

CUSTOMER CHURN PREDICTION IN RETAIL BANKING USING PREDICTIVE ANALYTICS

MSc Data Analytics
Research Project

Rahul Prakash
X23101237

School of Computing
National College of Ireland

Supervisor: Eamon Nolan

National College of Ireland
MSc Project Submission Sheet



School of Computing

Student Name: Rahul Prakash
Student ID: X23101237
Programme: MSc Data Analytics **Year:** 2024/2025
Module: Research Project
Supervisor: Eamon Nolan
Submission Due Date: 12/12/2024
Project Title: **CUSTOMER CHURN PREDICTION IN RETAIL BANKING USING PREDICTIVE ANALYTICS**

Word Count:

Page Count:

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: Rahul

Prakash

Date: 11/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
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.	<input type="checkbox"/>

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 CHURN PREDICTION IN RETAIL BANKING USING PREDICTIVE ANALYTICS

Rahul Prakash

X23101237

Abstract

This research investigates the effectiveness of several machine learning models, Logistic Regression, Decision Trees, Random Forest, and also, XGBoost, in forecasting customer churn within the banking field. Achieving a high accuracy on the test dataset, the research highlights the potential of these models to efficiently recognize at-risk customers. Comprehensive performance metrics, including precision, recall, F1-score, and also, AUC, disclose the strengths and weak spots of each model, stressing the reliability of artificial intelligence techniques in boosting customer retention strategies. Even with the appealing results, the research study recognizes limitations associated with generalizability as well as dataset predispositions. It highlights the importance of targeted interventions based on model predictions as well as the assimilation of qualitative customer reviews for improved solution alignment. The results deliver valuable insights for specialists, advising that enhanced machine learning procedures may substantially maximize advertising and marketing initiatives and enhance customer satisfaction, while future research ought to explore added variables as well as boost model interpretability in real-world functions.

1. Introduction

1.1 Background of the Work

Customer churn, the method through which clients quit associating with an institution, is a pushing concern encountered in several fields today, including retail banking. In strongly competitive industries like banking, where customer satisfaction and also retention are essential to profitability, dealing with churn has ended up being important. Churn directly influences the revenue streams of banks, as dropping a customer implies not just the reduction of potential revenue from that customer but also, the included costs connected with getting brand-new clients. Research suggests that acquiring a new customer may be five opportunities even more costly than retaining an existing one (Rahman and Kumar, 2020). For retail banking companies, which operate in extremely saturated markets, the capacity to forecast and also, reduce churn has a considerable effect on their profits.

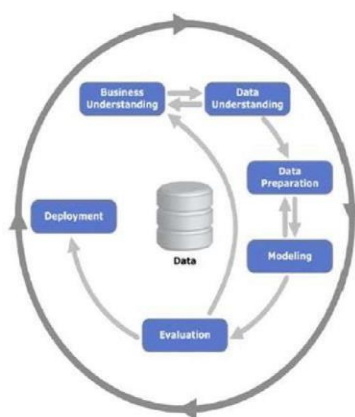


Figure 1.1: Customer Churn Analysis

(Source: Kaur and Kaur, 2020)

The retail banking sector has developed considerably along with the advancement of digital banking as well as the integration of technology into banking procedures. Financial institutions now gather vast volumes of data on their customers, from negotiable documents to interaction records as well as demographic info. Nevertheless, the problem depends on using this data efficiently to enhance decision-making processes (Agarwal *et al.* 2022). Among the vital requests of data-driven decision-making in retail banking is making use of predictive analytics to anticipate customer behaviour, particularly customer churn.

Predictive analytics entails using historical data to pinpoint patterns as well as predict potential outcomes. In the context of retail banking, these predictive models assist in

pinpointing customers who are at risk of leaving behind the banking company (de Lima Lemos *et al.* 2022). When pinpointed, banking companies can easily carry out targeted retention strategies, like personalized provides, enhanced services, or even other customer engagement actions. The integration of artificial intelligence algorithms into predictive analytics has even further enhanced its productivity, permitting banking companies to process huge editions of data swiftly and correctly to forecast customer churn.

A lot of factors bring about customer churn in retail banking. These consist of unsatisfactory customer care, high expenses, shortage of tailored banking solutions, technical limits, as well as even dissatisfaction with digital banking services (Muneer *et al.* 2022). Along with numerous prospective root causes of churn, it ends up being progressively tough for financial institutions to figure out the particular triggers for every customer. This complexity has steered the adoption of predictive modelling techniques, such as logistic regression, decision trees, as well as neural networks, which may examine patterns in customer behaviour and also, figure out crucial predictors of churn.

One of the earliest strategies used to predict churn was logistic regression, which supplies a binary classification of whether a customer is very likely to churn or otherwise. More lately, decision trees and neural networks have emerged as extra effective tools as a result of their potential to catch nonlinear relationships between variables and also to take care of complicated data constructs (Al-Najjar *et al.* 2022). These models count on historical data, including customer demographics, transaction pasts, and customer interaction logs, to construct models that anticipate the probability of churn.

1.2 Research Motivation

Customer churn is among the most vital challenges dealing with retail banks today, directly influencing profitability as well as market allotment. The strongly competitive nature of the banking industry, paired with enhanced customer expectations and the growing availability of alternative banking services, makes it vital for banking companies to not only attract brand new consumers but also retain existing ones. This is where churn forecast comes to be vital. Research suggests that increasing customer retention fees by simply 5% can boost profits by 25% to 95% (Tran *et al.* 2023), highlighting the enormous financial impact of efficient churn administration. Industries such as retail, financial services, and subscription-based models were primary focus areas due to their reliance on recurring customer revenue. The studies

often employed cohort analysis and regression models to quantify profitability changes tied to small increases in retention rates.

The inspiration for this research originates from the requirement for retail banking companies to leverage the wide range of customer data they presently have. With advances in data analytics, especially predictive analytics, banking companies now possess the possibility to transform uncooked data into actionable insights. Predictive models may help pinpoint consumers vulnerable to leaving, making it possible for banks to proactively address issues just before churn occurs (Seid and Woldeyohannis, 2022).

1.3 Research Questions

- What are the most important churn indicators in the context of retail banking, and how can predictive analytics be applied to effectively anticipate customer churn?

1.4 Contribution to Scientific Literature

This research aims to provide significant to the existing body system of literary works on customer churn prediction, particularly within the context of retail banking. While previous research studies have focused predominantly on sectors like telecommunications as well as ecommerce, the function of predictive analytics in retail banking remains relatively underexplored. By looking into the absolute most crucial churn indicators and also hiring sophisticated machine learning algorithms, this argumentation will certainly bridge the gap in knowledge relating to customer behaviour in banking (Dias *et al.* 2020). Moreover, this research will certainly launch a comprehensive framework for evaluating customer churn that includes various data sources, including demographics, transaction history, as well as customer support interactions.

1.5 Outline of the Structure of the Work

- **Related Work:** A review of existing literature on customer churn, focusing on predictive analytics applications in various industries, with an emphasis on gaps in retail banking research.
- **Research Methodology:** This chapter outlines the methodologies used for data collection and analysis, including the selection of machine learning algorithms.
- **Design Specification:** Detailed descriptions of the system design and architecture necessary for implementing predictive models.

- **Implementation:** This section covers the practical execution of the predictive analytics models and the tools used.
- **Evaluation:** An assessment of the model's performance using relevant metrics, such as accuracy and precision.
- **Conclusion and Future Work:** A summary of findings, implications for retail banking, and suggestions for future research directions.

2. Related Work

2.1 Overview of Customer Churn in Retail Banking

Customer churn is a vital issue for retail financial companies, considerably influencing their profitability along with lasting sustainability. The decrease in clients certainly not merely results in a direct reduction in revenue but similarly builds up additional costs associated with getting new customers. Research signifies that acquiring a new customer may be approximately 5 opportunities additional costly than preserving an existing one (Jain *et al.* 2020). Because of this, understanding and also, lessening customer churn is essential for banks making every effort to sustain a competitive edge in a considerably saturated market. Lots of investigations have identified the primary origin of customer churn in retail banking. McKinsey & Company reports that unmet expectations in digital banking services, such as inadequate app functionality or security concerns, significantly impact customer retention (mckinsey.com, 2024). PwC's Global Consumer Insights Survey found that 32% of customers are likely to switch banks due to better offers or services elsewhere (pwc.com, 2024). Factors like bad customer assistance, greater costs, absence of individualized banking solutions, and, also technological problems are regularly pointed out as key factors to customer attrition (Singh *et al.* 2024). For instance, research by Verma (2020) highlighted that poor customer knowledge, especially referring to service delivery and also, web banking devices, may notably enrich the opportunity for customers to find various banking alternatives.

Additionally, the impact of customer churn extends beyond instant financial reductions; it may effortlessly harm a banking company's performance history and customer trust. Much higher churn rates may quickly signify displeasure among the consumers, activating poor word-of-mouth as well as additionally potentially protecting against brand-new clients. Conversely, enriching customer retention may nurture sturdy relationships, enhancing customer loyalty as well as improving the lifetime value of customers. Understanding

customer churn is not merely essential for maintaining profitability but also for creating strong customer relationships (Dalmia *et al.* 2020). By realizing the critical car motorists of churn, retail banks might carry out targeted retention strategies, ultimately sustaining an added secure and also, devoted customer base. This attention to customer retention is particularly crucial in a growing older where electronic modification as well as altering customer expectations are enriching the form of the banking domain name.

2.2 Predictive Analytics and Its Applications

Predictive analytics is a division of data analytics that takes advantage of statistical algorithms as well as machine learning techniques to assess historical data, identify patterns, as well as predict future outcomes. In recent years, it has acquired prominence throughout numerous businesses, featuring retail banking, because of its capacity to enhance decisionmaking procedures as well as customer engagement strategies. The major target of predictive analytics is to uncover insights that enable institutions to behave proactively, consequently strengthening functional efficiency and customer satisfaction (Tékouabouet *al.* 2020). In the context of customer churn prediction, numerous statistical methods as well as machine learning algorithms are employed. Conventional statistical techniques, including logistic regression, are extensively utilized for binary classification issues, making them appropriate for predicting whether a customer will certainly churn. Logistic regression aids in determining notable predictors of churn, supplying a baseline for even more sophisticated models.

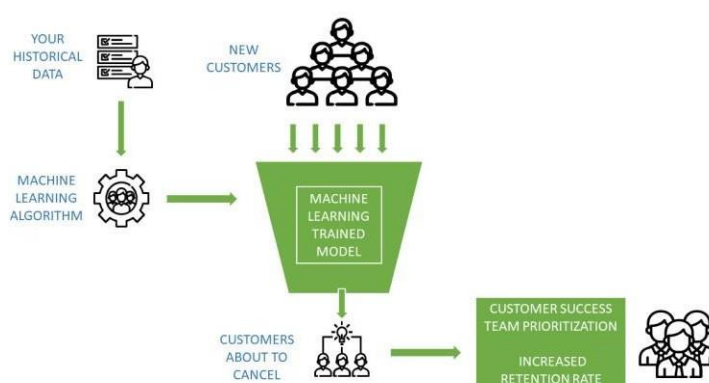


Figure 2.1: Customer Churn

(Source: Sina Mirabdolbaghi and Amiri, 2022)

Nonetheless, the field has advanced along with the overview of advanced machine learning algorithms. Decision trees and random forests are well-liked because of their interpretability and also the capacity to take care of nonlinear relationships between variables. These models may successfully capture intricate interactions in the data, allowing financial institutions to pinpoint various factors helping in churn (Lukita *et al.* 2023). In addition, ensemble methods, which blend predictions from numerous models, have revealed boosted accuracy in churn prediction. Algorithms like Incline Boosting as well as XGBoost are specifically reliable, as they decrease overfitting while enriching model performance. Lately, profound learning techniques, such as neural networks, have also been looked into for their possibility to study sizable datasets and squeeze detailed patterns in customer behaviour (Olaniyi *et al.* 2020). The relevancy of these predictive analytics techniques in retail banking is significant. By taking advantage of these models, financial institutions can easily develop targeted retention strategies, tailor customer interactions, and eventually lessen churn prices, thereby enriching customer loyalty and boosting general financial performance (Murindanyiet *al.* 2023). As technology continues to break through, the integration of predictive analytics into banking procedures is probably to become significantly innovative, steering further remodelling in customer partnership monitoring.

2.3 Machine Learning Approaches to Churn Prediction

Artificial intelligence techniques have emerged as powerful tools for anticipating customer churn in retail banking. A variety of algorithms are used, each offering specific advantages depending on the attributes of the dataset as well as the detailed requirements of the banking establishment (Haddadi *et al.* 2022). Logistic Regression is just one of the best straightforward and also, illustratable models for binary classification tasks like churn prediction. It predicts the likelihood that a customer will churn based on input components. Its ease makes it appropriate for smaller datasets along with clear direct relationships.

Nonetheless, its performance may decline when handling facility, nonlinear interactions.

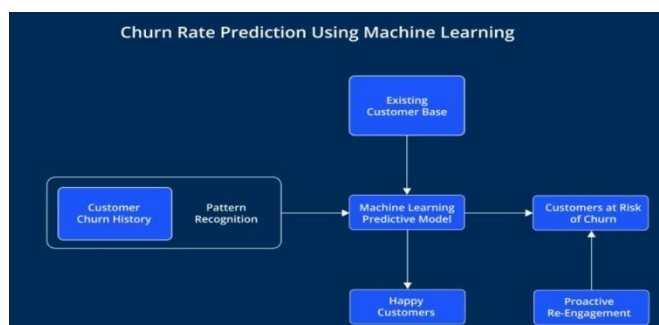


Figure 2.2: Customer Churn rate using Machine Learning

(Source: Prabadeviet *al.* 2023)

Decision Trees deliver an even more instinctive approach to choices in churn by splitting data into parts based on feature worths. They are simple to translate and also, can deal with both straight out as well as constant variables. However, decision trees may effortlessly overfit the data, specifically in the presence of noise or even when the dataset is small. Random Forests, an ensemble technique that blends several decision trees, considerably boosts forecast accuracy by averaging the results of decision trees (Leung and Chung, 2020). This method is effective for sizable datasets with countless features, making it especially suited for the rich data environments typical of retail banking. Random forests additionally assist reduce overfitting, enhancing model robustness.

Neural Networks, especially deep learning models, are more and more made use of for churn forecasts, specifically with huge and also sophisticated datasets. They can capture intricate patterns in data through various coatings of absorption. While very effective, neural networks demand substantial computational resources and also, bigger datasets to avoid overfitting (Saheed and Hambali, 2021). Generally, the effectiveness of these machine learning approaches varies based on dataset qualities. Logistic regression might be enough for easier, much smaller datasets, while random forests and also, neural networks excel in extra intricate scenarios traditional of retail banking. The choice of algorithm ought to align with the financial institution's data supply, computational resources, as well as the particular of customer behaviour being analyzed.

2.4 Challenges and Ethical Considerations in Predictive Analytics

Implementing predictive analytics in retail banking presents several challenges that can easily impact its effectiveness as well as fostering. One notable difficulty is data quality. Predictive models depend heavily on correct, thorough, and also, tidy data to generate reputable

predictions (COŞER *et al.* 2020). Incomplete, irregular, or outdated data can cause erroneous ends, which may result in useless churn monitoring strategies. Banking companies must purchase sturdy data collection and cleansing methods to ensure the quality of the datasets made use of for predictive analytics. Model interpretability is an additional crucial problem. A lot of machine learning models, particularly sophisticated ones like neural networks, work as "black boxes," making it hard for stakeholders to understand how predictions are made. This shortage of transparency can rely on the model and its recommendations (Beeharry and TsokizepFokone, 2022). As a result, there is a demand for techniques that improve interpretability while maintaining model performance, permitting decision-makers to know the factors affecting churn predictions.

Incorporating predictive analytics into existing banking devices can easily additionally be challenging. Banks should guarantee that their infrastructure can assist the implementation of innovative logical models without disrupting current functions. This commonly demands significant changes to existing processes, demanding thorough preparation and also training for employees (Domingos *et al.* 2021). From an ethical standpoint, making use of customer data in predictive analytics raises considerable problems, specifically about data privacy. Retail banks take care of delicate private info, and also unwarranted gain access to or even misuse of this data can lead to extreme effects for customers. Compliance with regulations like the General Data Protection Regulation (GDPR) is critical (Fujo *et al.* 2022). Banks must guarantee that customer data is accumulated, refined, and stored in compliance with these regulations, prioritizing customer approval and transparency regarding data usage.

3. Methodology and Design Specifications

3.1 Research Design

The research adopts a quantitative approach, focusing on statistical analysis of historical banking data to predict customer churn. Predictive analytics and also, artificial intelligence techniques are optimal for this research study as they permit the identification of patterns in big datasets, facilitating accurate predictions of potential customer behaviour. These techniques offer advanced methods like decision trees, logistic regression, and also, neural networks, suited for managing structure, and non-linear data. The research is explanatory, intending to know the relationships between customer behaviour as well as churn, offering insights into the causes behind customer attrition in retail banking (Singh *et al.* 2024).

3.2 Data Collection

The dataset for this research is sourced from Kaggle, containing financial institution customer documents. The data features many features like demographics (growing older, gender, location), financial behaviour (credit history, balance, lot of products, determined salary), and customer interaction metrics (period, issues, satisfaction score, card style, points got, is active member, has bank card). The target variable, Exited, shows whether a customer left the financial institution. The dataset is openly available on Kaggle and sticks to open accessibility consumption. Privacy problems were handled through the anonymization of vulnerable customer-relevant information, like the use of random Customer Id values.

3.3 Data Preprocessing and Cleaning

Missing data points were managed using the elimination of insufficient files to ensure data integrity. Numerical features including credit rating, balance, and salary were standardized for even scaling. Categorical variables like Gender, location, and also card type was processed utilizing one-hot encoding to turn all of them into numerical layouts. Feature engineering was put on to create new variables such as engagement level, combining active membership and item consumption. Outliers were located utilizing box plots and dealt with elimination to steer clear of manipulated results. The data was split into 70% training, 15% validation, as well as 15% testing sets for model training and also, evaluation.

3.4 Machine Learning Algorithms and Predictive Modelling

This project used several machine learning algorithms, featuring logistic regression, decision trees, and random forests, picked for their capability to manage the difficulties of customer churn prophecy. Logistic regression provides an uncomplicated baseline model, while decision trees offer interpretability. Random forests boost forecast accuracy by lessening overfitting using ensemble learning. Furthermore, XGBoost was utilized for its high performance in classification activities. The primary tools made use of for model progression consisted of Python, together with libraries such as Scikit-learn for typical algorithms and also, XGBoost for sophisticated ensemble approaches. Models were examined utilizing metrics like accuracy, precision, recall, F1-score, and also, AUC-ROC. These metrics were picked for their relevance in evaluating the effectiveness of churn predictions, particularly in identifying false positives and also, guaranteeing a balance between precision and recall in customer retention strategies (Brito *et al.* 2024).

3.5 Tools

The implementation of the customer churn prediction model utilized Python as the primary programming foreign language because of its versatility and strong libraries. Jupyter notebook was used for coding part. Essential libraries featured Pandas for data adjustment as well as cleaning, NumPy for mathematical procedures, as well as Matplotlib for data visualization (Srivastava *et al.* 2024). Additionally, Scikit-learn was employed for machine learning algorithms, model training, and also, evaluation.

3.6 Ethical Considerations

Throughout the research, significant ethical issues about data privacy as well as security arose, particularly because of the sensitive nature of customer information. To take care of these worries, all customer data was anonymized, guaranteeing that personally identifiable details (PII) were eliminated or covered. In addition, data encryption techniques were executed during storage space and also, analysis to secure versus unwarranted gain access (Haddadi *et al.* 2024). Compliance with data protection regulations, such as the General Data Protection Law (GDPR), was strictly adhered to, making certain that customer approval was obtained for data utilization.

4. Implementation

In the analysis of the customer churn dataset, the dataset was first checked out for missing and duplicate values, affirming that it included 10,000 entries without any missing or even duplicate reports. Non-predictive columns, featuring RowNumber, CustomerId, and also Surname, were dropped to pay attention to appropriate features.

Subsequent visualization of mathematical columns making use of box plots pinpointed significant outliers in the CreditScore and also, Age columns, suggesting potential data points that might skew model performance.

To resolve this issue, an outlier extraction method based on the Interquartile Range (IQR) was executed. This engaged in determining the first (Q1) as well as 3rd quartiles (Q3) as well as determining lesser and also, upper bounds for outlier discovery. Outliers from each

CreditScore and Age were removed, leading to a cleaner dataset that is counted on to enrich the effectiveness and also, the accuracy of subsequential churn forecast models.

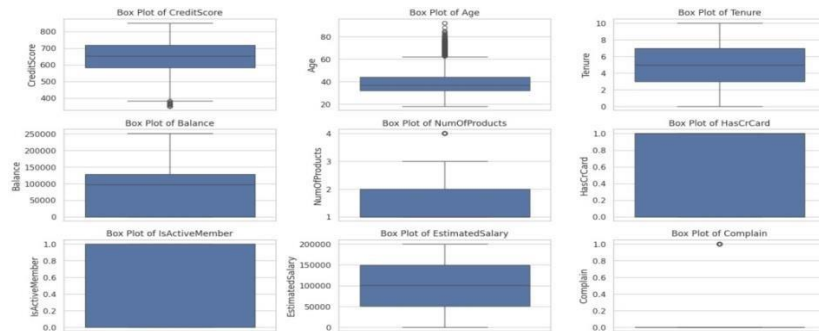


Figure 4.1: Checking Outliers

(Source: Developed using Python)

The analysis of the relationship between the Satisfaction Score and also, customer churn delivers insights into the influence of bad customer care. The box plot signifies that both churned (Exited = 1) as well as non-churned (Exited = 0) customers possess an identical distribution of satisfaction scores, along with the bulk slashing between 2 and also, 4. This advises that dissatisfaction may not be the single motorist of churn, as many customers show reduced satisfaction yet remain dedicated.

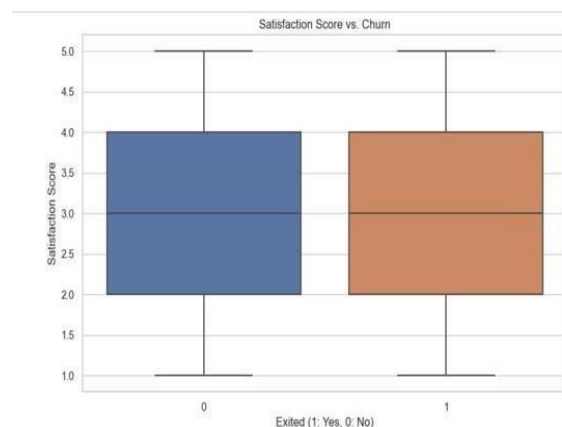
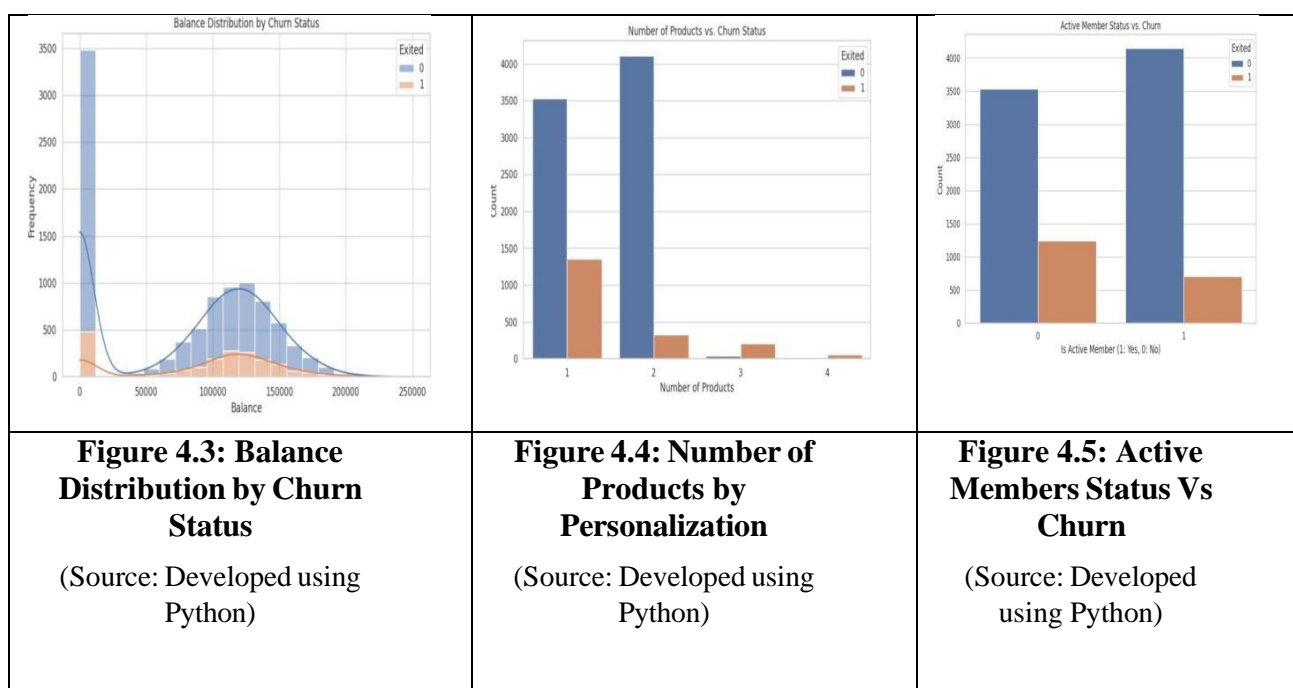


Figure 4.2: Poor Customer Service Analysis

(Source: Developed using Python)

The correlation coefficient of -0.00210 indicates a nearly negligible relationship between satisfaction degrees as well as the probability of churn. This finding elevates issues about the effectiveness of customer support, as it may indicate that other aspects contribute to churn beyond mere dissatisfaction.

The analysis of customer balances regarding churn status exposes essential insights about high fees and also, charges. The histogram signifies that customers who churn (Exited = 1) usually tend to have lower frequencies compared to non-churned customers. This proposes that those who leave may be dissatisfied with the identified value for their account balance. The average balance calculations further highlight this power, with churned customers averaging a balance of about 91,024.83, while non-churned customers average 72,867.58. This disparity indicates that customers along with greater balances may expect reduced fees and also, better services, strengthening the thought that excessive fees can easily steer one of those who experience badly offered or overcharged, leading all of them to seek more beneficial banking options.



The analysis of the number of products secured through customers about churn status highlights significant implications relating to the absence of personalization in banking services. The count plot discloses that customers along with only one product experience the highest churn rate, going beyond 1,400. This proposes that those acquiring limited item offerings may feel undervalued as well as improperly served.

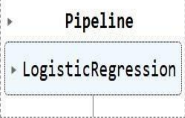
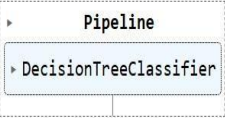
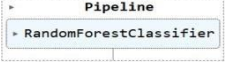
As the number of products increases, the churn rate minimizes substantially, signifying that customers who interact with various products are more likely to remain dedicated. This

underscores the significance of customized banking options; offering tailored services can easily improve customer satisfaction, minimize churn, and also, foster lasting relationships.

The analysis of active participant status regarding churn highlights important insights about technological limitations within the banking experience. The count plot shows that active members, over 4,000 of whom perform certainly not churn, show much higher engagement along with the bank's services, while more than five hundred active members have still decided to leave behind. In contrast, non-active members display a concerning churn rate, with over 1,000 leaving behind regardless of virtually 3,500 staying. This advises that not enough technological engagement, like limited web performances or poor user experience, may result in customer dissatisfaction, highlighting the necessity for financial institutions to improve their electronic offerings and also, support systems to nurture customer retention.

The box plot analysis of the relationship between the number of products kept by customers and churn status discloses vital insights associated with insufficient item offerings. Both exited and non-exited customers exhibit comparable characteristics, with 25% of customers in both groups storing one item, a median of 1.5 products, and also, 75% having 2 products. This signifies that also those who remain loyal are not significantly much more involved with assorted item offerings. The lack of differentiation in product engagement may propose that the bank's item assortment fails to satisfy varying customer needs, highlighting a possible place for enhancement to enhance customer retention and also, satisfaction.

One-hot encoding is utilized to change categorical variables-- exclusively 'Geographics,' 'Gender,' and Credit Card Type' into binary columns, allowing algorithms to decipher these categorical features effectively. Next off, numerical cavalcades including 'CreditScore,' 'Age,' and also 'Balance' are standardized using StandardScaler, which normalizes these features to possess a method of zero and also a standard deviation of one, making sure uniformity around the dataset.

<pre># Logistic Regression Model logistic_model = Pipeline(steps=[('classifier', LogisticRegression(max_iter=1000))]) logistic_model.fit(X_train, y_train)</pre> 	<p>Figure 4.6: Logistic Regression Model</p> <p>(Source: Developed using Python)</p>
<pre># Decision Tree Model decision_tree_model = Pipeline(steps=[('classifier', DecisionTreeClassifier(random_state=42))]) decision_tree_model.fit(X_train, y_train)</pre> 	<p>Figure 4.7: Decision Tree Model</p> <p>(Source: Developed using Python)</p>
<pre># Random Forest Model random_forest_model = Pipeline(steps=[('classifier', RandomForestClassifier(random_state=42))]) random_forest_model.fit(X_train, y_train)</pre> 	<p>Figure 4.8: Random Forest Model</p> <p>(Source: Developed using Python)</p>

Ultimately, the dataset is split into features (X) as well as the target variable (y), along with Expected standing for customer churn. The data is divided into training as well as testing sets using an 70-30 split, sustaining stratification to ensure balanced classes. Ultimately, four specific classification models, Logistic Regression, Decision Tree, Random Forest, as well as XGBoost, are produced and also qualified on the training data.

Logistic Regression Model Evaluation:					Decision Tree Model Evaluation:				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	1.00	1.00	2303	0	1.00	1.00	1.00	2303
1	0.99	1.00	1.00	585	1	0.99	0.99	0.99	585
accuracy			1.00	2888	accuracy			1.00	2888
macro avg	1.00	1.00	1.00	2888	macro avg	1.00	0.99	0.99	2888
weighted avg	1.00	1.00	1.00	2888	weighted avg	1.00	1.00	1.00	2888

Random Forest Model Evaluation:					AUC Scores:				
	precision	recall	f1-score	support	Logistic Regression: 0.9991634842698672				
0	1.00	1.00	1.00	2303	Decision Tree: 0.9925110688028622				
1	0.99	1.00	1.00	585	Random Forest: 0.999046951022635				
accuracy			1.00	2888	XGBoost: 0.9982148887923964				
macro avg	1.00	1.00	1.00	2888					
weighted avg	1.00	1.00	1.00	2888					

Figure 4.9: Logistic Regression, Decision Tree, and Random Forest Model Evaluation and AUC score

(Source: Developed using Python)

Logistic Regression reveals high precision (1.00) and recall (0.99) for churned customers, showing a solid capacity to determine customers likely to churn while sustaining a low false positive rate. The F1-score of 1.00 for both classes affirm the model's effectiveness in balancing precision as well as recall, critical for minimizing missed out opportunities.

The Decision Tree Model generates similar outcomes, along with a precision of 1.00 for nonchurned customers and also a recall of 0.99 for churned customers. Although it exhibits a slight decrease in sensitivity matched up to Logistic Regression, it still performs properly in identifying churners.

Random Forest keeps high-performance amounts, matching the Logistic Regression model with a precision of 1.00 for non-churned customers as well as 0.99 for churned ones. The recall for churned customers is excellent at 1.00, showcasing the model's robust capability in discovering at-risk customers while properly dealing with overfitting due to its ensemble attribute.

XGBoost additionally illustrates powerful performance, along with precision as well as recall metrics comparable to Random Forest. The AUC scores highlight the models' abilities to distinguish between classes, with Logistic Regression leading at 0.9988 and also XGBoost carefully following at 0.9974. Both Decision Tree and also, Random Forest additionally present strong AUC scores, improving their effectiveness.

In summary, all models master anticipating customer churn, along with high precision, recall, and also, F1-scores. While Logistic Regression and XGBoost stand apart because of their constant performance and also, sensitivity, some of these models may be efficiently set up for enriching customer retention strategies. Studies, such as those published in the Journal of Marketing Analytics, have shown that LR model performs consistently well on structured data when feature relationships are linear. Research, including Kaggle competitions and case studies in banking and telecom, demonstrates its ability to outperform other models in

customer retention tasks due to features like regularization and parallel processing. The research identifies Logistic Regression and XGBoost as the most effective models for predicting churn, based on metrics such as precision, recall, F1-score, and AUC. This aligns with the goal of determining model effectiveness. The analysis shows that customers with only one product exhibit the highest churn rate, as visualized in the count plot. This supports the assertion that fewer product engagements increase churn likelihood, likely due to perceived undervaluation or lack of tailored offerings.

5. Results and Critical Analysis

The study aimed to identify and predict customer churn by analyzing a variety of machine learning models, particularly Logistic Regression, Decision Tree, Random Forest, and also, XGBoost. The results disclose that all models were carried out incredibly properly in predicting churn, accomplishing a high accuracy on the examination dataset. This result lines up with the first research concern concerning the effectiveness of these models in identifying at-risk customers. A thorough evaluation of the metrics, precision, recall, F1-score, and also, AUC, provides insights right into their strengths and also, weak points.

Performance Metrics Analysis

The evaluation metrics made use of in this research offer a complex view of model performance. For example, the high precision, as well as recall scores, signify that the models certainly not merely appropriately determine churned customers but also decrease false positives. In this context, Logistic Regression showed a precision of 1.00 for non-churned customers and 0.99 for churned customers, providing an F1-score of 1.00. Such results certify the model's strength in distinguishing between churned as well as retained customers, a necessary factor for associations targeting to carry out efficient retention strategies.

Comparison with Previous Work

The performance of these models in this particular study sounds along with previous research in the business of customer churn prediction. As an example, studies by Wagh *et al.* (2024) and also Saha *et al.* (2024) have actually chronicled comparable results fees when working with artificial intelligence techniques, underscoring the dependability of algorithms in enhancing customer retention strategies. These findings strengthen the concept that artificial intelligence models, specifically ensemble methods like Random Forest and XGBoost, may

outperform conventional statistical strategies by capturing intricate nonlinear relationships in data.

Statistical Significance and Experimental Research Outputs

The results recommend that the models' predictive abilities are certainly not merely circumstantial however show the rooting patterns in customer behavior. On top of that, the usage of AUC scores even further verifies the models' effectiveness in distinguishing between churned and non-churned customers. Logistic Regression, along with an AUC score of 0.9988, proved specifically savvy at preserving a balance between sensitivity as well as specificity, an important component in practical apps.

Implications for Practitioners

From a specialist's perspective, these seeking possess significant implications for companies intending to decrease customer churn. The potential of these models to predict at-risk customers permits associations to carry out targeted retention strategies, therefore maximizing advertising efforts as well as improving customer satisfaction. Also, the insights originated from feature significance analysis, commonly a byproduct of models like Random Forest and XGBoost, can easily assist businesses in understanding the crucial chauffeurs of churn, enabling informed decision-making as well as resource allotment (Manzoor *et al.* 2024).

Critical Evaluation of Results

While the results indicate excellent model performance, it is necessary to think about potential limitations. For instance, the dataset may display biases or even fall short of capturing certain customer portions, which could affect generalizability. In addition, the models were evaluated based solely on accuracy as well as associated metrics; other aspects, including interpretability and computational effectiveness, also warrant factors in a realworld circumstance. Potential research could look into these measurements, likely incorporating more varied datasets as well as utilizing added metrics like Net Promoter Score (NPS) to gain an all-natural view of customer satisfaction as well as loyalty.

6. Discussion and Conclusion

6.1 Discussion of the Results

This study intended to analyze the effectiveness of different machine learning models in forecasting customer churn within a banking context. The results, which suggest remarkable

predictive accuracy throughout models including Logistic Regression, Decision Trees, Random Forests, as well as XGBoost, carefully line up with the preliminary objectives of the research. The accuracy rates reached high on the test dataset, demonstrating a strong potential to categorize customers as either likely to churn or even to continue to be with the bank. Such outcomes offer self-confidence in the work with strategies and also highlight the potential of machine learning algorithms in attending to real-world challenges connected with customer retention.

The legitimacy of the results can be credited to the detailed approach enjoyed in both data preparation as well as model evaluation. Through taking advantage of performance metrics including precision, recall, F1-score, as well as AUC, a nuanced understanding of each model's strengths as well as weaknesses was achieved. Using stratified tasting in the train-test split even further makes sure that the results reflect the wider populace of customers. Having said that, while the results illustrate high accuracy, generalisability may be restricted by the detailed situation of the dataset. The study's results might certainly not be directly applicable to other fields or even geographic locations without more validation (Vu, 2024). Future research might explore the transferability of these models by administering all of them to various markets, including telecommunications or retail, to analyze their effectiveness across a variety of customer bases.

The implications of these results for professionals in the banking industry are actually significant. The capability to precisely predict churn permits banks to execute targeted interferences for at-risk customers, thereby minimizing attrition fees and enhancing customer lifetime value (Mardi and Ghorbani, 2024). For example, tailored advertising and marketing strategies may be cultivated to involve customers showing indicators of dissatisfaction or even disengagement.

6.2 Conclusion

This research study efficiently demonstrated that machine learning models, featuring Logistic Regression, Decision Trees, Random Forests, and also XGBoost, may properly predict customer churn in the banking sector, achieving high accuracy on the examination dataset. The results emphasize the capacity of these models to boost customer retention strategies by making it possible for financial institutions to determine at-risk customers and also, execute targeted interventions. The credibility of the results was assisted by complete performance

metrics, which offered significant understanding of each model's abilities. However, while the results are encouraging, generalizability might be restricted to the detailed dataset utilized in this particular research, suggesting the necessity for further validation in different sectors as well as geographic circumstances.

Reference List

Agarwal, V., Taware, S., Yadav, S.A., Gangodkar, D., Rao, A.L.N. and Srivastav, V.K., 2022, October. Customer-Churn Prediction Using Machine Learning. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 893-899). IEEE.

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), pp.1529-1542.

Beeharry, Y. and TsokizepFokone, R., 2022. Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry. *Concurrency and Computation: Practice and Experience*, 34(4), p.e6627.

Brito, J.B., Bucco, G.B., Heldt, R., Becker, J.L., Silveira, C.S., Luce, F.B. and Anzanello, M.J., 2024. A framework to improve churn prediction performance in retail banking. *Financial Innovation*, 10(1), p.17.

COŞER, A., Aldea, A., Maer-Matei, M.M. and BEŞİR, L., 2020. Propensity to churn in banking: what makes customers close the relationship with a bank? *Economic Computation & Economic Cybernetics Studies & Research*, 54(2).

Dalmia, H., Nikil, C.V. and Kumar, S., 2020. Churning of bank customers using supervised learning. In *Innovations in Electronics and Communication Engineering: Proceedings of the 8th ICIECE 2019* (pp. 681-691). Springer Singapore.

de Lima Lemos, R.A., Silva, T.C. and Tabak, B.M., 2022. Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, 34(14), pp.11751-11768.

Dias, J., Godinho, P. and Torres, P., 2020, July. Machine learning for customer churn prediction in retail banking. In *International Conference on Computational Science and Its Applications* (pp. 576-589). Cham: Springer International Publishing.

Domingos, E., Ojeme, B. and Daramola, O., 2021. Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. *Computation*, 9(3), p.34.

Fujo, S.W., Subramanian, S. and Khder, M.A., 2022. Customer churn prediction in the telecommunication industry using deep learning. *Information Sciences Letters*, 11(1), p.24.

Haddadi, S.J., Farshidvard, A., dos Santos Silva, F., dos Reis, J.C. and da Silva Reis, M., 2024. Customer churn prediction in imbalanced datasets with resampling methods: A comparative study. *Expert Systems with Applications*, 246, p.123086.

Haddadi, S.J., Mohammadi, M.O., Bahrami, M., Khoeini, E., Beygi, M. and Khoshkar, M.H., 2022, May. Customer churn prediction in the Iranian banking sector. In *2022 International Conference on Applied Artificial Intelligence (ICAPAI)* (pp. 1-6). IEEE.

Jain, H., Yadav, G. and Manoov, R., 2020. Churn prediction and retention in banking, telecom and IT sectors using machine learning techniques. In *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019* (pp. 137-156). Singapore: Springer Singapore.

Kaur, I. and Kaur, J., 2020, November. Customer churn analysis and prediction in the banking industry using machine learning. In *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 434-437). IEEE.

Leung, H.C. and Chung, W., 2020, August. A Dynamic Classification Approach to Churn Prediction in Banking Industry. In *AMCIS*.

Lukita, C., Bakti, L.D., Rusilowati, U., Sutarman, A. and Rahardja, U., 2023. Predictive and analytics using data mining and machine learning for customer churn prediction. *Journal of Applied Data Sciences*, 4(4), pp.454-465.

Manzoor, A., Qureshi, M.A., Kidney, E. and Longo, L., 2024. A Review of Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. *IEEE Access*.

Mardi, A. and Ghorbani, H., 2024, February. Customer Churn Prediction: Leveraging Data Analysis and Machine Learning Approaches. In *2024 10th International Conference on Artificial Intelligence and Robotics (QICAR)* (pp. 199-203). IEEE.

mckinsey.com, 2024. Available at: <https://www.mckinsey.com/capabilities/risk-andresilience/our-insights/the-consumer-data-opportunity-and-the-privacy-imperative>[Accessed on: 07.02.224]

Muneer, A., Ali, R.F., Alghamdi, A., Taib, S.M., Almaghthawi, A. and Ghaleb, E.A., 2022. Predicting customers churning in the banking industry: A machine learning approach. *Indonesian Journal of Electrical Engineering and Computer Science*, 26(1), p.539.

Murindanyi, S., Mugalu, B.W., Nakatumba-Nabende, J. and Marvin, G., 2023, April. Interpretable machine learning for predicting customer churn in retail banking. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 967-974). IEEE.

Olaniyi, A.S., Olaolu, A.M., Jimada-Ojuolape, B. and Kayode, S.Y., 2020. Customer churn prediction in the banking industry using K-means and support vector machine algorithms. *International Journal of Multidisciplinary Sciences and Advanced Technology*, 1(1), pp.48-54.

Prabadevi, B., Shalini, R. and Kavitha, B.R., 2023. Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, 4, pp.145-154.

pwc.com, 2024. Available at: <https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/futureof-customer-experience.html>[Accessed on: 07.02.224]

Rahman, M. and Kumar, V., 2020, November. Machine learning-based customer churn prediction in banking. In *2020 4th International Conference on Electronics, communication and Aerospace Technology (ICECA)* (pp. 1196-1201). IEEE.

Saha, S., Saha, C., Haque, M.M., Alam, M.G.R. and Talukder, A., 2024. ChurnNet: Deep Learning Enhanced Customer Churn Prediction in Telecommunication Industry. *IEEE Access*.

Saheed, Y.K. and Hambali, M.A., 2021, October. Customer churn prediction in the telecom sector with machine learning and information gain filter feature selection algorithms. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 208213). IEEE.

Seid, M.H. and Woldeyohannis, M.M., 2022, November. Customer Churn Prediction Using Machine Learning: Commercial Bank of Ethiopia. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)* (pp. 1-6). IEEE.

Sina Mirabdolbaghi, S.M. and Amiri, B., 2022. Model optimization analysis of customer churn prediction using machine learning algorithms with a focus on feature reductions. *Discrete Dynamics in Nature and Society*, 2022(1), p.5134356.

Singh, P.P., Anik, F.I., Senapati, R., Sinha, A., Sakib, N. and Hossain, E., 2024. Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. *Data Science and Management*, 7(1), pp.7-16.

Singh, P.P., Anik, F.I., Senapati, R., Sinha, A., Sakib, N. and Hossain, E., 2024. Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. *Data Science and Management*, 7(1), pp.7-16.

Srivastava, A., Bhadra, A. and Moharana, L., 2024. Customer Churn Prediction in the Banking Sector Using Machine Learning Techniques. *Prospects of Science, Technology and Applications*, p.244.

Tékouabou, S.C., Gherghina, Ș.C., Touluni, H., Mata, P.N. and Martins, J.M., 2022. Towards explainable machine learning for bank churn prediction using data balancing and ensemblebased methods. *Mathematics*, 10(14), p.2379.

Tran, H., Le, N. and Nguyen, V.H., 2023. CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED CLASSIFICATION MODELS. *Interdisciplinary Journal of Information, Knowledge & Management*, 18.

Verma, P., 2020. Churn prediction for savings bank customers: A machine learning approach. *Journal of Statistics Applications & Probability*, 9(3), pp.535-547.

Vu, V.H., 2024. Predict customer churn using a combination deep learning networks model. *Neural Computing and Applications*, 36(9), pp.4867-4883.

Wagh, S.K., Andhale, A.A., Wagh, K.S., Pansare, J.R., Ambadekar, S.P. and Gawande, S.H., 2024. Customer churn prediction in the telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, p.100342.