Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/69/20230773

Bank Customer Churn Prediction with Machine Learning Methods

He Zhu[1,a,*]

[1]*Institute of Education, University College London, Gower St, London WC1E 6AE*
a. stnvh73@ucl.ac.uk
**corresponding author*

Abstract: This paper examines and analyses customer churn prediction in the banking sector using the data from ABC Bank. The analysis conducted will document the determinants of bank customer churn and provide insights to the most important factors which influence the customers decision to quit utilizing the services of a bank. The investigation is based on the results of two machine learning algorithms with k-fold-cross-validation and same boosting methods. The result of the analysis reveals that out of logistic regression and random forests algorithms, the random forest methods show a higher accuracy score which corresponds with the literature review studied. Furthermore, the statistic of the research indicates that customer's age has the highest association with the likelihood of customer churning, while whether the customer has a credit card at the bank has the lowest interconnection. The results of this research may provide valid explanations to customer churn in the banking sector and bring further intuitions of the advantages which machine learning methods may provide to future financial analysis.

Keywords: Bank Customer, Prediction, Machine Learning

1. Introduction

It is recently predicted that banks are at the turning point; with the numbers rising in neo-industries, banks need to restructure the services based on a future-oriented strategy [4]. It' estimated by the World Retail Banking Report in 2019 that 66.8 percent of current banking customers is expecting to switch to new forms of financial institutes for similar services [5] this phenomenon could undoubtedly lead to a pernicious effect on the banking sector. It is perceivable that such phenomenon is the outcome of both, potentially, banking service failures, and a change of tastes of the customers [6]. Accordingly, to minimize the loss due to customer discarding from the bank, it's necessary to analyze the factors which might influence the basis of churning. However, the internal flaws could be detected, it may be laborious to directly determine the pattern of customer preferences and make tailored solutions against it. and predict of such trend is crucial with the basis of customer taste since it's closely related to the success of banks [7].

Customer churn, or the loss of valuable customers can significantly impact a bank's revenue and growth, hence, it's crucial to make customer churn prediction as it can forestall potential churning while also improve bank services, along with the gains of the institute [1]. As mentioned above, customer churn, besides from being the result to enhance bank facilities, could also be the

DOI: 10.54254/2754-1169/69/20230773

consequences of customer preferences; as customers is a crucial dependency of banks, the risk of them leaving is important for bank development [6].

Consequently, the focus of this research lies in the development and comparative analysis of several machine learning algorithms to predict customer attrition in the banking sector. Customer churn, defined as the proportion of contractual customers or subscribers who leave a supplier during a given time period, represents a substantial revenue loss for banks [5]. These organizations must bear the cost of acquiring new customers to replace the ones who churn, making churn prediction a key strategic component for customer relationship management and business sustainability.

To conduct the research a dataset from Kaggle is extracted, which comprises of demographic and transactional data of bank customers, to develop and evaluate our prediction models. Our primary objective is to provide a comprehensive perspective on the application of various machine learning techniques, including supervised and unsupervised learning methods, for predicting customer churn.

The paper will illuminate the challenges and opportunities inherent in predicting bank customer churn, with a focus on the potential of machine learning to enhance accuracy, efficiency, and customer satisfaction in the banking industry. It further emphasizes the importance of customer retention strategies in ensuring long-term success in a dynamic and evolving financial marketplace. We believe that our research can serve as a foundation for future studies in this area and provide a road map for banking institutions seeking to mitigate customer churn using AI-based solutions.

2. Literature Review

In recent years, predicting customer churn in the banking sector has attracted considerable attention from researchers and practitioners alike. This area has been investigated from various perspectives, involving diverse methodologies ranging from traditional statistical models to advanced machine learning techniques, since research shows that the methods used to conduct such studies could significantly impact the results of the studies [10].

Seminal studies in the domain of churn prediction were initially grounded in the use of traditional statistical methods. For example, Buckinx and Van den Poel employed logistic regression to predict customer churn in the financial services sector, demonstrating its efficacy as a baseline model [2]. Examples from varying industries also shows that logistic functions have been effective in the prediction of churns, for example the study done in the telecommunication section [8]. Other conventional methods like decision trees and support vector machines have also been explored by Vafeiadis et al. [14]. In the study of de Lema Lemos et al. [5], it's shown that after running multiple regression models for the study of customer churn in a financial institution, there is a significant benefit with the usage of random forest technique, hence, this investigation will also attempt to verify that result. From the same study, it's shown that there is a strong relationship between customers and the institution, which proves that the topic of research is valid. Studies have begun to investigate more add more metrics to the prediction to study the link between customer churn and their individual experiences. For instance, Jamal and Bucklin [9] added time-varying covariates to the model and shows that the results of the model is significantly better when heterogeneity is added. Concerning methodology, K-fold cross-validation has been widely adopted as a standard method for model evaluation in churn prediction in other fields, such as Suh [14], It is widely accepted for its ability to provide unbiased performance estimates and prevent model overfitting.

Finally, research on feature importance in the context of customer churn has indicated that factors such as age, balance, credit score, and the number of products used often play significant roles [5]. However, the relative importance of these factors tends to vary across studies and datasets, suggesting the need for further investigation. In summary, while a considerable amount of research has been done in the area of customer churn prediction, there is still substantial scope for comparative studies, methodological refinement, and the investigation of feature importance. This paper contributes to the

DOI: 10.54254/2754-1169/69/20230773

existing literature by comparing logistic regression and random forest models, applying k-fold cross-validation, and examining feature importance in the context of bank customer churn prediction.

3. Data and Methodology

The section contains the information on the dataset, data preparation and machine learning methods including model building and outcome analysis. The data utilized for this study was sourced from a synthesized Kaggle dataset on bank customer churn. The dataset uses a wide range of criterions, including detailed information on customer demographics, member status, salaries, product usage, and customer service interactions. It also contains the churn status of each customer, providing a robust ground truth for the development and validation of our models.

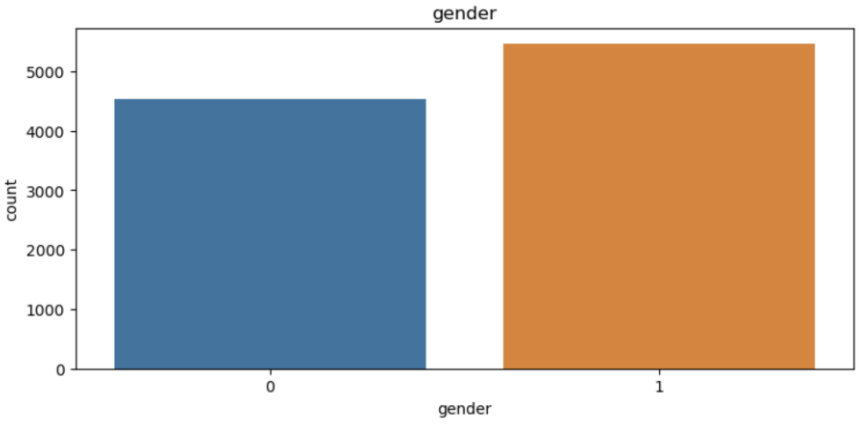
The dataset contains 120,000 entries, offering an ample sample size for training and testing our prediction models. After examining the data status of missing data, the extracted data is found to be well-documented and already cleaned, hence, after removing the customer ID column during the data pre-processing phase, approximately, 100,000 entries are utilized in the analysis.

3.1. Study Sample

Table 1: The variables and attribution used in the bank customer churn prediction analysis

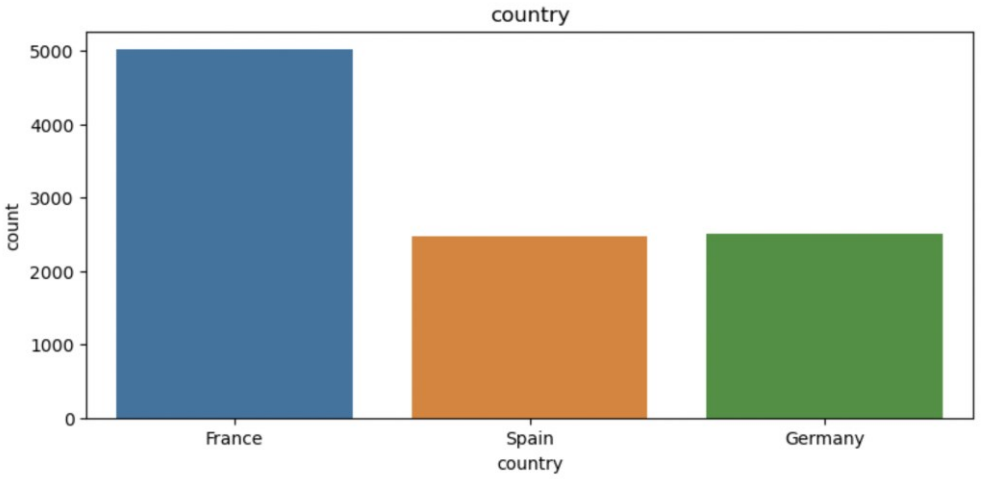
Variable	Data Type	Description
credit_score	Numerical	The credit score of the customer
country	Categorical	Country of residence
gender	Binary (Male or Female)	Sex of the customer
age	Numerical	The age of the customer
tenure	Numerical	For how many years he/she is having bank account in ABC bank
balance	Numerical	The account balance of the customer
products_number	Numerical	Number of products from bank
credit_score	Binary (Yes or No)	Whether if the customer own a credit card at the bank
active_member	Binary (Yes or No)	Is the customer an active member of the bank
estimated_salary	Numerical	Estimated salary of the customer
churn	Binary (Yes or No)	Whether if the customer's account has been closed

3.2. Understanding the Data



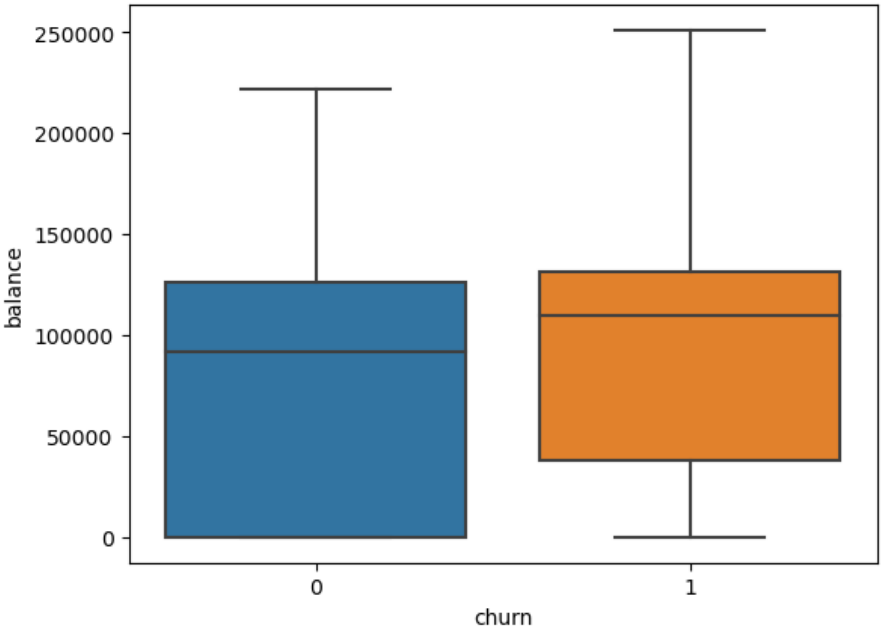
Graph 1: Distribution of customer gender

DOI: 10.54254/2754-1169/69/20230773



Graph 2: Distribution of customer demographics

The preceding graphs present the distribution of the customers' gender and residence, showing that there are relatively even distributions of each attributes of the variable. For the detail of the distribution, it's visible that there are a larger proportion of customers from France and a slightly higher numbers of male customers in the dataset.



Graph 3: The relationship between customer churn and the balance of the customer in the bank

The boxplot above shows that customers who have a higher balance in the bank is more likely to churn in comparison to those with a lower balance at the bank.

Table 2: The numbers of customers that have churned

Churned (Yes/No)	Yes	No
Counts	2,037	7,963

DOI: 10.54254/2754-1169/69/20230773

Table 2 shows that there is about 20% of the customers that have churned, hence, there is an imbalanced classification, which feature engineering needs to be performed before building the predictive model.

3.3. Model building

In the research, two widely applied machine learning methods are used: Logistic Regression and Random Forest. Logistic Regression is a statistical model used for binary classification problems. It estimates the probability of a binary outcome based on one or more predictor (or independent) variables. It provides a robust baseline model for this type of prediction task [3].

$$p(X) = \frac{1}{1 + e^{-(a+bX)}}$$

The fundamental equation for logistic regression relates to the log-odds of the probability of an event to a linear combination of independent variables [11]. The preceding equation of the logit regression transforms a linear combination of binary variables, where $p(X)$ denotes the probability of an event occurring given the independent variable X. While e stands for the base of natural logarithm, a and b each acts the intercept and the coefficient of the independent variable X.

On the other hand, the Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training time and outputting the majority vote of individual trees for classification problems [12].

$$Y(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

The Random Forest, being an ensemble of decision trees, is a collection of equation from individual trees, in the equation above, $Y(x)$ signifies the prediction of the Random Forest model for a given input x, and B suggests the quantity of individual decision trees comprised in the forest. Finally, $T_b(x)$ is the representation of the prediction of the bth tree of the same input x. The model, hence, calculates the average of the prediction from all the individual trees in the forest. Given its ability to handle high dimensional spaces and multicollinearity between variables effectively, Random Forest is highly suited to our dataset.

Data pre-processing involved handling missing values, scaling numerical attributes, and encoding categorical variables, as mentioned prior, there is a imbalanced classification in the dataset, hence, a feature engineering needs to be performed. The data set is divided into a training set (80% of the data) and a test set (20% of the data) for model training and evaluation.

To prevent overfitting and ensure the generalizability of our models, we used k-fold cross-validation technique during the model training process. In k-fold cross-validation, the training dataset is randomly partitioned into 'k' equal-sized subsamples. Of the 'k' subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 'k-1' subsamples are used as training data. The cross-validation process is then repeated 'k' times, with each of the 'k' subsamples used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation, providing a more comprehensive assessment of the model performance.

The performance of the Logistic Regression and Random Forest models was evaluated based on several metrics, including accuracy, precision and ROCAUC scores. Furthermore, the feature importance is examined, provided by the Random Forest model to identify the most influential factors driving customer churn.

DOI: 10.54254/2754-1169/69/20230773

By following this rigorous methodology, the study aims to provide reliable, robust, and informative insights on the application of machine learning in predicting customer churn in the banking sector.

4. Results

The findings of the analysis, based on the analysis of the bank customer dataset using logistic regression and random forest models, offer insightful revelations into the field of bank customer churn prediction.

Table 3: The Cross-Validated Results of the Random Forest and Logistic Regression Models

	Random Forest	Logistic Regression
Cross-validation Scores	0.863 0.869 0.86 0.864 0.853	0.795 0.788 0.793 0.792 0.796
Mean Score	0.862	0.793
Standard Deviation	0.00522	0.00266

After building the random forest and logit model and running a k-fold-cross-validation score with 5 folds on the model, it's visible that the random forest model has a relatively higher mean score of 0.862, while the logit model obtains a lower score of 0.793. Both model acquired low standard deviations, with 0.00522 for the random forest model, and 0.00266 for the logistic model, suggesting that both models are fairly accurate and there are few variations in the model.

Consequently, the random forest model is selected as the better model, hence, an accuracy test is performed on this selected model. In terms of accuracy, the Random Forest model achieved an accuracy of 85.9%, and a precision score of 76.9%. The AUROC, a comprehensive measure of model performance, was found to be 0.852 for the Random Forest model, showing that the model has a predictive accuracy of 85.2%, which is a relatively high score for a model. This aligns with findings from previous literature which suggests that the random forest model outperforms the other algorithms.

The superiority of the random forest model can be attributed to its ability to handle high-dimensional data and capture complex, non-linear relationships between variables. This robust performance underscores the potential of ensemble learning methods in customer churn prediction. Further, the k-fold cross-validation confirmed the robustness of our models by providing consistent performance metrics across different folds. This process effectively mitigated the risk of overfitting, ensuring the reliability and generalizability of our results.

When analyzing feature importance, the Random Forest model indicated that customer age was the most significant factor influencing customer churn. Other significant variables included account balance, credit score, and the number of products used, underlining the multi-dimensional nature of churn prediction.

These results not only shed light on the prediction of customer churn in the banking sector but also provide valuable insights for both researchers and practitioners. The superior performance of the Random Forest model highlights its potential for real-world applications in customer relationship management and retention strategies. Furthermore, the importance of customer age, among other factors, can inform the design of targeted intervention strategies to mitigate churn.

5. Conclusion

The study, aimed at predicting customer churn in the banking sector using machine learning techniques, provides several important findings and implications for both academic research and industry practices. Firstly, the results show that machine learning methods, specifically the Random

Forest model, can effectively predict customer churn. Compared to the Logistic Regression model, the Random Forest model delivered superior performance across all metrics. This underlines the significant potential of advanced machine learning techniques in tackling customer retention challenges in the banking sector. Secondly, our research offers insights into the factors influencing customer churn. Specifically, customer age was found to be the most significant variable impacting churn. This indicates that age-specific interventions could be an effective part of churn prevention strategies in banking institutions. In conclusion, this study represents a step forward in understanding customer churn in the banking sector through machine learning. The knowledge generated from this research can contribute to enhancing customer relationship management strategies, improving customer retention, and ultimately driving sustained growth in the banking industry.

Despite these contributions, our study is not without limitations. The findings are based on a single dataset, which may not fully capture the diversity and complexity of the global banking customer base. Furthermore, while the study explored two machine learning techniques, other methods such as gradient boosting or neural networks may offer additional insights into churn prediction. Future research could extend this work by exploring other machine learning techniques, incorporating a wider range of customer variables, and using more diverse datasets. Also, the integration of unsupervised learning methods to identify hidden customer segments at high risk of churn may offer additional insights.

References

[1] *Briker, V., Farrow, R., Trevino, W., & Allen, B. (2019). SMU Data Science Review, 2(3).*

[2] *Buckinx, W., & Van den Poel, D. (2005). c. European Journal of Operational Research, 164(1), 252–268. doi:10.1016/j.ejor.2003.12.010*

[3] *Cole, A. (2020). Retrieved from https://towardsdatascience.com/predicting-customer-churn-using-logistic-regression-c6076f37eaca*

[4] *Czímer, B., Dietz, M., László, V., & Sengupta, J. (2022). Retrieved from https://www.mckinsey.com/industries/financial-services/our-insights/the-future-of-banks-a-20-trillion-dollar-breakup-opportunity*

[5] *de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. Neural Computing and Applications, 34(14), 11751–11768. doi:10.1007/s00521-022-07067-x*

[6] *Guliyev, H., & Yerdelen Tatoğlu, F. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. Journal of Applied Microeconometrics, 1(2), 85–99. doi:10.53753/jame.1.2.03*

[7] *J, S., Gangadhar, Ch., Arora, R. K., Renjith, P. N., Bamini, J., & Chincholkar, Y. devidas. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. Measurement: Sensors, 27, 100728. doi:10.1016/j.measen.2023.100728*

[8] *Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167, 101–112. doi:10.1016/j.procs.2020.03.187*

[9] *Jamal, Z., & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. Journal of Interactive Marketing, 20(3–4), 16–29. doi:10.1002/dir.20064*

[10] *Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204–211. doi:10.1509/jmkr.43.2.204*

[11] *Stoltzfus, J. C. (2011). Logistic regression: A brief primer. Academic Emergency Medicine, 18(10), 1099–1104. doi:10.1111/j.1553-2712.2011.01185.x*

[12] *Suh, Y. (2023). Machine learning based customer churn prediction in home appliance rental business. Journal of Big Data, 10(1). doi:10.1186/s40537-023-00721-8*

[13] *Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using Random Forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access, 7, 60134–60149. doi:10.1109/access.2019.2914999*

[14] *Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. Ch. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1–9. doi:10.1016/j.simpat.2015.03.003*