performed the worst because of the characteristic of the dataset (Figure 5). Even though other three models (RandomForest, AdaBoost, XGBoost) gained relative fair score, NN+RF yield satisfying result, and outperform all other models with respect to Recall 1. Deep-learning methods are representation-learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level (LeCun, 2015). Given that neural networks consist of multiple neurons and hidden layers, providing high flexibility and non-linear expressiveness. They can learn complex patterns and relationships, enabling better adaptation to training data. In contrast, random forests, AdaBoost, and XGBoost are decision tree-based methods with relatively weaker performance.

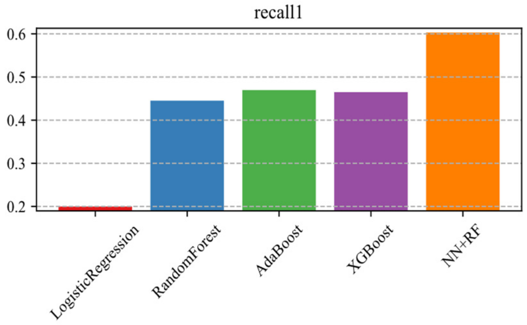

Figure 5: The result of Recall1 (Photo/Picture credit : Original).

5 CONCLUSION

In conclusion, this paper combines machine learning and deep learning to seek an appropriate method for predicting the probability of bank customer churn. The models employed include Logistic Regression, Random Forest, Neural Network, XGBoost, Decision Tree, Gradient Boosting, and Ada Boost. Among all these models, the ensemble learning approach combining Random Forest and Neural Network yielded the best results in predicting bank customer churn rate, achieving satisfactory outcomes with a recall1 value of 0.6 and a precision0 value of 0.9. Another advantage of ensemble learning methods is their ability to assess the relative importance of each attribute during training. Experimental results indicate that the customer's age, geographic location, account balance, active membership status, and the number of bank products used are the top five significant features influencing bank customer churn. The paper also explains how these factors can be utilized to devise different strategies in real-world scenarios to address bank customer churn.However, the dataset only includes three regions and may not be fully applicable to every country. Additionally, obtaining relevant bank-related data can be quite challenging, as it is only accessible through public platforms. Therefore, predicting bank customer behavior based on this dataset may not be comprehensive enough. It is hoped that future research will optimize existing problems and algorithms to reduce sensitivity to class imbalance data and improve churn prediction accuracy. churn.

The paper also explains how these factors can be utilized to devise different strategies in real-world scenarios to address bank customer churn.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

Chandar M., Laha A, Krishna P. 2006. *Modeling Churn Behavior of Bank Customers Using Predictive Data Mining Tools*, Business Intelligence Journal,5(1):96-101.

He, B, L., Y. Shi, Q. Wan, et al. 2014. *Prediction of Customer Attrition of Commercial Banks Based on SVM Model.* Procedia Computer Science. 31(3), 423~430.

Huang, K.Z., D.Zheng, J.Sun, et al. 2014. *Sparse Learning for Support Vector Classification.* Pattern Recognition Letters,31(13),1944~1951.

Huang, J.F.and L.L.H, 2018. *Application of Improved PSO-BP Neural Network in Customer Churn Warning.* Procedia Computer Science,131.1238~1246.

LeCun, Y., Bengio, Y. & Hinton, G. 2015. *Deep learning.* Nature 521, 436–444.

Li Y. X., Chai Y, Hu Y Q, et al. 2019, *A review of classification methods for unbalanced data(in Chinese)*. Control and decision. 34(04). 673-688.

Liao W Y. 2012. *The Credit Customers Churn Analysis Based on Improved Boosting Decision Tree.* Computer Knowledge and Technology.8(18).4306-4307+4319.

Sandeepkumar,H. and M.R.Mundada. 2020. *Optimized Deep Neural Network Based Predictive Model for Customer Attrition Analysis in the Banking Sector.* Recent Patents on Engineering,14(3), pp.412~421.

Tang C L, Chen H, Tang H C, et al. 2021, *Research and application of data preprocessing methods under the background of big data(in Chinese).* Information recording material. 22(09). 199-200.

Wang L C, Liu S S. 2022. *Research on improved random forest algorithm based on mixed sampling and feature selection(in Chinese).* Journal of Nanjing University of Posts and Telecommunications(Natural Science Edition). 42(01), 81-89.