Customer Churn in Bank – Classification By Use Of Machine Learning

Kent State University

Ambassador Crawford College of Business and Entrepreneurship

Course - Capstone Project in Business Analytics

Professor Name : Mostafa Arkdani

Name : Vaishnavi Donde

Email ID: vdonde@kent.edu

KSU ID: 811300129

Group : 11

Abstract

This study explores the factors of customer churn in the banking sector using a comprehensive dataset, with the objective of identifying key factors influencing a customer's decision to leave their bank. This study was conducted to provide insights into the process of analyzing and interpreting data from an openly accessible dataset. It focuses on evaluating the features based on their significance and importance in predicting customer churn among bank clients. The initial stages of this report clearly highlighted the challenges involved in selecting the relevant features. Using these variables, with some embedded and specified holdout and test percentages ranging from 10-20%, this experiment required 3 runs of different adjustments to thoroughly analyze over 24 machine learning methods tailored for classification tasks (A Customer exiting the bank). For future research, examining which combinations of features yield more detailed insights through the 3 runs would be valuable. The analysis includes various features such as CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, and EstimatedSalary, each evaluated for its potential impact on churn rates. Through a series of experimental runs, the study aimed to optimize the predictive model, enhancing its accuracy in classifying customers who are likely to exit the bank. This research underscores the importance of understanding customer behavior and the interplay of various factors in developing effective churn-prevention strategies. By identifying the key drivers of churn, banks can tailor their loyalty programs and retention campaigns to better address customer needs and reduce attrition rates. The findings contribute valuable insights for the banking industry, highlighting actionable areas for improving customer satisfaction and loyalty.

Key Words: Machine Learning, Classification, 3 Runs, Ohio Super Computer (OSC), Deep Learning, IQR, Confusion Matrix, Churn.

1. Introduction

The study to be presented applies multiple classification runs. It employs a set of 24 methods for each type of run. The purpose of this analysis is to evaluate whether the runs are accurate enough to predict customer churn in a banking context. This involves considering various factors such as CreditScore, Geography, Gender, Age, Tenure, Balance, Number of Products, possession of a Credit Card, Active status, and Estimated Salary (Mehmet Akturk et al., 2020). The study also addresses the time required for each of these runs, translating all four types of results (True-Positive, True-Negative, False-Positive, and False-Negative) from the Test Confusion Matrix into comprehensible visuals (Wikipedia, 2024). These visuals depict the cumulative results across all three runs for each type of outcome. This report will provide insights into why the results help us identify customers who are likely to leave the bank (classification), as well as examine various other studies that have explored similar and different approaches to deep machine learning. The beginning will review the different types of classification methods and their significance for identifying customer churn in banks. The next section will cover the data preparation process for training using modern methods and methodologies. It will present the features, methods, and results in an organized manner, making it easy to understand and extract key insights from this research report.

2. Literature Review

Customer churn, the phenomenon where customers discontinue their relationship with a service provider, is a significant issue for banks (Pahul et al.,2024). Predicting and mitigating churn is crucial for maintaining customer base and profitability. Various machine learning techniques have been explored for this purpose, each offering unique insights and advantages. This literature review provides an overview of key methods and studies in this domain.

Machine Learning Techniques for Customer Churn Prediction - Machine learning (ML) has become a foundation in predictive analytics for customer churn (Brain John et al.,2023). The flexibility of ML allows for the application of various techniques tailored to specific datasets and problem contexts.

Binary Classification and Economic Indicators - Binary classification models are commonly used in churn prediction, categorizing customers into churners and non-churners (Mzanatta et al.,2024). Studies have highlighted the importance of incorporating economic indicators into these models. For example, the Mishkin test, traditionally used to detect economic shifts, can enhance model accuracy by integrating macroeconomic factors. By aligning these economic variables with customer behavior, models can provide a more holistic view of churn dynamics.

Advanced Classification Methods - Beyond traditional techniques, advanced methods such as Naive Bayes, Support Vector Machines (SVM), and Neural Networks have been employed. Naive Bayes, despite its simplicity, has proven effective in many churn prediction scenarios due to its probabilistic nature (Selva et al.,2024). On the other hand, SVM and Neural Networks offer powerful tools for capturing nonlinear relationships and high-dimensional data patterns, even with increased computational complexity.

ARDL Models and Stratification - Auto-Regressive Distributed Lag (ARDL) models provide a nuanced approach by examining the impact of past values of independent variables on current churn rates. (Dave et al.,2017) This sequential dimension helps in understanding the long-term effects of various predictors. Similarly, stratification techniques categorize data into meaningful subgroups, improving the roughness of predictions (Muhammet et al.,2023). For instance, stratifying customers by income or education level can yield more precise insights.

Socioeconomic Factors and Churn - Socioeconomic factors play a key role in customer churn. Studies have shown that variables such as income, occupation, and education significantly influence churn rates (Pahul et al.,2024). Visual representations of these factors, through methods like heatmaps or cluster analysis, aid in identifying critical segments at risk of churning.

Impact of the COVID-19 Pandemic - The COVID-19 pandemic has profoundly affected customer behavior and churn. The economic instability and lifestyle changes induced by the pandemic have reshaped customer priorities and financial stability (Rui et al.,2024). Consequently, models developed pre-pandemic may require recalibration to account for these shifts. The pandemic's impact on global economic structures underscores the necessity of adaptive and strong churn prediction models.

Equity and Bias in Churn Prediction - Finally, addressing biases in churn prediction models is crucial (Rapid Canvas, 2024). Research indicates that demographic factors, including race and gender, can introduce biases into predictions, potentially leading to unfair treatment of certain groups. Ensuring equitable model performance across diverse customer segments is essential for ethical and effective churn management.

The landscape of customer churn prediction in banking is rich with diverse methodologies and considerations (Datrics, 2024). From traditional decision trees to sophisticated neural networks, each technique offers unique strengths. Incorporating socioeconomic variables and adapting to dynamic economic conditions, such as those induced by the COVID-19 pandemic, are critical for accurate and fair churn predictions. Future research should continue exploring the interplay between customer demographics, economic factors, and advanced machine learning techniques to enhance predictive capabilities and ensure equitable outcomes.

The research involved conducting three runs of various methods using the Ohio Supercomputer (OSC) to classify whether a client would leave the bank. My specific contribution focused on ensuring accuracy and consistency. By employing multiple functions and methods, I identified and selected the most impactful ones, monitored their performance, and recorded the CPU time. I trained the data and analyzed the results using confusion matrices to determine the effectiveness of each method. The goal was to identify the most reliable methods for predicting customer churn. (Explicitly my contribution)

3. Data Preparation

The dataset was sourced from Kaggle, a website that provides easy access to freely downloadable open datasets. The dataset encompasses several columns, each contributing differently to the prediction of customer churn. Here is a detailed breakdown of the dataset's columns and their potential impact on customer churn:

1) RowNumber: This column corresponds to the record number and serves merely as an identifier without any effect on the model's output. It helps in tracking the rows but does not influence the prediction of whether a customer will leave the bank.
2) CustomerId: Similar to RowNumber, this column contains random values specific to each customer. It acts as a unique identifier for customers but does not contribute to the customer's decision to leave the bank.
3) Surname: This column records the surname of the customers. While it is useful for identifying individual customers, it has no impact on predicting customer churn. The surname of a customer does not influence their likelihood of staying with or leaving the bank.
4) CreditScore: Credit score is a significant variable as it can affect customer churn. Typically, customers with higher credit scores are considered less likely to leave the bank. A higher credit score often reflects better financial stability and satisfaction with the bank's services, thereby reducing the likelihood of churn.

5) Geography: The location of a customer can influence their decision to leave the bank. Geographical factors such as local economic conditions, competition among banks in the area, and cultural preferences can affect customer loyalty and retention.
6) Gender: This column records the gender of the customers. It is interesting to explore whether gender plays a role in customer churn.
7) Age: Age is a relevant factor in predicting customer churn. Generally, older customers are more likely to remain loyal to their bank compared to younger customers. Older customers often have longer relationships with the bank and may find it inconvenient to switch banks.
8) Tenure: This column refers to the number of years a customer has been with the bank. Tenure is an important indicator as longer-term customers are usually more loyal and less likely to leave. They have established trust and familiarity with the bank's services over the years.
9) Balance: The balance in a customer's account is a strong indicator of customer churn. Customers with higher account balances are generally less likely to leave the bank. A higher balance signifies a greater investment in the bank's services and possibly higher satisfaction levels.
10) NumOfProducts: This column records the number of products a customer has purchased through the bank. Customers with multiple products are typically less likely to leave, as they have more integrated services and benefits from the bank.
11) HasCrCard: This column indicates whether a customer has a credit card issued by the bank. Customers with a credit card are usually less likely to leave the bank, as they may rely on the credit card for various transactions and enjoy associated benefits.
12) IsActiveMember: Active members are those who frequently use the bank's services. Active customers are less likely to leave the bank because they are engaged with the bank's offerings and have a higher level of interaction and satisfaction.
13) EstimatedSalary: Similar to the balance, the estimated salary of a customer can influence their likelihood of leaving the bank. Customers with lower salaries might be more inclined to leave, possibly seeking better financial deals or services elsewhere. Conversely, higher-salary customers are more likely to stay with the bank.
14) Exited: This is the target column indicating whether or not a customer has left the bank. 0=No,customer did not exit and 1=Yes, customer did exit.

The dataset highlights an important business insight, it is significantly more expensive to acquire a new customer than keeping an existing one. This understanding supports the value of churn analysis. Banks benefit greatly from understanding the factors that drive customers to leave, as it allows them to design effective retention strategies. Preventing churn not only helps in maintaining the customer base but also enables banks to develop loyalty programs and targeted campaigns aimed at keeping customers satisfied and engaged. These initiatives can range from personalized services, and financial incentives, to improved customer support, all geared towards enhancing customer loyalty and reducing the likelihood of churn.

In summary, the dataset from Kaggle provides a comprehensive view of various factors that can influence customer churn in a banking context. By analyzing these variables, banks can identify at-risk customers and implement measures to improve customer retention. The ultimate goal is to

find the most reliable methods for predicting customer churn, allowing banks to proactively address issues and enhance customer satisfaction.

4. Methodology

In the context of predicting customer churn, various statistical methods were applied to determine the importance and influence of different variables. The variables include Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Credit Card Ownership, Active Membership Status, and Estimated Salary. Each variable was evaluated using four methods: MRMR (Minimum Redundancy Maximum Relevance), Chi-Square Test (Chi2), ANOVA (Analysis of Variance), and Kruskal-Wallis Test (Kruskal). The median ranking of each variable across these methods was also calculated to provide a consolidated importance score.

The analysis of various variables using different statistical methods reveals the key factors influencing customer churn in the banking sector. After applying all the methods, below yellow highlighted are the 7 features that I will be using further – Age, NumOfProducts, IsActiveMember, Geography, Gender, Balance, CreditScore. This comprehensive evaluation helps banks identify at-risk customers and tailor retention strategies to reduce churn and enhance customer loyalty.

Table 1: Feature Selection Table

Variable	MRMR	Chi2	ANOVA	Kruskal	Median
CreditScore	8	7	7	7	**7**
Geography	3	3	6	6	**5**
Gender	4	6	4	5	**5**
Age	5	2	1	1	**2**
Tenure	9	8	8	8	**8**
Balance	6	5	3	4	**5**
NumOfProducts	1	1	5	3	**2**
HasCrCard	7	9	10	10	**10**
IsActiveMember	2	4	2	2	**2**
EstimatedSalary	10	10	9	9	**10**

Table 2: Results of 3 Runs defining Holdout, Test Size, and Feature Set.

Set 1 – 7 Features			
Runs	Run 1	Run 2	Run 3
Holdout	15%	15%	15%
Test	10%	15%	20%

Table 3: ML methods.

Label	Machin Learning Methods
ML1	Fine Tree
ML2	Medium Tree
ML3	Coarse Tree
ML4	Binary GLM Logistic Regression
ML5	Efficient Logistic Regression
ML6	Efficient Linear SVM
ML7	Gaussian Naïve Bayes
ML8	Kernel Naïve Bayes
ML9	Linear SVM
ML10	Quadratic SVM
ML11	Cubic SVM
ML12	Fine Gaussian SVM
ML13	Medium Gaussian SVM
ML14	Coarse Gaussian SVM
ML15	Boosted Trees
ML16	Bagged Trees
ML17	RUSBoosted Trees
ML18	Narrow Neural Network
ML19	Medium Neural Network
ML20	Wide Neural Network
ML21	Bilayered Neural Network
ML22	Trilayered Neural Network
ML23	SVM Kernel
ML24	Logistic Regression Kernel

Table 4: Medians and IQRs for CPU times across 3 runs in ascending order in 3 separate tables for viewing pleasure.

Methods	ML7	ML6	ML5	ML1	ML4	ML9	ML15	ML2
Median	0.96	1.02	1.04	1.49	2.03	2.29	2.48	2.50
IQR	0.17	0.1104	0.25	301.21	0.24	0.37	0.1100	1.20

Methods	ML17	ML3	ML8	ML24	ML14	ML13	ML23	ML16
Median	3.55	4.19	4.21	4.99	5.25	6.00	6.19	6.43
IQR	2.10	301.10	0.82	0.87	0.86	1.22	1.48	1.45

Methods	ML12	ML18	ML22	ML19	ML21	ML20	ML10	ML11
Median	9.52	10.82	11.94	12.09	12.24	19.38	119.23	198.11
IQR	1.22	1.06	0.75	1.64	0.27	2.76	22.40	19.94

5. CPU Time

This table presents CPU time statistics for 24 machine learning methods (ML1-ML24) used in customer churn classification of 3 runs. The data reveals significant variations in computational efficiency among the methods. The fastest methods are ML7(Gaussian Naïve Bayes), ML6(Efficient Linear SVM), and ML5(Efficient Logistic Regression), with median CPU times of 0.96, 1.02, and 1.04 seconds respectively. In contrast, ML10 (Quadratic SVM) and ML11 (Cubic SVM) are the slowest, with median times of 119.23 and 198.11 seconds, reflecting their higher complexity. Consistency in performance is indicated by the Interquartile Range (IQR). ML15 (Boosted Trees) shows the most consistent performance with the lowest IQR of 0.1100 seconds, followed by ML6 (Efficient Linear SVM) and ML7(Gaussian Naïve Bayes). Conversely, ML3(Coarse Tree) and ML1(Fine Tree) exhibit high variability with large IQRs of 301.10 and 301.21 seconds respectively.

The wide range of CPU times, from 0.96 to 198.11 seconds, underscores the importance of algorithm selection in machine learning projects, especially for applications requiring rapid predictions or dealing with large datasets.

This analysis is crucial for identifying methods that offer both good predictive performance and computational efficiency, essential for practical applications in customer churn prediction.

Performance Assessment

The comprehensive evaluation of various machine learning models, based on CPU time, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), offered detailed insights into their efficiency and accuracy for customer churn classification in a bank. During N+P Testing, the results consistently showed approximately 80% positive (P) and 20% negative (N), closely matching the distribution detailed in the data card provided on Kaggle. Initially, the Gaussian Naïve Bayes model was notable for its excellent computational efficiency, with a CPU time of just 0.05718 seconds, followed by the Efficient Linear SVM at 0.06783 seconds and the Efficient Logistic Regression at 0.08547 seconds. Less efficient models included the Quadratic SVM and Cubic SVM, with CPU times of 7.88533 and 13.20733 seconds, respectively. Feature selection methods such as MRMR, Chi2, ANOVA, and Kruskal-Wallis consistently identified Age, NumOfProducts, IsActiveMember, Geography, Gender, Balance, and CreditScore as significant predictors. The RUSBoosted Trees model achieved a TP rate of 65.8%, TN rate of 13.5%, FP rate of 6.9%, and FN rate of 13.85%, with balanced weighted metrics resulting in a total weight of 0.44166. The models were rigorously validated through three experimental runs, demonstrating consistent metrics, and further tested using holdout and test sets to ensure robustness. This comprehensive analysis underscores the importance of considering multiple factors when selecting the optimal model for customer churn classification. While the Gaussian Naïve Bayes model excelled in computational efficiency, the RUSBoosted Trees model provided the best balance of accuracy and efficiency according to the weighted metrics.

Table 5: ML rankings based on confusion matrix elements.

	Sec.					Weight	0.2	0.1	0.3	0.4	
Method	**CPU Time**	**TP**	**TN**	**FP**	**FN**		**TP**	**TN**	**FP**	**FN**	**Total_W**
Fine Tree	0.10	75.60	9.73	10.67	4.10		-0.43	0.77	-0.76	0.46	-0.05
Medium Tree	0.20	75.50	9.30	11.00	4.20		-0.47	0.67	-0.68	0.49	-0.03
Coarse Tree	0.42	78.70	4.93	15.47	0.93		0.66	-0.38	0.39	-0.65	-0.05
Binary GLM Logistic Regression	0.16	76.50	4.30	16.00	3.20		-0.12	-0.54	0.52	0.14	0.14
Efficient Logistic Regression	0.09	79.60	0.00	20.40	0.00		0.98	-1.57	1.58	-0.98	0.12
Efficient Linear SVM	0.07	79.60	0.00	20.40	0.00		0.98	-1.57	1.58	-0.98	0.12
Gaussian Naïve Bayes	0.06	76.90	5.90	14.40	2.73		0.02	-0.15	0.14	-0.02	0.02
Kernel Naïve Bayes	0.30	78.60	5.60	14.70	1.00		0.62	-0.22	0.21	-0.63	-0.09
Linear SVM	0.15	79.60	0.00	20.40	0.00		0.98	-1.57	1.58	-0.98	0.12
Quadratic SVM	7.89	77.53	8.30	12.00	2.07		0.25	0.43	-0.44	-0.25	-0.14
Cubic SVM	13.21	77.40	9.27	11.13	2.30		0.20	0.66	-0.65	-0.17	-0.16
Fine Gaussian SVM	0.63	76.40	7.50	12.80	3.20		-0.15	0.23	-0.25	0.14	-0.02
Medium Gaussian SVM	0.40	77.75	8.50	11.80	1.85		0.32	0.47	-0.49	-0.33	-0.17
Coarse Gaussian SVM	0.37	79.60	0.93	19.47	0.00		0.98	-1.35	1.35	-0.98	0.07
Boosted Trees	0.17	76.85	9.20	11.20	2.75		0.01	0.64	-0.63	-0.01	-0.13
Bagged Trees	0.43	74.87	9.60	10.70	4.73		-0.69	0.74	-0.75	0.68	-0.02
RUSBoosted Trees	0.24	65.80	13.50	6.90	13.85		-3.89	1.68	-1.67	3.88	0.44
Narrow Neural Network	0.58	75.75	10.07	10.33	3.85		-0.38	0.85	-0.84	0.37	-0.10
Medium Neural Network	0.81	75.87	10.20	10.20	3.73		-0.34	0.88	-0.87	0.33	-0.11
Wide Neural Network	1.29	74.13	9.73	10.67	5.50		-0.95	0.77	-0.76	0.95	0.04
Bilayered Neural Network	0.78	76.27	10.30	10.10	3.33		-0.20	0.91	-0.90	0.19	-0.14
Trilayered Neural Network	0.80	76.00	9.85	10.55	3.60		-0.29	0.80	-0.79	0.28	-0.10
SVM Kernel	0.41	79.60	0.00	20.40	0.00		0.98	-1.57	1.58	-0.98	0.12
Logistic Regression Kernel	0.33	79.53	0.00	20.40	0.07		0.95	-1.57	1.58	-0.96	0.12

Average	1.24	76.83	6.53	13.84	2.79
Standard Deviation	2.92	2.84	4.15	4.16	2.85

Conclusion

Running multiple iterations of a dataset using different feature sets within various holdout and test percentages can significantly influence the determination of the optimal testing approach for larger datasets and complex issues. Features were categorized into different sets based on their cumulative significance, determined through various correlation methods. Processing the 10,000 data points through numerous training and testing methods, and analyzing the results using confusion matrices and computing times for each method, was highly successful. Examining these tables provides valuable insights into the consistencies and variations depending on specific parameters. The comprehensive evaluation of various machine learning models, based on CPU time, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), offered detailed insights into their efficiency and accuracy for customer churn classification in a bank. Key factors were identified as significant predictors, offering valuable insights into the drivers of churn. The Gaussian Naïve Bayes model stood out for its excellent computational efficiency, with a CPU time of just 0.05718 seconds, while the RUSBoosted Trees model provided an optimal balance with a CPU time of 0.23678 seconds. The RUSBoosted Trees model was recognized as the best in terms of accuracy and efficiency based on weighted metrics. Leveraged the Ohio Supercomputer Center (OSC) for high-performance computing, ensuring efficient handling of extensive data and complex algorithms needed for model training and evaluation. The model and insights from the analysis can guide data-driven decisions in customer churn classification, enabling banks to proactively address issues and enhance customer satisfaction.

References

1. Mehmet Akturk, (2020). *Churn for Bank Customers - Predict customer churn in a bank* Churn for Bank Customers (kaggle.com)
2. Wikipedia,(2024). *Confusion matrix* Confusion matrix - Wikipedia
3. Pahul Preet Singh , Fahim Islam Anik, (2024). *Investigating customer churn in banking: a machine learning approach and visualization app for data science and management* Investigating customer churn in banking: a machine learning approach and visualization app for data science and management - ScienceDirect
4. Brain John, (2023). *How to Implement Customer Churn Prediction [Machine Learning Guide for Programmers]* How to Implement Customer Churn Prediction [Machine Learning Guide for Programmers] (neptune.ai)
5. Mzanatta,(2024). *Strategic Customer Retention: A Comparative Analysis of Churn Prediction Models* Strategic Customer Retention: A Comparative Analysis of Churn Prediction Models | by mzanatta | Data Reply IT | DataTech | Medium
6. Selva Prabhakaran, (2024). *How Naive Bayes Algorithm Works? (with example and full code)* How Naive Bayes Algorithm Works? (with example and full code) | ML+ (machinelearningplus.com)
7. Dave Giles, (2017). *AutoRegressive Distributed Lag (ARDL) Estimation. Part 1 - Theory* EViews: AutoRegressive Distributed Lag (ARDL) Estimation. Part 1 - Theory
8. Muhammet, (2023). *Slice and Dice Your Dataset: An Introduction to Data Stratification* Slice and Dice Your Dataset: An Introduction to Data Stratification | by Muhammet | Medium

9. Rui Chen, Tong Li & Yong Li (2024*). Analyzing the impact of COVID-19 on consumption behaviors through recession and recovery patterns* Analyzing the impact of COVID-19 on consumption behaviors through recession and recovery patterns | Scientific Reports (nature.com)
10. Rapid Canvas (2024). *Common Pitfalls in Churn Prediction and How to Avoid Them* Common Pitfalls in Churn Prediction and How to Avoid Them (rapidcanvas.ai)
11. Datrics (2024). *Bank Churn Prediction: Using ML to Retain Customers* Bank Churn Prediction: Using ML to Retain Customers (datrics.ai)