JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol.11 No.3 S1, pISSN:2303-0577 eISSN:2830-7062 http://dx.doi.org/10.23960/jitet.v11i3%20s1.3394

UNVEILING CHURN PREDICTION AT BANK IVORY

Marsella Patricia[1], Raymond Sunardi Oetama[2*], Iwan Prasetiawan[3]

[1-3]Universitas Multimedia Nusantara; Scientia Garden Boulevard Raya Gading Serpong, Tangerang, Indonesia.

Riwayat artikel:
Received: 21 Juli 2023
Accepted: 25 Agustus 2023
Published: 11 September 2023

Keywords:
customer churn;
CRISP-DM;
gradient boosting;
Random Forest

Corespondent Email:
raymond@umn.ac.id

Abstrak. Industri perbankan menghadapi tantangan signifikan dalam menangani pergantian nasabah dalam layanan kartu kredit. Pergantian nasabah terjadi ketika nasabah berhenti menggunakan layanan dari bank tertentu dan beralih ke bank lain. Untuk mengatasi masalah kritis ini, penelitian saat ini bertujuan untuk memprediksi pergantian nasabah dalam layanan kartu kredit. Dalam mencapai tujuan ini, penelitian ini menggunakan kerangka kerja CRISP-DM secara luas dan membandingkan kinerja dua model prediktif, yaitu Gradient Boosting dan Random Forest. Penelitian ini berusaha untuk mengidentifikasi nasabah yang berpotensi keluar dengan menganalisis variabel-variabel penting, seperti usia nasabah, status pernikahan, jenis kelamin, kategori pendapatan, batas kredit, dan total transaksi. Pendekatan pemodelan yang dipilih ditentukan berdasarkan tingkat kesalahan klasifikasi terendah, menjadi komponen penting dalam proses analisis penelitian. Hasil dari model Gradient Boosting menunjukkan tingkat kesalahan klasifikasi sebesar 0.1118 dalam memprediksi pergantian nasabah ini.

Abstract. *The banking industry faces significant challenges in tackling customer churn within its credit card services. Customer churn refers to the situation where customers discontinue using a bank's services and migrate to another financial institution. To proactively address this critical issue, the present research endeavors to predict customer attrition in credit card services. To achieve this goal, the study extensively employs the CRISP-DM framework and diligently compares the performance of two predictive models, namely Gradient Boosting and Random Forest. The research endeavors to identify potential churn customers by analyzing crucial variables, including customer age, marital status, gender, income category, credit limit, and total transactions. The preferred modeling approach, determined based on the lowest misclassification rate, serves as a vital component of the research's analytical process. Remarkably, the research findings unequivocally demonstrate the superior performance of the Gradient Boosting model, which attains a misclassification rate of 0.1118 in predicting customer attrition.*

1. INTRODUCTION

At a financial institution, a business manager is facing challenges with an increasing number of customers leaving their credit card services [1]. As a result, they sought assistance from a data scientist to predict potential customers who might churn from their credit card services [2]. The business manager also requires technology to help process and analyze the data effectively. This research aims to analyze the customers using credit card services and predict whether they are likely to churn or not. The outcomes of this study will provide the business manager with insights into which customers are prone to churning. Armed with this knowledge, they can approach these customers with better services or offers, thus changing their decision to leave

and encouraging them to remain loyal to the credit card services. To facilitate the research process, the data science technique of the CRISP-DM framework will be utilized, along with Machine Learning algorithms, namely Gradient Boosting and Random Forest. The required tools for the research include SAS Studio and SAS Visual Analytics. By comparing the misclassification rates of the models generated using Gradient Boosting and Random Forest algorithms, the research aims to select the model with the lowest misclassification rate. A lower error rate or misclassification rate indicates better model performance. The goal of this research is to successfully predict customers likely to churn from credit card services, leading to a reduction in customer churners at the financial institution when the predictions are implemented.

2. METHODS

Customer churn, also known as customer attrition, occurs when customers stop using a bank's services [3]. Several reasons contribute to churn, such as no longer needing the services, finding more appealing options from other banks, or experiencing dissatisfaction with the bank's offerings, leading customers to discontinue. Data mining projects involve interlinked stages, converting the process into an iterative and collaborative form [4]. As can be seen in Figure 1, CRISP-DM (Cross-Industry Standard Process for Data Mining) is a well-structured methodology used for data modeling and analysis [5]. Business Understanding involves comprehending the business objectives and identifying the specific problem to be addressed. Data Understanding entails gaining a comprehensive understanding of the available data, identifying pertinent variables for the research, and preparing for the subsequent stage, which is Data Preparation. During Data Preparation, the data is readied for analysis, which may include processes like data cleaning and transformation to facilitate further analysis. Modeling encompasses the application of diverse data modeling techniques to create an appropriate analysis model. This involves model selection, testing various models, and adjusting model parameters. Evaluation involves assessing the model's performance and checking if it fulfils the initial objectives. Finally, Deployment entails implementing the model into a production environment and integrating it with existing systems.

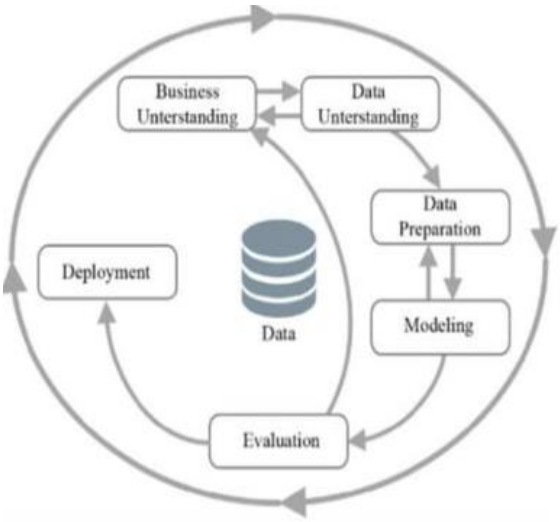

Figure 1. CRISP-DM [6]

Gradient Boosting and Random Forest algorithms are commonly applied in predicting customer churn in various industries, including banking and finance [7], [8], [9]. Gradient Boosting is a form of supervised learning that combines base learners in an additional ensemble [10]. The model is learned in a sequential manner, where each subsequent base learner is adjusted to the residual of the current ensemble's training objective. The outcome of the adjusted base learner is then scaled by the learning rate and added to the ensemble. Random Forest, on the other hand, is a widely used machine learning method for creating prediction models in various research scenarios. Its primary goal is often to minimize the number of necessary variables for predictions, thereby reducing data collection burdens and improving overall efficiency. The Random Forest Classification method utilizes ensemble learning, merging numerous decision trees that are constructed randomly [11]. Each decision tree in the ensemble is built using a random subset of the training data and a random subset of predictor variables.

Misclassification Rate, also known as the error rate, is a common evaluation metric for classification models. The misclassification rate is calculated as the ratio of the total misclassified instances to the total instances in

the dataset [12]. In this research, the model with the lowest misclassification rate will be selected, as a lower value indicates better model performance. The method of calculating the misclassification rate is as follows:

$$Misclassification\ Rate = \frac{FN + FP}{N} \quad (1)$$

Where FN is False negative, FP is False Positive, and N is the amount of data.

3. RESULTS AND DISCUSSION

3.1. Business Understanding

The first stage, known as business understanding, will begin by comprehending the problem faced by Ivory Bank, which involves an increasing number of customers leaving their credit card services and a decrease in the customer base, as explained in the earlier section, 'Background & Business Understanding.' Additionally, the research aims to predict customers who are likely to churn from Ivory Bank's credit card services. The ultimate objective is to approach these customers with improved services to change their decisions and encourage them to continue using Ivory Bank's credit card services.

3.2. Data Understanding

In the data understanding stage, data related to Ivory Bank's customers will be collected for analysis. Visualization techniques will also be employed to facilitate a better understanding of the data. The dataset is available in .csv format and will be uploaded to SAS Visual Analytics. As can be seen in Figure 2, it consists of approximately 10,127 rows and twenty-three columns, containing customer information such as age, gender, education level, salary, marital status, credit card limit, card category, and more. The goal of this stage is to gain in-depth insights into the data, which will aid in data preparation and modeling since a profound understanding of the data is crucial at this stage.

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
CLIENTNUM	10127	739177606	36903783	7.48565E12	708082083	828343083
Customer_Age	10127	46.32596	8.01681	469143	26.00000	73.00000
Dependent_count	10127	2.34620	1.29891	23760	0	5.00000
Months_on_book	10127	35.92841	7.98642	363847	13.00000	56.00000
Total_Relationship_Count	10127	3.81258	1.55441	38610	1.00000	6.00000
Months_Inactive_12_mon	10127	2.34117	1.01062	23709	0	6.00000
Contacts_Count_12_mon	10127	2.45532	1.10623	24865	0	6.00000
Credit_Limit	10127	8632	9089	87415795	1438	34516
Total_Revolving_Bal	10127	1163	814.98734	11775818	0	2517
Avg_Open_To_Buy	10127	7469	9091	75639977	3.00000	34516
Total_Amt_Chng_Q4_Q1	10127	0.75994	0.21921	7696	0	3.39700
Total_Trans_Amt	10127	4404	3397	44600182	510.00000	18484
Total_Trans_Ct	10127	64.85869	23.47257	656824	10.00000	139.00000
Total_Ct_Chng_Q4_Q1	10127	0.71222	0.23809	7213	0	3.71400
Avg_Utilization_Ratio	10127	0.27489	0.27569	2784	0	0.99900
Naive_Bayes_Classifier_Attritio	10127	0.16000	0.36530	1620	7.6642E-6	0.99958
VAR23	10127	0.84000	0.36530	8507	0.0004200	0.99999

Figure 2. Ivory Bank's Dataset Basic Statistics

3.3. Data Preparation

The data preparation stage is a crucial step taken to ensure that the data is well-prepared before proceeding to the modeling phase. This preparation process is executed in SAS Data Studio, and the output is presented as a comprehensive plan. In the context of the Ivory Bank dataset, several essential steps are conducted to prepare the data for further analysis. The first step involves dropping unnecessary columns from the Ivory Bank dataset. Specifically, there are two columns, namely 'Naive_Bayes_Classifier_Attritio…', that do not contribute to the analysis and are, therefore, deemed unnecessary. These columns are promptly dropped from the dataset during the initial planning phase, streamlining the data for subsequent modeling. The second step focuses on removing any duplicate data that may be present in the Ivory Bank dataset. To achieve this, rows with the same CLIENTNUM are identified and subsequently dropped, ensuring that there are no duplicated entries in the dataset. This data cleansing process enhances the data's quality and integrity, providing a reliable foundation for further analysis. Next, the data is filtered to specifically target customers who possess Blue Cards. This decision is based on prior data visualization, which indicated that customers with Blue Cards exhibit the highest churn rate. As a result, the filter is applied to isolate and present only those customers with Blue Cards, enabling a more targeted analysis of this group. Lastly, the data is partitioned into training and testing sets. The training set comprises 80% of the data, while the remaining 20% constitutes the testing set. This partitioning ensures that the model will be trained on a substantial portion of the data,

enabling it to learn patterns and relationships effectively. Subsequently, the testing set can be used to evaluate the model's performance and generalization capabilities. By meticulously executing these data preparation steps, the Ivory Bank dataset is primed for the subsequent modeling process in SAS Data Studio. The carefully curated data, devoid of unnecessary columns and duplicate entries, allows for more accurate and focused analysis, leading to more insightful and reliable predictions and conclusions.

3.4. Modeling

In the modeling stage, the focus shifts to building the two selected models, Gradient Boosting, and Random Forest, as previously mentioned. These models will be used to predict customer churn in the credit card services of Ivory Bank. To begin the modeling process, specific roles are assigned to the variables in the dataset. The "Attrition Flag" variable is designated as the Response variable, representing the target variable that the models will aim to predict accurately. On the other hand, several predictor variables are identified, including "Customer_Age," "Marital_Status," "Gender," "Income_Category," "Credit_Limit," and "Total_Trans_Ct." These predictor variables will be used to make predictions and understand their influence on the "Attrition Flag." Both the Gradient Boosting and Random Forest models will be developed and evaluated using these designated roles. The goal is to assess their performance in accurately predicting the "Attrition Flag," which is essential for understanding customer churn patterns and identifying influential factors within the credit card services offered by Ivory Bank. By leveraging these advanced modeling techniques and incorporating crucial predictor variables, the research aims to enhance the bank's understanding of customer churn behavior. The insights gained from these models will enable Ivory Bank to take initiative-taking measures and design effective strategies to retain valuable customers, optimize services, and foster long-term customer relationships. The success of these predictive models will contribute significantly to Ivory Bank's efforts in reducing customer churn and ensuring sustainable growth in its credit card services.

3.5. Evaluation

In the evaluation stage, the models will be compared based on their misclassification rates, and the model with the lowest misclassification rate will be considered the best and chosen for predicting customer churn.

Following the evaluation, the model with the smallest misclassification rate will be selected to proceed with the implementation of customer churn prediction at Ivory Bank. During the deployment phase, the chosen model will be utilized to make predictions on customer churn in Ivory Bank's credit card services. The insights and predictions generated by the selected model will help the bank in taking initiative-taking measures to address customer churn, retain valuable customers, and enhance overall service quality. By implementing the most accurate and effective model, Ivory Bank aims to optimize its strategies, improve customer satisfaction, and foster long-term loyalty among its clientele. As can be seen in Table 1, the results of the analysis showed that the Gradient Boosting model achieved a remarkably low misclassification rate of 0.1118, indicating its superior performance in accurately predicting customer churn. On the other hand, the Random Forest model, while still performing well, had a comparatively higher misclassification rate of 0.1359. This means that the Gradient Boosting model exhibited better precision in classifying customers, making it a more reliable choice for predicting potential churners and retaining valuable customers. The success of the Gradient Boosting model can be attributed to its ability to build a strong ensemble of base learners sequentially. By adjusting subsequent learners to the residual errors of the previous ones, the model can iteratively improve its predictive capabilities, leading to more precise outcomes. In contrast, the Random Forest model's strength lies in its ability to create a diverse set of decision trees by using random subsets of data and predictor variables, which helps in reducing overfitting and improving generalization.

Table 1. Misclassification Rate

Model	Misclassification Rate
Gradient Boosting	0.1118
Random Forest	0.1359

3.6. Deployment

In this stage, the model with the lowest misclassification rate will be selected to proceed with the implementation of customer churn prediction at Ivory Bank. As can be seen in Figure 3, there is a variable importance measure that assesses the contribution or impact of each predictor variable on the model's performance. This measure helps identify the most influential variables for accurate predictions. In the Gradient Boosting model, the most impactful variable is "Total_Trans_Ct," while "Income Category" has less significance. Additionally, partial dependence is employed to understand the relationship between predictor variables and prediction outcomes in the predictive model. It identifies how predicted outcomes change with different values of a specific predictor while keeping other predictors constant.

Furthermore, the Confusion Matrix is a commonly used evaluation metric for classification models. It provides a concise summary of the model's predictions compared to the actual class labels in the data. The Confusion Matrix for the Gradient Boosting model shows True Positive 7683, True Negative 698, False Positive 234, and False Negative 821. Using SAS Visual Analytics and Model Comparison from SAS Visual Statistics, a superior model was identified and selected for predicting bank churners. The chosen model is the Gradient Boosting model with a misclassification rate of 0.1118. In comparison, the Random Forest model has a misclassification rate of 0.1359. The Gradient Boosting model provides a better prediction. Therefore, for predicting customer churn at Ivory Bank, the Gradient Boosting model is preferred.

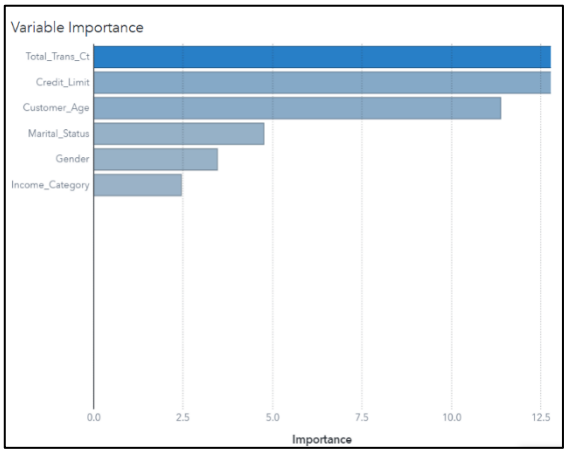

Figure 3. Variable Importance

4. CONCLUSION

The CRISP-DM framework was instrumental in the successful development of an appropriate model for predicting customer churn at Ivory Bank, with the Gradient Boosting model emerging as the preferred choice. This model highlighted superior performance compared to the Random Forest model, evident from its lower Misclassification Error rate. Specifically, the Gradient Boosting model achieved an impressive misclassification rate of 0.1118, outperforming the Random Forest model, which had a higher misclassification rate of 0.1359.

Due to its more accurate predictive capabilities, the Gradient Boosting model is suggested for future customer churn predictions at Ivory Bank. By leveraging this model, the bank can proactively identify customers who are likely to churn from their credit card services. Armed with these valuable insights, the business manager can take initiative-taking measures to retain these customers, design personalized retention strategies, and enhance customer satisfaction. Ultimately, implementing the Gradient Boosting model empowers Ivory Bank to optimize its customer retention efforts and bolster its overall performance in the competitive banking landscape.

For more accurate research findings, it is recommended to compare several types of classification models, such as Decision Trees, Naive Bayes, Neural Networks, and others. Additionally, using different evaluation metrics like F1-Score, AUC-ROC, Recall, and others would be beneficial.

ACKNOWLEDGEMENT

Thank you to Multimedia Nusantara University for the funding and support provided, which made this research possible from start to finish.

REFERENCES

[1] S. Nurjannah, "Digital Transformation In The Banking Industry Challenges And Opportunities," *Iternational J. Accounting, Manag. Econ.*, vol. 1, no. 01, 2023.

[2] A. Palandurkara, D. Bawankule, A. Bagde, F. Deshpande, S. Kamble, and S. Shende, "Customer Churn Prediction in Banking Environment using Data Science".

[3] H. Jain, G. Yadav, and R. Manoov, "Churn prediction and retention in banking, telecom and IT sectors using machine learning techniques," in *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019*, Springer, 2020, pp. 137–156.

[4] H. Hardjadinata, R. S. Oetama, and I. Prasetiawan, "Facial Expression Recognition Using Xception And DenseNet Architecture," in *2021 6th International Conference on New Media Studies (CONMEDIA)*, 2021, pp. 60–65.

[5] N. T. Msweli, T. Mawela, and H. Twinomurinzi, "Massifying Data Science Education through Immersive Datathons," 2023.

[6] W. H. Wijaya, R. S. Oetama, and F. A. Halim, "Implementation of Backpropagation Method with MLPClassifier to Face Mask Detection Model," *IJNMT (International J. New Media Technol.*, vol. 9, no. 2, pp. 48–55, 2022.

[7] S. J. Haddadi, M. O. Mohammadi, M. Bahrami, E. Khoeini, M. Beygi, and M. H. Khoshkar, "Customer churn prediction in the iranian banking sector," in *2022 International Conference on Applied Artificial Intelligence (ICAPAI)*, 2022, pp. 1–6.

[8] A. Kinge, Y. Oswal, T. Khangal, N. Kulkarni, and P. Jha, "Comparative study on different classification models for customer churn problem," in *Machine Intelligence and Smart Systems: Proceedings of MISS 2021*, Springer, 2022, pp. 153–164.

[9] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, "Propension to customer churn in a financial institution: A machine learning approach," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11751–11768, 2022.

[10] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H. S. Alghamdi, "Prediction of employee attrition using machine learning and ensemble methods," *Int. J. Mach. Learn. Comput*, vol. 11, no. 2, pp. 110–114, 2021.

[11] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," *Comput. Chem. Eng.*, vol. 128, pp. 392–404, 2019.

[12] J. L. M. Arruda, R. B. C. Prudêncio, and A. C. Lorena, "Measuring instance hardness using data complexity measures," in *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part II 9*, 2020, pp. 483–497.