National College of Ireland

Customer Retention Management with Predictive Data Mining: An Indonesian Banking Case Study

MSc Research Project
Data Analytics

Thi Thanh Thuy Nguyen
Student ID: x22107720

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Thi Thanh Thuy Nguyen
Student ID:	x22107720
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	14/12/2023
Project Title:	Customer Retention Management with Predictive Data Mining: An Indonesian Banking Case Study
Word Count:	6716
Page Count:	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Thi Thanh Thuy Nguyen
Date:	30th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Customer Retention Management with Predictive Data Mining: An Indonesian Banking Case Study

Thi Thanh Thuy Nguyen
x22107720

Abstract

This research project is centered on Indonesian bank's customer loyalty. It employs innovative analytics techniques such as random forests (RF), Catboost, XGBoost (extreme gradient boosting), and deep neural networks to explore transactional data. To make the analysis more understandable, SHAP (SHapley Additive exPlanations) values should be utilized to provide actionable information from historical customer behaviors. The primary objective of the study is to determine the significant attributes of loyalty-based strategies and how it affects customer retention. SHAP values would benefit in analyzing the importance of features, making predictions at the individual basis, and comprehending features relationship. By conducting a comparative analysis and showing outcomes, the study seeks to discern critical factors that shape customers' churn behavior. The research is significant in assisting banking service providers to develop customer-oriented strategies that will improve customer satisfaction and increase profits. The use of applying SHAP values shall make the analysis more transparent and clearer, enabling banks to make evidence-based decisions and prevent their customers from switching to a competitor. The findings show that XGBoost classifier is proposed as the most suitable one in predicting churner for customer retention strategy with actionable insights obtained.

1 Introduction

Nowadays, in a fierce competitive free market, businesses understand that customer relationships are indispensable for long-term success and recognizing the critical necessity of customer retention over customer acquisition is grounded in several strategic considerations for businesses (Pfeifer Phillip E; 2005). The advancements in data science and digitalization have enabled enterprises to easily accumulate a massive database filled adequately with the demographics, economic characteristics, and behavioral patterns of their customer base (Asadi et al.; 2017). As an integral component in retaining customers, identifying potential churners – those more likely to put an end to services - poses challenges for conventional analytical tools due to the large amount of information, commonly known as 'big data' (Bhadani and Jothimani; 2016). To address this challenge, taking advantage of historical data and employing machine learning (ML) algorithms for customer attrition prediction becomes essential to alleviate revenue loss through personalized marketing and service customization strategies (Verbeke et al.; 2011). Consequently, customer retention, defined as the ability to keep existing customers advocated, gratified,

and faithful to a product or service, has garnered significant attention from scholars and practitioners alike (De and Prabu; 2022).

In the constantly altering banking domain, where technological innovations, governing changes and consumer aspirations meet, the idea of customer retention serves as the critical necessity for sustained growth and stayed gamesmanship (Mbithi; 2013). Efforts to hold onto customers and avert their potential shift to competitors are often hampered by the sheer magnitude of transactional data at hand. In fact, handling customer retention strategies in a bank with a massive customer base poses challenges in cost-effective edge and resource allocation. Thus, advanced techniques are indispensable for churn prediction in banking when it comes to dealing with personalized retention efforts on a timely basis which can be a more cost-effective and sustainable approach in the era of data-driven customer centricity (Ngai and Wu; 2022). In this context, this study concentrates on the important pattern of customer retention within the Indonesia's banking domain.

1.1 Background and Motivation

An preliminary examination of previous research has pointed out disadvantages in the model appraisal, interpretative assessment, and data-informed implementation. Previous studies on identifying churners in the banking field have particularly focused on addressing the significant challenge of data imbalance and demonstrating the benefit of churn prediction, as exemplified by works such as Kumar, Tirupathaiah and Krishna Reddy (2019), and Kavyarshitha, Sandhya and Deepika (2022). Additionally, retention efforts may be challenged by budget constraints that may arise during the implementation of churn prevention campaigns. Especially, when dealing with imbalanced data that has not been thoroughly considered despite model development completed. It is appear to be not noteworthy that Kavyarshitha et al. (2022) uniquely used 'Accuracy' as a metric to measure model effectiveness. Furthermore, the growing popularity of deep learning in banking customer attrition research, driven by its enhanced correctness extraordinarily, introduces intricate issues related to encompassing front-to-end and applying these constructed models in practical, data-driven framework, as also emphasized by Kavyarshitha et al. (2022). Literally, prior studies have not fully gotten to grips with these challenges, and the objective of this project lies in bridging this gap.

1.2 Research Question

The formulation of hypotheses serves as a crucial aspect of our inquiry. Drawing upon established theories in banking and insights from previous empirical research, our hypotheses seek to test and refine our understanding of the drivers of customer retention specific to the banking sector. Through meticulous empirical analysis, we aspire to validate or challenge existing assumptions, offering a nuanced perspective on the intricate dynamics that underpin customer loyalty in a sector characterized by constant innovation and evolving customer preferences.

The formulation of hypotheses with following key questions serves as an important aspect of our investigation will be addressed in this project:

1. How to benefit advanced developments of analytical techniques for preventing churn in the banking domain ?

2. How to the obtained model fulfill its role as a data-informed customer loyalty management?

3. How do the valuable insights acquired from SHAP values establish the development of personalized retention efforts for institutions?

4. How can regular practices approaches be integrated with comprehension from transactional data for building a better customer experience?

1.3 Research Objectives

This research's objective is to fill the gap in the existing literature related to customer retention within in the fast-paced environments of Indonesian's retail banking, aiming to figure out valuable comprehend and solutions for maximizing the customer lifetime value in the banking industry through an investigation of contemporary data mining technologies and advanced analytics techniques. The positive impact lies in introducing a step wise framework for building a model to predict and mitigate churn leveraging Machine Learning (ML) and Deep Learning (DL) techniques. This framework enables banks to proactively recognize clients at risk of churning, allowing for targeted marketing efforts more adaptable to evolving scenarios. The second contribution implies in integrating SHAP values to support the model's explanatory power. This technique accelerate the interpretation of the model's forecast, offering banking stakeholders a clearer understanding of factors influencing customer loyalty. The data-driven approach upheld by this study is ready to assist in making informed decisions, resulting in more effective and adaptive targeted retention strategies. Moreover, by identifying effective retention strategies, the study empowers banks to maximizing the customer lifetime value, ultimately leading to increased profits and contributing positively to the bank's reputation.

This remainder of this document is organized as follows: Section 2 is a comprehensive assessment of the existing literature on customer retention within this domain. Section 3 outlines the research methodology and specifications employed in this study. Following that, Section 4 presents the techniques and framework that will be employed in this study. In Section 5, a comprehensive analysis of the results and main findings of the study is provided. Ultimately, Section 6 encapsulates the conclusions.‘

2 Related Work

Even if the customer turnover rate of the banking sector is not as high as other industries, the financial implications of losing a customer are substantial (Elyusufi and Kbir; 2022). Consequently, effectively calculating and detecting potential churners becomes a crucial aspect of customer retention management in banking business aiming to enhance long-term profitability (Elyusufi and Kbir; 2022). Embracing this challenge, there is a growing trend of using machine learning and data mining techniques within the banking sector to detect churn risks and refine customer retention efforts (Kumar Hegde et al.; 2023).

This initial literature review provides a detailed exploration of prior research regarding customer retention management practices within the banking domain. It puts forward valuable insights into fundamental concepts and findings about research gaps. Additionally, it calls attention to the utilization of data mining techniques in earlier studies as a powerful tool to escalate customer retention strategies.

2.1 Customer Retention Management Practice in the Banking Sector

Banking businesses employ customer retention management as a master plan to retain at-risk customers and improve overall customer satisfaction, with the success of this approach relying on the use of proactive retention strategies (Krishna et al.; 2022). Numerous previous studies have stated the importance of client-centric approach, fully integrating high-quality services, leveraging new technology, and addressing client inquiries to maximize customer lifetime value (Sugiato et al.; 2023). Notably, relationship marketing has been identified as a high-powered tool for mitigating churn and winning the benefits of customer loyalty, as asserted by (Ganaie and Bhat; 2020).

The role of service quality in customer retention has been generally known in the literature (Supriyanto et al.; 2021). Enhanced service quality not only leads to greater customer experience but also paves for long-term customer satisfaction and loyalty. However, Supriyanto et al. (2021) suggests that various other characteristics can dominate this relationship, implying that analyzing customer behavior through transaction history data can offer valuable insights for implementing personalized retention initiatives.

Customer Relationship Management (CRM) is identified as a pivotal component of customer retention within the banking sector, as explored by Krishna et al. (2022). The study asserts that effective CRM practices greatly contribute to building sustainable and profitable customer base, meeting customer expectations, enhancing brand reputation, and staying competitive in the market. Given the impracticality of servicing all customers equally, especially for businesses with wide customer bases, Krishna et al. (2022) indicated the necessity of advanced analytics for dynamically identifying niche segments to keep high-value customers.

While research on customer retention strategies in the banking sector remains somewhat limited, consistent findings emphasize the relevance of factors such as service quality, product innovation, and customer service in building customer loyalty. The suggestion to derive valuable insights into customer behavior before formulating retention strategies (Mecha et al.; 2015) demonstrated the importance of integrating predictive data mining techniques into targeted retention initiatives.

2.2 Data Mining Techniques for Customer Retention in the Banking Sector

Empirical studies focusing on the prediction of potential churners in the banking sector have been prevalent since the early twentieth century, as exemplified by the case study conducted by Bounsaythip and Rinta-Runsala (2001). Over the past decade, the evolution of this field has transitioned from conventional machine learning approaches, as highlighted in works such as Özden Gür Ali and Arıtürk (2014), Sundarkumar and Ravi (2015), Van den Poel and Larivière (2004), A. and D. (2016), and He et al. (2014), towards more advanced methods with the advent of deep learning. This emergence represents a contemporary trending in technological innovation within the big data era, as evidenced by research on customer retention from G. Ravi Kumar (2019), Geiler et al. (2022), Bilal Zorić (2016), Patil et al. (2022), AL-Najjar et al. (2022), and Tariq et al. (2022). In general, machine learning remains a widely used and straightforward approach in this domain. However, the contemporary shift towards deep learning signifies a more sophisticated methodology. Deep learning, while offering the advantage of reduced effort

in determining feature importance, comes with the trade-off of increased computational resources required for training (Xin et al.; 2018).

Random Forest (RF) is a tree-based algorithm composed of a collection of trees, where each tree depends on a random selection of features independently sampled for all the trees in the forest from the same distribution (Breiman; 2001). Numerous studies have explored the application of RF in predicting banking churn. Notably, the Random Forest model introduced by Geiler et al. (2022) demonstrated enhanced prediction accuracy and successfully addressed challenges associated with imbalanced data.

Catboost or Categorical Boosting is a robust and efficient gradient boosting library explicitly developed to handle categorical features in machine learning models. Formulated by Dorogush et al. (2018), it is a relatively recent entrant in the churn prediction field but is gaining attention across various prediction domains for its ability in effectively dealing datasets containing both numerical and categorical features. In empirical studies, such as those referenced by Sagala and Permai (2021), the algorithm demonstrated stable performance even at default parameters. Moreover, Catboost achieved the highest accuracy of 97.85% compared to other boosting algorithms in this study.

Extreme Gradient Boosting, commonly known as XGBoost, is a tree-based machine learning technique designed for supervised learning tasks (Chen et al.; 2015). The fundamental concept of this algorithm involves boosting predictive power by amalgamating multiple decision trees using ensemble learning techniques. Notably, XGBoost has gained popularity for its proficiency to carry out large datasets with high accuracy and efficiency, making it a scalable and effective method increasingly favored in data science challenges and real-world applications (Sanders et al.; 2022).

Artificial neural networks (ANN), which are deep learning algorithms, have become more prevalent in recent churn studies, as highlighted in works like Kavyarshitha, Sandhya and Deepika (2022) and Kellner, Nagl and Rösch (2022). These studies demonstrated that neural network system models can handle a broader range of complexities compared to traditional machine learning methods. However, the 'black box' nature of this technique poses challenges in understanding model results, as mentioned by Kellner et al. (2022).

In addition to the insights gained into the effectiveness of applied algorithms in churn studies, a notable limitation identified is the insufficient exploration of models for imbalanced datasets. This oversight could pose challenges in cost management when implementing the model in customer retention efforts. It worthy noted that Kavyarshitha et al. (2022) solely used 'Accuracy' as a metric to gauge the model's effectiveness.

In conclusion, the research landscape on customer retention in the banking industry, employing data mining techniques, has gone through significant growth. Early studies relied on traditional and machine learning approaches, with recent research integrating more advanced techniques such as deep learning. Noteworthy methods include Random Forest, Catboost, XGBoost, and Artificial Neural Networks (ANNs), each of which offers distinct advantages. Catboost excels in handling categorical features, XGBoost demonstrates scalability and high accuracy, Random Forests combine interpretability with performance. While deep learning methods have demonstrated highly accurate predictions, they face challenges due to their 'black box' nature, characterized by a lack of transparency and interpretability in data-informed processes.

2.3 SHAP Values in Improving Model Interpretability and Feature Importance Analysis

In recent years, machine learning models have shown remarkable predictive capabilities, especially churn prediction for customer retention (De and Prabu; 2022). However, the inherent black-box nature of advanced algorithms, including deep learning algorithms, can pose a problem to interpretability, making it challenging to explain the rationale behind each prediction (Kellner et al.; 2022). Addressing this challenge, SHAP (SHapley Additive exPlanations) values have emerged as an effective and intuitive approach for understanding complex machine learning models (Ekanayake et al.; 2022). While studies on SHAP values in the context of bank churn for customer retention management are relatively recent, some noteworthy examples include the work of Jovanovic, Kljajic, Mizdrakovic, Marevic, Zivkovic and Bacanin (2023) and Cao (2021).

In the study by Cao (2021), SHAP values has been utilized to explain the importance of features which emphasize the necessity of model explainability in achieving reliable outcomes from AI churn prediction models. Similarly, in the research conducted by Jovanovic et al. (2023), SHAP analysis has been employed to assess the significance of attributes modeled by XGBoost. Importantly, the insights derived from the model explained by SHAP values offer reliable concepts and recommendations on key elements that directly influence the decision to discontinue using the bank's services (Jovanovic et al.; 2023).

In the general landscape of studies, previous research efforts have mainly concentrated on constructing models that deal with challenges related to imbalanced data and not addressed the model constraints. Additionally, numerous studies have fallen short in furnishing a data-driven fundamental explanation for their models. In attending this research gap, the motivation for this project arises from the need to develop an effective framework for mitigating client attrition, aiming to bridge the existing gaps in understanding and interpretation.

2.4 Research niche

This study seeks to rectify a gap present in the existing literature regarding customer retention within the Indonesian retail banking sector. Its primary objective is to introduce a data mining framework capable of yielding actionable insights for the improvement of proactive retention initiatives. Furthermore, this research aims to utilize SHAP values to address the complexities and interpretability challenges associated with the application of data mining techniques in practical business contexts. By doing so, the study endeavors to enhance interpretative clarity, thereby enabling financial institutions to effectively leverage insights derived from data mining for informed decision-making. This strategic approach is anticipated to contribute to the formulation and implementation of more successful, customer-centric strategies within the banking sector.

3 Methodology

3.1 Research Method

It is imperative for customer retention management in the banking sector to pinpoint clients more prone to discontinuing services, enabling the concentration of retention

strategies on the appropriate target customers. Consequently, the experimental research aimed at anticipating churners will leverage a publicly available banking dataset within the Indonesian context to construct a predictive model. In essence, churn prediction is treated as a supervised mathematical problem, with the primary goal of identifying potential churners through the application of Machine Learning (ML) and Deep Learning (DL) in model training. Subsequently, the selected model will undergo ongoing examination to elucidate and provide insights for customer retention management.

The procedural framework for executing data mining projects adheres to the Cross Industry Standard Process for Data Mining (CRISP-DM) (Schröer et al.; 2021), customized to align with the banking business. This framework, delineated into six phases, serves as the structure for the CRISP-DM domain and is visually represented in Figure 1.

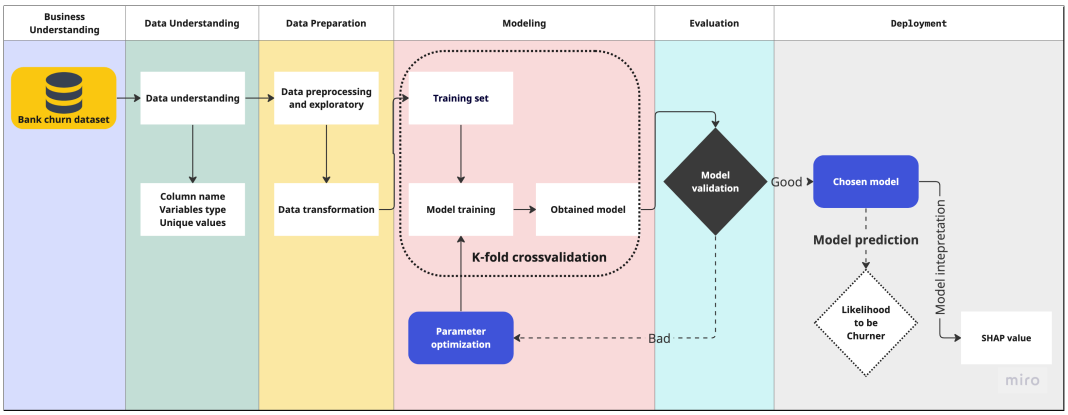

Figure 1: Flowchart of conducting customer retention efforts

3.2 Dataset and Business Understanding

Customer turnover, or churn rate, as used in the banking industry, is a person who stop all of their accounts or stops doing business with a bank. Customer turnover can negatively impact on revenue streams, brand advocates and earnings depreciation. Consequently, finding and detecting clients that have a high propensity to quit the service; or predicting and mitigating churn is a crucial part of customer-oriented retention, which aims to reduce the turnover rate (Silveira, Pinheiro and Junior; 2021).

The sample 'banking churn' collected from Kaggle (2019) comprises data pertaining to the demographics, accounts, and transactional characteristics of its client. There are 28,382 rows in the dataset, each with a distinct identification, and 21 columns of customer's features. More specifically, the dataset incorporates variables of numerical, categorical, textual, and temporal nature, covering information pertaining to the following areas: Details about clients, Data related to segments, Behavior within customer segments, Account balance history and Account transaction history. The target feature 'churn', being a binary variable, provides information on whether or not the customer has stopped the service. This feature is instrumental in employing data mining techniques to predict potential churners.

In overall, the dataset includes 18.5% of churning customers and 81.5% of non-churning customers as illustrated in the Figure 2. This ratio indicates that the distribution of two classes is imbalanced. As a result, the 'Accuracy' will not suffice to assess the model's performance and the use of k-fold during model development is necessary. Furthermore, the data set contains 5 categorical variables and 14 numerical

variables. In which numerical variables mainly reflect account balance history and some categorical variables represent demographic information. These initial observations lay the groundwork for further analysis and model development.

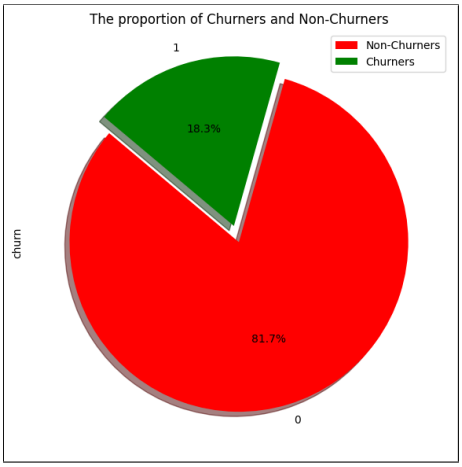

Figure 2: The imbalance of Churners and Non-Churners in the dataset

3.3 Data Preparation

Following the completion of data cleaning, univariate, multivariate, and correlation analyses will be conducted to assess the efficacy of variables and gain deeper insights into the dataset.

3.3.1 Data pre-processing

Data pre-processing is an important preliminary step that ensures the banking dataset is in standard condition for analysis and modeling, resulting in more accurate and reliable results. Table 1 summarizes the data pre-processing stage of the given dataset. Generally, the bulk of variables are complete with no duplicated records. However, there are substantial missing values in the 'gender', 'dependents', 'occupation', and 'city' columns of the dataset. The handling of missing values will be decided by the data type of each feature. Regarding the two categorical variables 'gender' and 'occupation', they should be treated separately as a new category 'other' or 'unknown' based on the assumption that the customer did not wish to share their personal information. The purpose of creating this new data is to avoid removing data or filling in with an existing value, which could lead to bias. For 'dependents' variables, a continuous numeric data, missing values would be replaced by the median value of itself variable.

The number of unique values of each variable will be considered to reclassify variables as either numerical or categorical. This approach is preferred over relying solely on data types, so that it can reduce the risk of ambiguity, especially when dealing with categorical variables that may comprise of numeric values. As a result, the sample data is enhanced by a wide range of numerical variables such as 'vintage', 'age', and numerous transaction-related indicators. Categorical attributes such as 'gender', 'occupation' and 'customer_nw_category' provide a clearer perspective on customer segmentation, while 'churn' serves as the binary outcome of interest, with 0 representing 'non-churners' and 1 representing 'churners' which is the anticipated target in this research.

Some columns have been dropped out of the dataset because of unnecessary or inadequate information for model development such as customer id, last transaction, city and branch code.

Table 1: Summary of pre-processing stage

Column name	**Count**	**# Missing value**	**# Unique value**	**Dropped column**
customer_id	28,382	0	28,382	**Yes**
vintage	28,382	0	1,459	No
age	28,382	0	90	No
gender	27,857	**525**	2	**No**
dependents	25,919	**2463**	15	**No**
occupation	28,302	**80**	5	**No**
city	27,579	**803**	1,604	**Yes**
customer_nw_category	28,382	0	3	**No**
branch_code	28,382	0	3,185	**Yes**
current_balance	28,382	0	27,903	**No**
previous_month_end_balance	28,382	0	27,922	**No**
average_monthly_balance_prevQ	28,382	0	27,801	**No**
average_monthly_balance_prevQ2	28,382	0	27,940	**No**
current_month_credit	28,382	0	10,411	**No**
previous_month_credit	28,382	0	10,711	**No**
current_month_debit	28,382	0	13,704	**No**
previous_month_debit	28,382	0	14,010	**No**
current_month_balance	28,382	0	27,944	**No**
previous_month_balance	28,382	0	27,913	**No**
churn	28,382	0	2	**No**
last_transaction	28,382	0	361	**Yes**

3.3.2 Univariate analysis

The statistical analysis of features is presented in Table 2, as well as illustrated by corresponding histograms in Figure 3. Notably, most variable distributions, as evidenced by histogram shapes, deviate from the bell shape. Specifically, the 'vintage' feature, which represents the length of a client's engagement with the bank, ranges from 73 to 2,476, with a mean of 2,091. In addition, the left tail is longer than another tail stating that the variable has negatively-skewed distributions. This also indicates that the majority of customers have been in relationship with the bank for a long time as the customers with fresher vintage only account for a small proportion.

Customer's age in this dataset has a mean of 48.2, ranging from 1 to 90. While customers's age under 16 might seem uncommon, actually certain banks offer specialized programs that allow parents to open savings accounts for their children.

Table 2: Statistical Analysis

Feature name	**Mean**	**Min**	**25%**	**50%**	**75%**	**Max**
vintage	2,091.1	73.0	1,958.0	2,154.0	2,292	2,476
age	48.2	1.0	36.0	46.0	60.0	90
dependents	0.3	0.0	0.0	0.0	0.0	52
customer_nw_category	2.2	1.0	2.0	2.0	3.0	3
current_balance	7,380.6	-5,504.0	1,784.5	3,281	6,636	5,905,904
previous_month_end_balance	7,495.8	-3,149.6	1,906.0	3,380	6,657	5,740,439
average_monthly_balance_prevQ	7,496.8	1,428.7	2,180.9	3,543	6,667	5,700,290
average_monthly_balance_prevQ2	7,124.2	-16,506	1,832.5	3,360	6,518	5,010,170
current_month_credit	3,433.3	0.0	0.3	0.6	707.3	12,269,845
previous_month_credit	3,261.7	0.0	0.3	0.6	749.2	2,361,808
current_month_debit	3,658.7	0.0	0.4	91.9	1,360	7,637,857
previous_month_debit	3,339.8	0.0	0.4	110.0	1,358	1,414,168
current_month_balance	7,451.1	-3,374.2	1,996.8	3,448	6,668	5,778,184
previous_month_balance	7,495.2	-5,171.9	2,074.4	3,465	6,655	5720145
churn	0.2	0.0	0.0	0.0	0.0	1
balance_difference	2,055.2	0.0	43.4	373.7	1,394	879,985
credit_utilization_ratio	72.0	-61.2	0.0	0.0	0.2	482,031
debit_credit_ratio	4347.1	0.0	1.0	1.0	31.2	11,000,001
balance_change_percentage	874.2	-3347.6	-12.2	0.0	7.3	7,317,767
avg_monthly_balance_change	186.3	-2,138,251	-233.5	49.7	605.6	511,383

There are significant differences observed between the values of the variables related to account balances as well as the fluctuations in account balances over a given period. Particularly, the customer's current account balance reveals a mean value of 7,380, ranging from -5,503 to 5,905,904, an extremely high value. And the same thing appears in the features of average monthly balance and credit utilization ratio. This characteristic provides an overview of the numerical landscape of the obtained data, emphasizing the diverse and non-normal nature of the variables under the banking context.

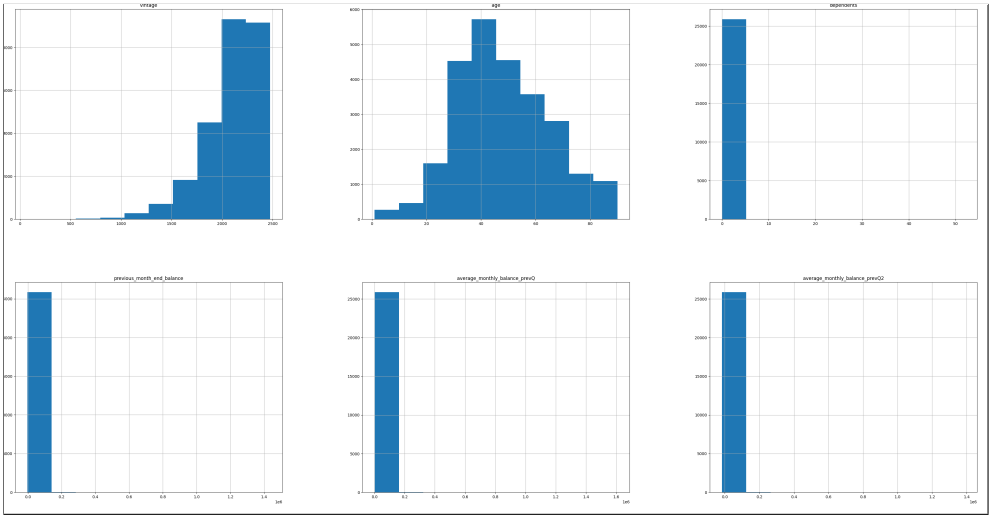

Figure 3: Exploratory Data Analysis

3.3.3 Multivariate analysis

Figure 4 displays the histogram and boxplot of the independent variables towards two classes of 'Churn', exploring the potential power of variables. No significant gap were found in the average age between churners and non-churners by box-plot. However, the histogram displayed a noticeable difference in the churn rate among different age groups, particularly, the ages between 25 and under 35. This age group appears to have a higher likelihood of stopping the service in comparison to other age groups. In addition, the bar plot of customer's gender also illustrates the fact that female customer are more likely to switch the services than male ones. Regarding the client's job type and net worth group, there is no significant difference between churners and non-churners.

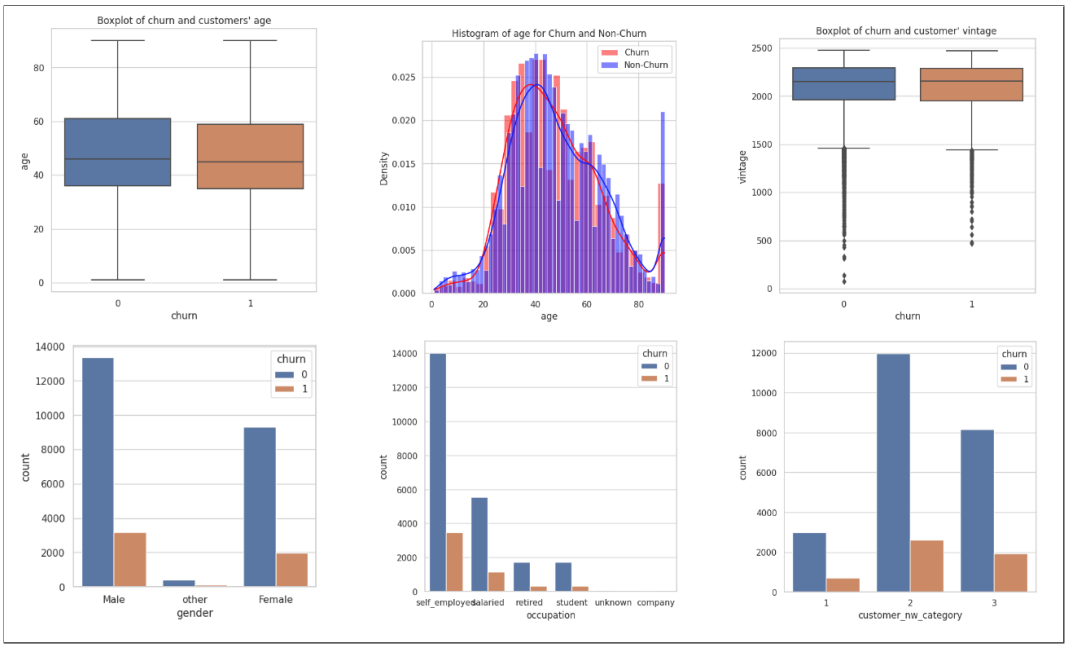

Figure 4: Mean comparison between independent features and churn

There is statistically significant association among feature regarding customers' account balance displayed in Figure 5. This can be assumed that there had no prominent change in the customer's account balance in the two consecutive months. Notably, the strong positive correlation between transaction of deposits and withdrawals with bank account stated that the more money customers deposit, the more they will withdraw. This suggests the idea that promoting account deposit or spending activities using bank accounts is beneficial in strengthening customer life time value.

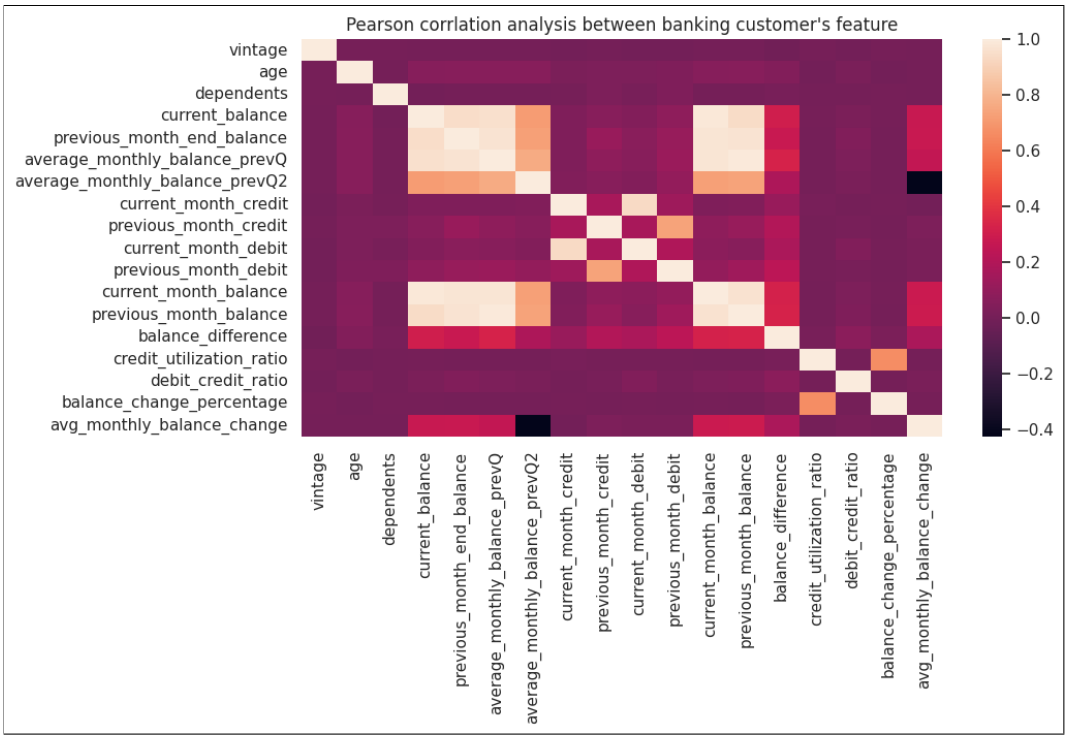

Figure 5: Correlation analysis between customers' features

These above observation can be of great importance in understanding consumer behavior and developing customer retention strategies that specifically address customer segmentation or interests based on the factors such as age, gender or obvious changes in their banking account balance.

3.3.4 Transformation

The process of transforming data used to numerically represent non-numerical variables so that they can be fed to machine learning algorithms is known as one-hot coding, or one-hot encoding. Categorical variables have a limited number of possible values within their categories. Consequently, these variables cannot be used directly in machine learning techniques required numerical inputs. One-hot coding shall be chosen as a method to split the original variable into binary columns to represent these categorical variables as numerical inputs (Zhang; 2023) (Al-Shehari and Alsowail; 2021).

One of the main advantage using one-hot coding is that it may allow the use of categorical variables as inputs to machine learning algorithms without imposing any kind of ordering on the categories. Therefore, it shall not alter the interpretable output of the model. In this study, customer's features in terms of **occupation**, **gender** and **net-worthy category** will be converted to numerical by using one-hot coding.

Customer_id	gender
1	Male
2	Female
5	other

Customer_id	gender_Male	gender_Female	gender_other
1	1	0	0
2	0	1	0
5	0	0	1

Figure 6: Transformation of attribute ■gender■ after using one-hot coding

Certain machine learning models perform better when the fed data is normalized. The normalization method in the applied models is the Min Max Scaler (MMS), because the data is essentially scaled down by MMS inside the intervals of [0, 1] and [-1, 1]. Following is a representation of the mathematical formula for min max scaling:

$$X' = \frac{x - min(x)}{max(x) - min(x)}$$

3.4 Evaluation

The confusion matrix is a widely used approach for measuring performance in binary classification problems. It comprises four initial indicators, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Kelleher et al.; 2015, p.402-404). More specifically, TP represents the number of churners accurately predicted to be 'Churn'; TN denotes the number of non-churners accurately predicted to be 'Non Churn'; FP represents the number of churners inaccurately predicted to be 'Non Churn'; and FN denotes the number of non-churners inaccurately predicted to be 'Churn'. In addition to the confusion matrix, precision, recall, specificity, and F_1 are more frequently used performance evaluation metrics that are directly taken out from this matrix. How effective each classifiers work shall be measured using these four indicators (Kelleher et al.; 2015, p.414-429). Furthermore, another measurement that will be included in this study is AUC-ROC, which assesses how well the developed model can classify between churner and non-churners. It is supposed to be one of the key metrics in various machine learning classification problem (Kaur and Kaur; 2020).

4 Design Specification and Implementation

The fundamental steps for putting the end-to-end retention initiatives into effect in the banking context are illustrated in Figure 7. This process also refers Jeong et al. (2023) and follow the structure of the CRISP-DM domain from the stage of data collection and preparation to the prediction model development and model deployment.

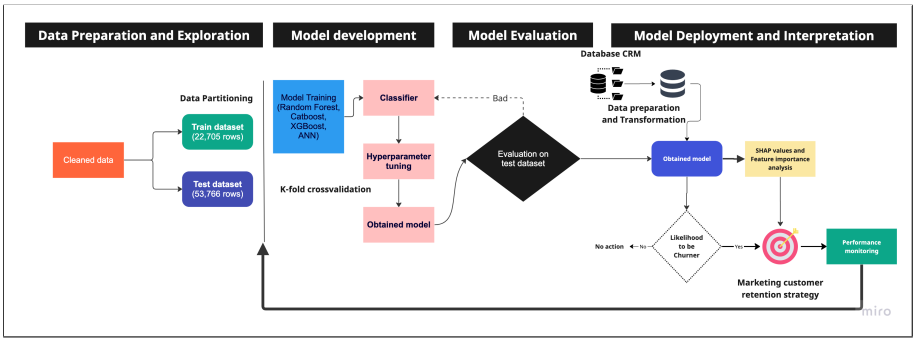

Figure 7: Model Baseline Training Output

Various sophisticated machine learning models, such as RF, Catboost, XGBoost, and ANN, would be studied to determine the best machine learning algorithm that can more accurately identify churners. Accordingly, several machine learning libraries, including Scikit-learn, numpy, matplotlib, pandas, tensorflow.js, catboost and xgboost should be called in Google Colab's notebook for model development. A cutoff ratio of 70% and

30% would be used to partition the dataset into training and testing, respectively. This produces 22,705 observations assigned to the training set and 5,677 instances distributed to the testing set. Finally, the training phase and testing phase with 30 customer features would be fed to machine learning and deep learning techniques for model development.

On notebook, the random seed would be set to ensure that the same sequence of random number shall be generated every time the code run. Due to the asymmetry of binary prediction, this study also utilize a stratified K-fold cross-validation during modelling process, following the earlier studies from Karvana, Yazid, Syalim and Mursanto (2019), where k-Fold has been found to be computationally advantageous and efficient to assess the model's generalization performance.

To examine the impact of sampling to classifers' performance, all models shall undergo baseline training for two input dataset (with sampling and without sampling), and performance metrics shall be monitored for comparative purposes. Prior to conduct hyperparameter optimization, this baseline training stage is supposed to provide an overview of how capable the models are toward the given problem. The baseline training results were illustrated in Table 3. Catboost outperformed other algorithms in terms of Accuracy, Recall and F1 Score. This preliminary result demonstrates Catboost's potential in churn prediction problems.

Table 3: Model Baseline Training Output

	Accuracy	**Recall**	**F1 Score**	**AUC**
CatBoost	**86.5766%**	**48.0355%**	**57.0139%**	83.6402%
RandomForest	86.5297%	44.2966%	54.9312%	**84.7317%**
XGBoost	86.4474%	47.0849%	56.2879%	82.4476%

Following the initial training, a procedure known as hyperparameter tuning will be applied to every single of the chosen classifier. Hyperparameter tuning is known as the process of choosing the best hyperparameter values for a machine learning classifier. It is imperative to fine-tune these hyperparameters in order to maximize the efficiency of machine learning algorithms. While there are several ways to tune hyperparameters, random search should be used in this study due to its computational friendliness. And the best set of tuning parameters will be considered after taking into account the characteristics of the dataset and the computational power at hand. K-fold cross validation also gets involved at this point for performance evaluation and avoiding over-fitting purpose. Figure 8 displays the tuning parameters chosen for each machine model along with the tuning outcomes.

```
params = {'XGBoost':{'learning_rate': [0.001, 0.01, 0.1, 0.2],
                     'n_estimators': [50, 100, 200, 500],
                     'max_depth': [3, 6, 9],
                     'objective': ['binary:logistic'],
                     'eval_metric': ['logloss']
                    },
          'CatBoost':{'iterations': [50, 100, 200, 500, 1000],
                      'learning_rate': [0.0001, 0.001, 0.01, 0.1],
                      'depth': [6, 8, 10],
                      'loss_function': ['Logloss', 'CrossEntropy']
                      },
          'RandomForest':{'n_estimators': [50, 100, 500, 1000],
                          'max_depth': [6, 8, 10],
                          'max_features': [2, 3, 4, 5],
                          'min_samples_split': [2, 5, 10],
                          'min_samples_leaf': [1, 2, 4]
                          }
          }
```

```
# Best params output
model_rs_best_param_1
```

```
{'RandomForest': {'n_estimators': 500,
  'min_samples_split': 5,
  'min_samples_leaf': 2,
  'max_features': 5,
  'max_depth': 10},
 'CatBoost': {'loss_function': 'CrossEntropy',
  'learning_rate': 0.01,
  'iterations': 1000,
  'depth': 8},
 'XGBoost': {'objective': 'binary:logistic',
  'n_estimators': 500,
  'max_depth': 6,
  'learning_rate': 0.01,
  'eval_metric': 'logloss'}}
```

Figure 8: Hyperparameter tuning setting and output

The model architecture for ANN referred to Munkhdalai, Munkhdalai and Ryu (2020) in handling imbalanced dataset, and optimized by using Bayesian techniques as it is computationally resource friendly compared to other techniques like Gridsearch or Random Search (Cho, Nam, Choi, Choi, Kim, Bae and Moon; 2021), and then improved during experimental. More specifically, the algorithm has been applied to choose the optimized number of epoches and learning rate. One of outputs from using Bayesian algorithms to optimize ANN is illustrated in Figure 9, and the architecture proposed is displayed in Figure 10.

iter	target	epochs	learni...
1	0.8147	169.6	0.003575
2	0.8147	122.7	0.005962
3	0.8147	171.9	0.004808
4	0.8146	198.1	0.007163
5	0.8147	154.9	0.006478

Figure 9: Bayesian optimization stages

```
Model: "sequential_39"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_207 (Dense)           (None, 30)                930

 dense_208 (Dense)           (None, 30)                930

 dense_209 (Dense)           (None, 20)                620

 dense_210 (Dense)           (None, 20)                420

 dense_211 (Dense)           (None, 10)                210

 dense_212 (Dense)           (None, 1)                 11

=================================================================
Total params: 3121 (12.19 KB)
Trainable params: 3121 (12.19 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 10: ANN architecture

After optimizing the classifier, SHAP or SHapley Additive Explanation shall be employed figure out the features that play an important role by the obtained model. SHAP value is a method for interpreting machine learning models. It estimates the impact of

each feature on the model output for a given instance to provide a local interpretation of the input space. The SHAP value is computed as the weighted sum of each feature, scaled by the magnitude of the feature, and multiplied by the predicted class label. In Python, the *shap* library is used to calculate SHAP values for the machine learning models specified.

The findings of this study indicate that there is no discernible difference between the sampling and non-sampling data based on baseline training and hyperparameter tuning output. In addition, to be more straightforward to explain model results in terms of SHAP value, the outcomes shown below are by data without sampling.

5 Evaluation

5.1 Customer Retention with Random Forest

With an AUC of 85% measured on testing set, the random forest model demonstrates its capacity for discrimination, displayed in Figure 11 and 12. Its classier shows a precision of 73% in identifying potential churners (Class 1), meaning that 73% of the time it is true when it predicts a customer will churn. Class 1 recall is 42%, meaning 42% of real churn instances are captured by the model. Additionally, the precision and recall-balancing F1-score is 54%, it implies a trade-off between reducing false positives and accurately detecting churn cases. The total number of correctly predicted cases was 667 (True Positive). Furthermore, the model shows no signs of overfitting based on the AUC of the training and testing set (90% and 85%).

	precision	recall	f1-score	support
0	0.88	0.96	0.92	6937
1	0.73	0.42	0.54	1578
accuracy			0.86	8515
macro avg	0.81	0.69	0.73	8515
weighted avg	0.85	0.86	0.85	8515

Figure 11: Model accuracy

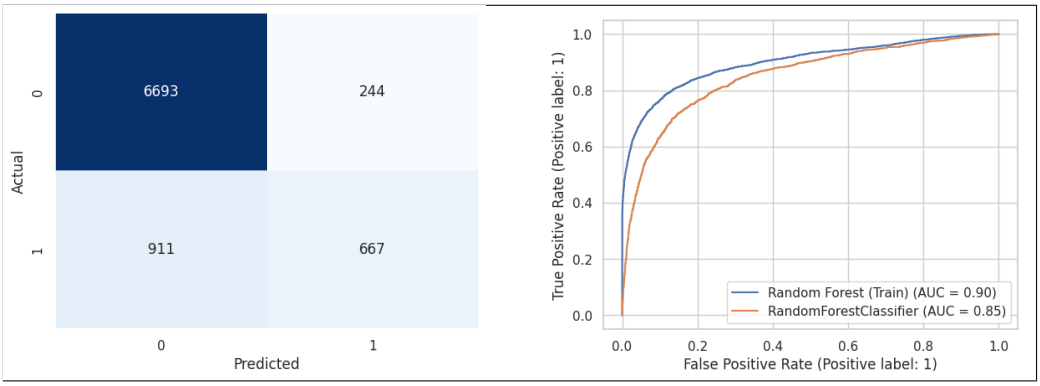

Figure 12: Random Forest performance in banking churn prediction

The following explanation illustrates the specific roles played by every single feature in modifying the model output from the base value. Features that have a positive influence

on risk of churn are shown in red, while features that have a negative influence are shown in blue.

The Random Forest classifier's top five features are shown in Figure 13, along with an explanation of one observation using the SHAP value. Predictions are significantly impacted by the current account balance, the percentage of changing balance, and the difference between the amount of credit and debit. More specifically, customers are more likely to discontinue using the service if the debit credit ratio or debit amount of current month is higher (>9,797 and >5,486, respectively).

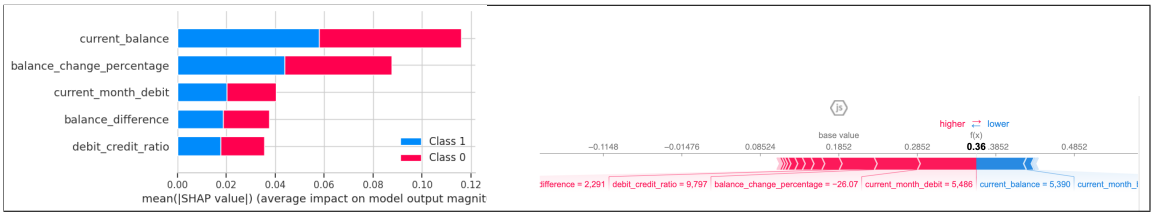

Figure 13: Random Forest classifier prediction's explanation using SHAP in banking churn prediction

5.2 Customer Retention with Catboost

In the output of Catboost classifier illustrated in Figure 14 and 15, the model can cover nearly half of the churn cases, as evidenced by its precision of 72% and recall churn of 46%. The F1 score, which trade-off precision and recall, is 56%. The model's ability to distinguish between the two classes was demonstrated by the overall accuracy of 87% and the reported Area Under the Curve (AUC) of 85%. A total of 729 cases (True Positive) were successfully predicted. Moreover, the model's AUCs for the training and testing sets indicate that there were no indications of overfitting.

	precision	recall	f1-score	support
0	0.89	0.96	0.92	6937
1	0.72	0.46	0.56	1578
accuracy			0.87	8515
macro avg	0.80	0.71	0.74	8515
weighted avg	0.86	0.87	0.85	8515

Figure 14: Model accuracy

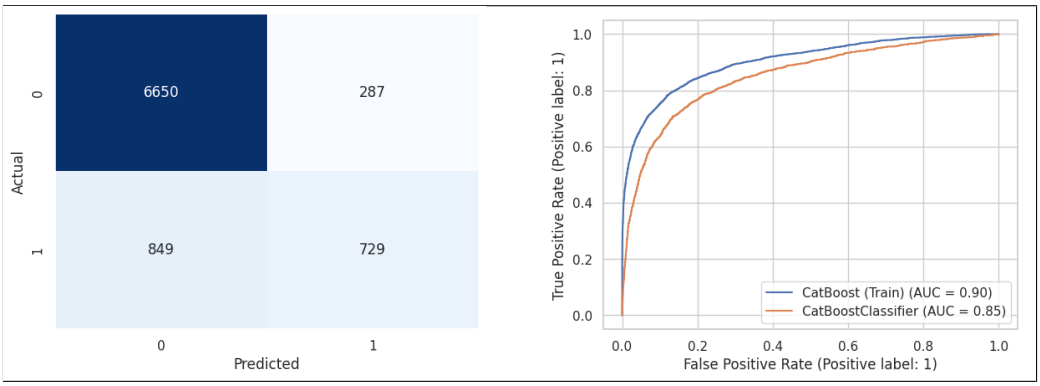

Figure 15: Catboost classifier performance in banking churn prediction

Figure 16 displays top five important feature extracted from the Catboost classifier and one observation explained by SHAP value. Current account balance, the percentage of changing balance, and monthly average balance last quarter have a significant influence on the prediction. In particularly, bank account holders with low current balance (<1,858) or negative changes (<-75%) in recent cycles while high monthly average balance past quarter are more likely to become churners in the future.

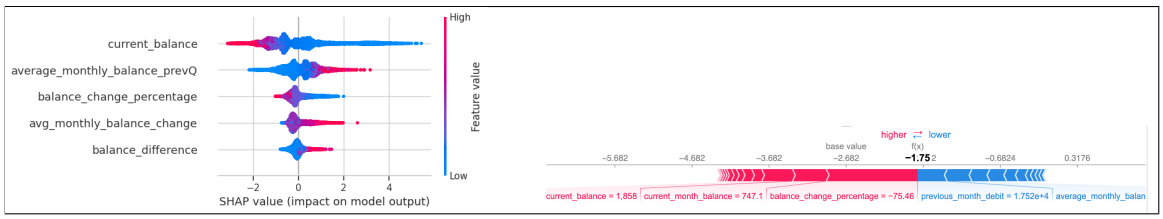

Figure 16: Catboost classifier prediction's explanation using SHAP in banking churn prediction

5.3 Customer Retention with XGBoost

Figure 17 and 18 describe the XGBoost's performance in classifying potential churners. In overall, the XGBoost classifier performs well in forecasting which customers are more likely to stop using the bank service. The precision in the XGBoost model is 72%, meaning that the model is able to predict accurately 72% churners. The recall is 0.47, indicating that 47% of real risk of churn are captured by the model. Furthermore, the precision-recall balance (F1-score) is 57% and the AUC is 85%, which also denotes strong classification power. There were 742 cases that were correctly predicted overall. Additionally, based on the AUC of the training and testing sets shown, the model does not pose any sign of overfitting.

	precision	recall	f1-score	support
0	0.89	0.96	0.92	6937
1	0.72	0.47	0.57	1578
accuracy			0.87	8515
macro avg	0.80	0.71	0.75	8515
weighted avg	0.86	0.87	0.86	8515

Figure 17: Model accuracy

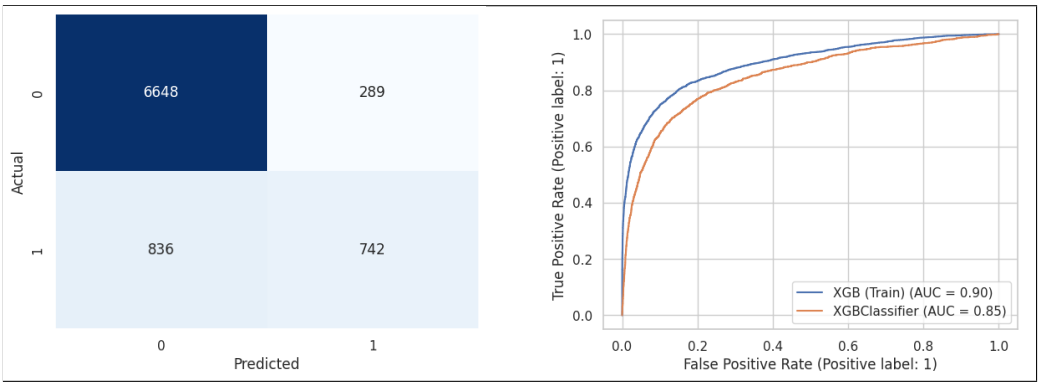

Figure 18: XGBoost classifier performance in banking churn prediction

The top five significant features gleaned from the XGBoost classifier are shown in Figure 19, along with an explanation of one observation based on the SHAP value. The prediction is significantly impacted by the current account balance, average monthly balance change, and monthly average balance from the previous quarter. Particularly, those who have a high monthly average balance from the previous quarter (>3,410) but a low current balance (<123) or significant changes (<-62.47 %) in recent cycles are at a higher risk of becoming churners in the future.

Figure 19: XGBoost classifier prediction's explanation using SHAP in banking churn prediction

5.4 Customer Retention with ANN

The artificial neural network model's (ANN) AUC is 78%, shown by Figure 20. With regard to accuracy in classifying churners, the model scores 68%, meaning that 68% of the time it is right when it forecasts a customer will leave. The model appears to capture 39% of the real churn instances, according to the recall for class 1 of 39%. In addition, the precision-recall balance represented by the F1-score is 49%. In terms of predicting customers who are likely to leave, the ANN model performs moderately overall. 609 instances (True Positive) were correctly predicted overall. The AUC of the training and testing sets (87% and 78%) further indicate that the model pose to overfitting.

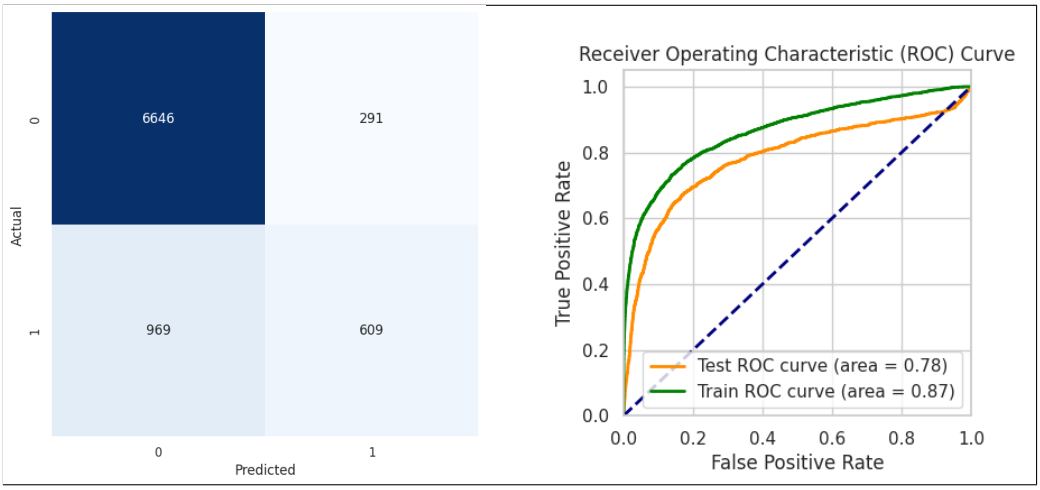

Figure 20: ANN classifier performance in banking churn prediction

The simulation results are shown in Figure 21, with 100 randomly selected samples from the testing dataset representing significant feature values and 500 perterbation samples used to estimate the SHAP value for specific predictions. This is necessary due to the high computational requirements of interpreting a deep learning model by using SHAP in this context.

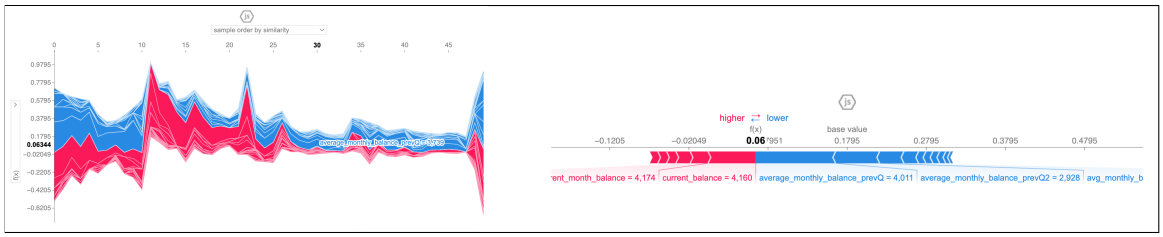

Figure 21: Sample prediction's explanation for ANN using SHAP in banking churn prediction

5.5 Discussion

The use of advanced analytics techniques, particularly machine learning models including CatBoost, RandomForest, XGBoost and ANN, is critical in preventing churn in the banking domain. These models are widely used by their ability to process massive amounts of data, identify patterns, and make predictions, enabling banking business to mitigate customer churn proactively. Their interpretability, more specifically, aided by techniques such as SHAP values, provides important insights into the factors influencing churn decision. Banks can implement timely interventions thanks to real-time decision-making capabilities, and personalized customer loyalty strategies improve the overall customer experience. The cost-effectiveness of these techniques stems from their ability to efficiently allocate resources, focusing on high-value customers for personalized retention efforts. Finally, in the dynamic banking landscape, these advanced analytical approaches also provide a comprehensive solution for proactively addressing churn and strengthening customer loyalty in banking domain.

The implemented models' performance metrics demonstrate their effectiveness in churn prevention in the banking domain as shown in Table 4. XGBoost has the highest Recall (47%), AUC (85%), and F1 score (57%) making it a strong overall contender in customer retention strategies. CatBoost achieves a competitive F1 score (56%), while also excelling in recall (47%). Random Forest and the artificial neural network (ANN) perform well, with Random Forest achieving a good balance of accuracy and recall. Accordingly, the model by XGBoost is proposed as the most suitable one in predicting churner for customer retention strategy with highest AUC, recall and f1-score based on model evaluation. Furthermore, the applied models show no signs of being affected by data imbalance and over fitting. This result supports the research of Sanders et al. (2022) regarding the potential of XGBoost in predicting potential churner.

Table 4: Performance Summary on Testing Set

Classifier	**Accuracy**	**Recall**	**Precision**	**F1 Score**	**AUC**
Catboost	87%	46%	72%	56%	85%
XGBoost	87%	**47%**	**72%**	**57%**	85%
Random Forest	86%	42%	73%	54%	85%
ANN	85%	39%	68%	49%	78%

The SHAP values analysis used in this study stated its useful in determining the complex relationships between various features and the likelihood of customer churn. By investigating the significance of individual features via SHAP values, valuable insights

have be gained by deep diving into simple model predictions. These insights pave the way for the creation of personalized retention efforts such as targeted marketing campaigns, customized services, and loyalty programs. In this regard, specific features identified by SHAP values, such as those from the XGB classifier, show that significant changes in the current month for customers with stable account balances in the previous quarter provide a more nuanced understanding of potential churn risks. Furthermore, office workers have a lower churn rate than other customers, whereas customers with low net worth have a higher churn rate than other groups. This understanding also enables institutions to tailor their retention strategies to each customer's unique characteristics, fostering a more effective and personalized approach to customer loyalty management. These findings from this study's aspect is also consistent with research by (Mecha et al.; 2015) Supriyanto et al. (2021) and Krishna et al. (2022) on the use of valuable insights in personalized retention efforts.

In conclusion, using advanced analytic techniques to integrate traditional banking practices with a deep understanding of transactional data is critical for improving the overall customer experience. Transactional data provides valuable insights into individual customer behaviors, allowing institutions to improve interactions, streamline service delivery, and customize communication strategies. Based on transactional history, this integration enables personalized product recommendations, targeted promotions, and timely communication. Banking institutions can create a more comprehensive and improved customer experience by recommending strategies that seamlessly blend traditional approaches with transactional insights, fostering satisfaction and brand loyalty.

6 Conclusion and Future Work

Overall, the study provided a conceptual framework for customer retention management along with a thorough analysis of the business practices and literature pertaining to customer retention in the context of the Indonesian banking industry. The author's particular goal is to suggest a data mining strategy for client retention. The second goal is to explore how to apply SHAP value for gaining practical insight into customer attrition strategies. It is anticipated that the study's findings will be helpful in creating improved churn management practices for the banking industry as well as other industries.

Generally, the obtained models not only fulfill their role in predicting churn but also serve as pivotal tools for customer loyalty management. Through a detailed evaluation of model performance, it is evident that data-informed strategies derived from these models is potentially impact and improve customer loyalty in banking institutions. The integration of regular practices with transactional data comprehension has emerged as a key aspect, promising a holistic approach to building a better customer experience.

Finally, this study delves into the complex landscape of preventing churn in the banking domain, employing advanced analytical techniques and taking use of valuable insights from SHAP values. The research has given us a thorough understanding of how these techniques contribute to data-informed customer loyalty management. The developed models, which include CatBoost, RandomForest, XGBoost, and ANN, have shown promising results in preventing churn, highlighting their potential applications in the banking industry.

Future Work:

While this study has made significant advances, there is still room for additional stud-

ies that may be beneficial to the ongoing development in churn prevention and customer loyalty management in the banking industry. Limitation in computational resources was not able to take advantage of the power of ANN in predicting churners, as well as in expanding the search area of Bayesian algorithms when performing model optimization. Alternatively, based on direct customer input, consider integrating customer feedback and sentiment analysis for further refine and personalize retention strategies. By addressing these issues, future research can help to improve and innovate customer retention practices in the banking domain, promoting a more resilient and customer-centric financial ecosystem.

References

A., S. K. and D., C. (2016). A survey on customer churn prediction using machine learning techniques, *International Journal of Computer Applications* **154**(10): 13–16.
URL: *http://www.ijcaonline.org/archives/volume154/number10/26526-2016912237*

AL-Najjar, D., Al-Rousan, N. and AL-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction, *Journal of Theoretical and Applied Electronic Commerce Research* **17**(4): 1529–1542.

Al-Shehari, T. and Alsowail, R. A. (2021). An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques, *Entropy* **23**(10).
URL: *https://www.mdpi.com/1099-4300/23/10/1258*

Asadi, S., Nilashi, M., Husin, A. R. C. and Yadegaridehkordi, E. (2017). Customers perspectives on adoption of cloud computing in banking sector, *Information Technology and Management* **18**: 305–330.
URL: *https://doi.org/10.1007/s10799-016-0270-8*

Bhadani, A. K. and Jothimani, D. (2016). Big data: challenges, opportunities, and realities, *Effective big data management and opportunities for implementation* pp. 1–24.
URL: *https://doi.org/10.4018/978-1-5225-0182-4.ch001*

Bilal Zorić, A. (2016). Predicting customer churn in banking industry using neural networks, *Interdisciplinary Description of Complex Systems: INDECS* **14**(2): 116–124.

Bounsaythip, C. and Rinta-Runsala, E. (2001). Overview of data mining for customer behavior modeling, *VTT Information Technology Research Report, Version* **1**: 1–53.

Breiman, L. (2001). Random forests, *Machine Learning* **45**: 5–32.
URL: *http://dx.doi.org/10.1023/A:1010950718922*

Cao, N. (2021). Explainable artificial intelligence for customer churning prediction in banking.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T. et al. (2015). Xgboost: extreme gradient boosting, *R package version 0.4-2* **1**(4): 1–4.

Cho, H. U., Nam, Y., Choi, E. J., Choi, Y. J., Kim, H., Bae, S. and Moon, J. W. (2021). Comparative analysis of the optimized ann, svm, and tree ensemble models using bayesian optimization for predicting gshp cop, *Journal of Building Engineering* **44**: 103411.

De, S. and Prabu, P. (2022). Predicting customer churn: A systematic literature review, *Journal of Discrete Mathematical Sciences and Cryptography* **25**(7): 1965–1985.
URL: *https://doi.org/10.1080/09720529.2022.2133238*

Dorogush, A. V., Ershov, V. and Gulin, A. (2018). Catboost: gradient boosting with categorical features support, *arXiv preprint arXiv:1810.11363* .
URL: *https://doi.org/10.48550/arXiv.1810.11363*

Ekanayake, I., Meddage, D. and Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using shapley additive explanations (shap), *Case Studies in Construction Materials* **16**: e01059.
URL: *https://doi.org/10.1016/j.cscm.2022.e01059*

Elyusufi, Y. and Kbir, M. A. (2022). Churn prediction analysis by combining machine learning algorithms and best features exploration, *International Journal of Advanced Computer Science and Applications* .
URL: *https://dx.doi.org/10.14569/IJACSA.2022.0130773*

G. Ravi Kumar, K. Tirupathaiah, B. K. R. (2019). Client churn prediction of banking and fund industry utilizing machine learning techniques, *International Journal of Computer Sciences and Engineering* **7**: 842–846.
URL: *https://doi.org/10.26438/ijcse/v7i6.842846*

Ganaie, T. A. and Bhat, M. A. (2020). Relationship marketing practices and customer loyalty: A review with reference to banking industry, *International Journal of Engineering and Management Research* **10**.
URL: *https://papers.ssrn.com/sol3/papers.cfm?abstract*$_i d = 3682528$

Geiler, L., Affeldt, S. and Nadif, M. (2022). An effective strategy for churn prediction and customer profiling, *Data Knowledge Engineering* **142**: 102100.
URL: *https://www.sciencedirect.com/science/article/pii/S0169023X2200091X*

He, B., Shi, Y., Wan, Q. and Zhao, X. (2014). Prediction of customer attrition of commercial banks based on svm model, *Procedia Computer Science* **31**: 423–430. 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050914004633*

Jeong, Y., Hwang, M. and Sung, W. (2023). Training data selection based on dataset distillation for rapid deployment in machine-learning workflows, *Multimedia Tools and Applications* **82**(7): 9855–9870.

Jovanovic, L., Kljajic, M., Mizdrakovic, V., Marevic, V., Zivkovic, M. and Bacanin, N. (2023). Predicting credit card churn: Application of xgboost tuned by modified sine cosine algorithm, *2023 3rd International Conference on Smart Data Intelligence*

(ICSMDI), pp. 55–62.
URL: *https://doi.org/10.1109/ICSMDI57622.2023.00018*

Kaggle (2019). Abc bank customer churn dataset, `https://www.kaggle.com/datasets/sainathreddys/banking-churn`.

Karvana, K. G. M., Yazid, S., Syalim, A. and Mursanto, P. (2019). Customer churn analysis and prediction using data mining models in banking industry, *2019 International Workshop on Big Data and Information Security (IWBIS)*, pp. 33–38.
URL: *https://doi.org/10.1109/IWBIS.2019.8935884*

Kaur, I. and Kaur, J. (2020). Customer churn analysis and prediction in banking industry using machine learning, *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp. 434–437.
URL: *http://dx.doi.org/10.1109/PDGC50313.2020.9315761*

Kavyarshitha, Y., Sandhya, V. and Deepika, M. (2022). Churn prediction in banking using ml with ann, *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, pp. 1191–1198.
URL: *https://doi.org/10.1109/ICICCS53718.2022.9788456*

Kelleher, J. D., Namee, B. M. and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, The MIT Press.

Kellner, R., Nagl, M. and Rösch, D. (2022). Opening the black box–quantile neural networks for loss given default prediction, *Journal of Banking & Finance* **134**: 106334.
URL: *https://doi.org/10.1016/j.jbankfin.2021.106334*

Krishna, S. H., Vijayanand, N., Suneetha, A., Basha, S. M., Sekhar, S. C. and Saranya, A. (2022). Artificial intelligence application for effective customer relationship management, *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, pp. 2019–2023.
URL: *https://doi.org/10.1109/IC3I56241.2022.10073038*

Kumar, G. R., Tirupathaiah, K. and Krishna Reddy, B. (2019). Client churn prediction of banking and fund industry utilizing machine learning techniques, *International Journal of Computer Sciences and Engineering* **7**(6): 842–846.
URL: *http://dx.doi.org/10.26438/ijcse/v7i6.842846*

Kumar Hegde, S., Hegde, R., Nanda, S. S., Phatak, G., Hongal, P. and V, D. G. (2023). Customer churn analysis in financial domain using deep intelligence network, *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pp. 362–370.
URL: *https://doi.org/10.1109/IDCIoT56793.2023.10053473*

Mbithi, W. N. (2013). Challenges of customer retention in the kenyan banking sector. a case study of kcb treasury square branch, mombasa, *International Journal of Sciences: Basic and Applied Research* **12**(1): 124–132.
URL: *https://www.gssrr.org/index.php/JournalOfBasicAndApplied/article/view/1278*

Mecha, E. K., Martin, O. and Ondieki, S. M. (2015). Effectiveness of customer retention strategies: A case of commercial banks, kenya, *International journal of business and management* **10**(10): 270.
URL: *http://dx.doi.org/10.5539/ijbm.v10n10p270*

Munkhdalai, L., Munkhdalai, T. and Ryu, K. H. (2020). Gev-nn: A deep neural network architecture for class imbalance problem in binary classification, *Knowledge-Based Systems* **194**: 105534.
URL: *https://www.sciencedirect.com/science/article/pii/S095070512030037X*

Ngai, E. W. and Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda, *Journal of Business Research* **145**: 35–48.
URL: *https://doi.org/10.1016/j.jbusres.2022.02.049*

Patil, K., Patil, S., Danve, R. and Patil, R. (2022). Machine learning and neural network models for customer churn prediction in banking and telecom sectors, *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2021*, Springer, pp. 241–253.

Pfeifer Phillip E (2005). The optimal ratio of acquisition and retention costs, *Journal of Targeting, Measurement and Analysis for Marketing* **13**(2): 179—-188.
URL: *https://doi.org/10.1057/palgrave.jt.5740142*

Sagala, N. T. M. and Permai, S. D. (2021). Enhanced churn prediction model with boosted trees algorithms in the banking sector, *2021 International Conference on Data Science and Its Applications (ICoDSA)*, Vol. 1, pp. 240–245.
URL: *https://doi.org/10.1109/ICoDSA53588.2021.9617503*

Sanders, W., Li, D., Li, W. and Fang, Z. N. (2022). Data-driven flood alert system (fas) using extreme gradient boosting (xgboost) to forecast flood stages, *Water* **14**(5): 747.
URL: *https://doi.org/10.3390/w14050747*

Schröer, C., Kruse, F. and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model, *Procedia Computer Science* **181**: 526–534. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050921002416*

Silveira, L. J., Pinheiro, P. R. and Junior, L. S. d. M. (2021). A novel model structured on predictive churn methods in a banking organization, *Journal of Risk and Financial Management* **14**(10).
URL: *https://www.mdpi.com/1911-8074/14/10/481*

Sugiato, B., Riyadi, S. and Budiarti, E. (2023). The effects of customer relationship management, service quality and relationship marketing on customer retention: The mediation role of bank customer retention in indonesia, *Accounting* **9**(2): 85–94.
URL: *http://doi.org/10.5267/j.ac.2022.12.004*

Sundarkumar, G. G. and Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance, *Engineering Applications of Artificial Intelligence* **37**: 368–377.
URL: *https://www.sciencedirect.com/science/article/pii/S0952197614002395*

Supriyanto, A., Wiyono, B. B. and Burhanuddin, B. (2021). Effects of service quality and customer satisfaction on loyalty of bank customers, *Cogent Business & Management* **8**(1): 1937847.
URL: *https://doi.org/10.1080/23311975.2021.1937847*

Tariq, M. U., Babar, M., Poulin, M. and Khattak, A. S. (2022). Distributed model for customer churn prediction using convolutional neural network, *Journal of Modelling in Management* **17**(3): 853–863.

Van den Poel, D. and Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models, *European Journal of Operational Research* **157**(1): 196–217. Smooth and Nonsmooth Optimization.
URL: *https://www.sciencedirect.com/science/article/pii/S0377221703000699*

Verbeke, W., Martens, D., Mues, C. and Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques, *Expert Systems with Applications* **38**(3): 2354–2364.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417410008067*

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., Gao, M., Hou, H. and Wang, C. (2018). Machine learning and deep learning methods for cybersecurity, *IEEE Access* **6**: 35365–35381.
URL: *https://ieeexplore.ieee.org/document/8359287*

Zhang, J. (2023). Customer churn prediction based on a novelty hybrid random forest algorithm, *in* T. Zhang and T. Yang (eds), *Third International Conference on Computer Vision and Data Mining (ICCVDM 2022)*, Vol. 12511, International Society for Optics and Photonics, SPIE, p. 125112J.
URL: *https://doi.org/10.1117/12.2660705*

Özden Gür Ali and Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking, *Expert Systems with Applications* **41**(17): 7889–7903.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417414003595*