UNIVERSITY OF BIRMINGHAM | BIRMINGHAM BUSINESS SCHOOL

Assessment and Feedback: Student Template

Student ID Number: 2817392

Programme: MSc Business Analytics

Module: 38157 LM Data Analytics and Predictive Modelling

Name of Tutor: Dr Hannan Amoozad Mahdiraji

Assignment Title: Individual Report

Proposal Title: Churn Rate Evaluation to Power-up Bank Loyalty – Data Analysis Framework

Date and Time of Submission: 16/01/2025 11:50AM

Actual Word Count: 3000

Extension: N **Extension Due Date:** -

I do not wish my assignment to be considered for including as an exemplar in the **School Bank of Assessed Work.**

The purpose of this template is to ensure you receive targeted feedback that will support your learning. It is a requirement to complete to complete all 3 sections, and to include this completed template as the first page of every assignment that is submitted for marking (your School will advise on exceptions).
Section One: Reflecting on the feedback that I have received on previous assessments, the following issues/topics have been identified as areas for improvement: (add 3 bullet points). *NB – for first year students/PGTs in the first term, this refers to assessments in your previous institution* • Sub Headings • Conclusion • References
Section Two: In this assignment, I have attempted to act on previous feedback in the following ways (3 bullet points) • Analytical Answer • Precise Writing • Proper Referencing

Section Three: Feedback on the following aspects of this assignment (i.e. content/style/approach) would be particularly helpful to me: (3 bullet points)

- Data Analysing Skills
- Reading Research papers
- Sentence Formation

Please ensure that you complete and attach this template to the front of all work that is submitted.

By submitting your work online you are confirming that your work is your own and that you understand and have read the University's rules regarding authorship and plagiarism and the consequences that will arise should you submit work not complying with University's Code of Practice on Academic Integrity.

I confirm that I have / have not used a proof-reader(s) (delete as appropriate). If I have used a proof-reader(s) I confirm that I understand the guidance on use of a proof-reader, as specified in the Code of Practice and School guidance.

I hereby declare that I have used Gen AI occasionally to refine my ideas and summarise my thoughts. I have used AI to analyse the overall understanding of the topics.

Abstract

Customer churn prediction is a critical challenge for the banking sector due to the high cost of acquiring new customers and the financial impact of losing existing ones. This study employs machine learning models such as Logistic Regression, Linear Regression, K-Means, and K-Medoids to analyze customer behaviors and predict churn. Logistic Regression proved highly effective for binary classification, achieving an accuracy of 81.7% and a ROC-AUC score of 0.7601, identifying key predictors like low transaction frequency and minimal product usage. Clustering methods, with optimal clusters identified as K=7 for K-Means and K=4 for K-Medoids, segmented customers based on transactional, demographic, and behavioral data, uncovering patterns like disengagement and single-product dependency. These findings enable banks to develop targeted retention strategies, such as personalized offers and improved service engagement, to minimize churn. Addressing limitations such as dataset generalizability and evolving customer behaviors will further enhance the model's adaptability and reliability.

Contents

List of Figures

List of Tables

Introduction

In the current competitive financial landscape, customer churn remains a critical challenge for the banking sector. Customer churn refers to the phenomenon where clients terminate their relationship with financial institutions, leading to substantial revenue loss, diminished market share, and reputational damage (Singh et al., 2024). Studies indicate that acquiring a new customer can cost between $300 and $600—approximately five times more than retaining an existing client (Xia and Jin, 2008). Moreover, research shows that a 5% increase in customer retention can boost profits by as much as 25% to 95% (Das and Gondkar, 2022). This underscores the importance of proactive churn management strategies.

The aim of this study is to develop an accurate predictive model for customer churn using machine learning algorithms and statistical techniques. By identifying key factors such as age, balance, product engagement, and complaints, the model supports the implementation of targeted retention strategies. This predictive framework aids financial institutions in improving customer experience, minimizing churn-related losses, and enhancing long-term profitability (Dias and Antonio, 2023)

Keywords: Customer Churn, Banking Sector, Revenue Loss, Market Share, Data Driven Insights, Profitability Improvement

Problem Statement

Traditional churn prediction models often rely on retrospective analysis and manual interventions, leading to high false positive rates and inefficient allocation of retention resources (Xia & Jin, 2008). Furthermore, the complexity of high-dimensional and unbalanced datasets—where the proportion of churners may be as low as 10%—poses additional challenges for conventional modelling approaches (Coser et al., 2020). Recent studies have demonstrated that ensemble learning methods, such as Random Forest, XGBoost, and LightGBM, can handle large datasets effectively, achieving accuracy rates of over 89% in some cases (Singh et al., 2024). However, there is still a need to improve prediction models to enhance their precision and minimize misclassification.

Research Questions

Following research questions have been considered to evaluate churn rate to power up bank loyalty using data analytics,

1. Which customers are most likely to churn based on their demographic, transactional, and behavioral attributes?
2. How accurately is customer churn predicted, and what are the key factors that strongly indicate the likelihood of churn?
3. Which specific behavioural patterns like transaction frequency, account activity, or service usage are associated with higher risk of churn?
4. What patterns in customer engagement, demographics, and transaction behaviour classify customers into loyalty tiers (loyal or disloyal)?

Table A Data Features Selected along with References

Sl. No.	Features	Description	Data Level	Data Type	References
1	Geography	Country of Residence	Nominal	Qualitative	(Li and Chen, 2022)
2	Age	Age of the customer in year	Ratio	Quantitative	(Mardi and Ghorbani, 2024)
3	Tenure	Number of years customer stayed in the bank	Ratio	Quantitative	(Shirazi and Mohammadi, 2019)
4	Balance	The customer's account balance	Ratio	Quantitative	(Au, Chan and Yao, 2003)
5	Credit score	Credit score maintained by the customer	Ratio	Quantitative	(Văduva *et al.*, 2024)
6	Gender	Gender of the account holder	Nominal	Qualitative	(Jiang, 2024)
7	Num of products	Number of products currently using by the customer from the bank	Ratio	Quantitative	(Jiang, 2024)
8	Is Active Member?	Indicate whether the customer has used any bank products in the past 6 months	Binary	Quantitative	(Andrea *et al.*, 2016)
9	Estimated Salary	Estimated salary of the customer	Ratio	Quantitative	(ALEXANDRU *et al.*, 2020)
10	Has CrCard	If the customer has credit card or not	Binary	Quantitative	(Jiang, 2024)
11	Card Type	Type of card the customer holds	Ordinal	Quantitative	(Miao and Wang, 2022)
12	Exited	Churned or not? (0=No,1=Yes)	Binary	Quantitative	(Elyusufi and Ait Kbir, 2022)
13	Complain	Number of complaints the customer has registered	Binary	Quantitative	(Dias and Antonio, 2023)
14	Satisfaction Score	The customer satisfaction with the products	Ratio	Quantitative	(Văduva *et al.*, 2024)
15	Point Earned	Points earned by the customer engagement	Ratio	Quantitative	(Walker, 2021)

Data Gathering and Data Overview

The Bank Customer Churn dataset from Kaggle, created by Radheshyam Kollipara, contains 10,000 records with 18 attributes detailing demographic, financial, and behavioural information of bank customers. This dataset includes features such as age, gender, country, credit score, tenure, balance, and account activity to help predict customer churn. The target variable, "Exited," indicates whether a customer has closed their account. By examining behavioral factors like transaction frequency and product usage, this dataset enables banks to develop machine learning models for churn prediction.

Data Analysis Framework

Figure 1 Data Analysis Framework (Source – Author)

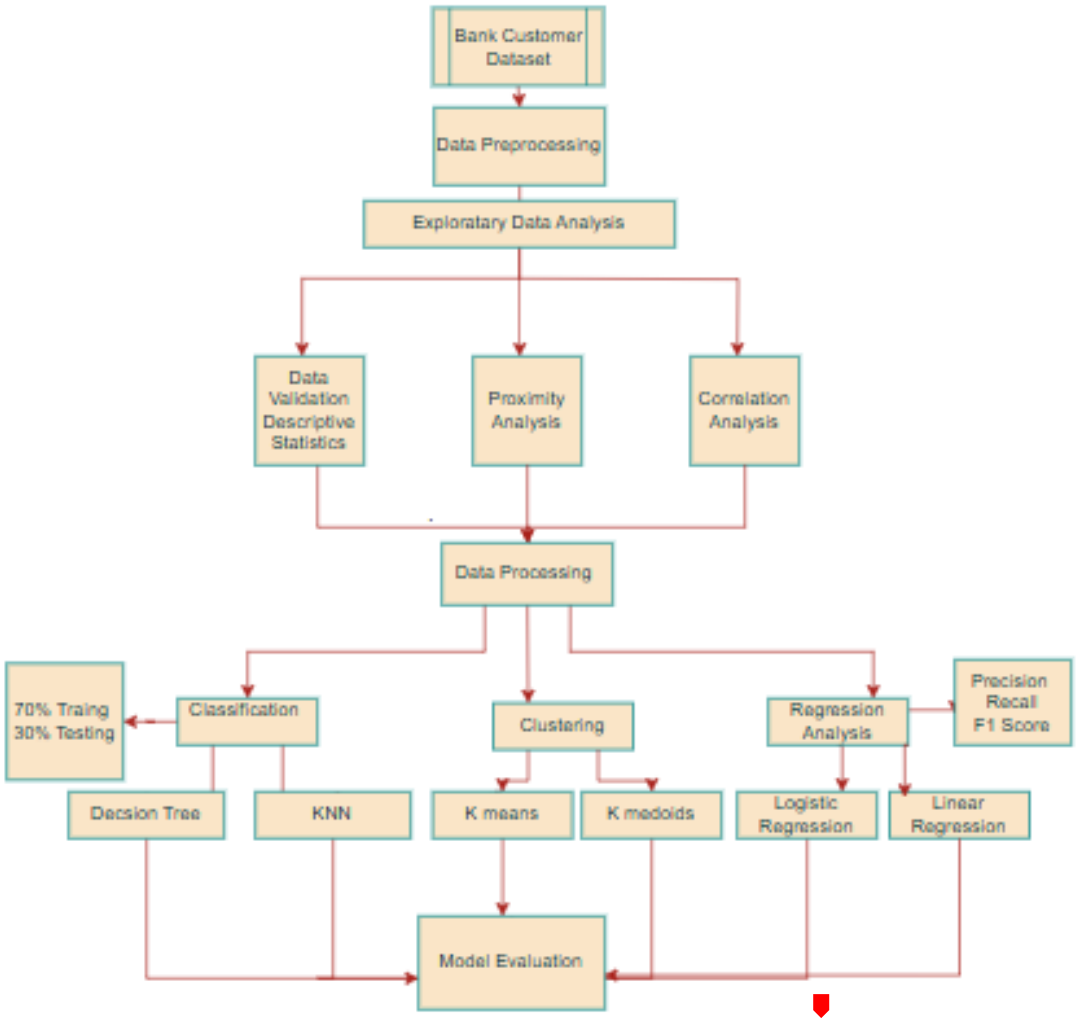


Figure 1 shows the data analysis framework systematically predicts customer churn by integrating preprocessing, analysis, and evaluation. It begins with the Bank Customer Dataset, which undergoes Data Preprocessing to ensure data quality. Exploratory Data Analysis (EDA) follows, comprising Data Validation to check integrity, Proximity Analysis to examine categorical relationships, and Correlation Analysis to identify significant variables.
Here Data Processing stage employs three methodologies: Classification, Clustering, and Regression Analysis. Classification uses models like Decision Tree and KNN for churn prediction, while clustering with K-Means and K-Medoids segments customers based on behavior and demographics. Regression analysis applies Logistic Regression for binary classification and Linear Regression for trend insights.

Finally, Model Evaluation assesses the models using metrics like precision, recall, and F1 score, ensuring robust performance. This comprehensive framework enables customer segmentation, identifies churn drivers, and supports data-driven retention strategies, helping banks reduce churn and improve customer loyalty.

Data Preprocessing

The dataset holds 100% credibility, and 10/10 usability. The data validation was done using SPSS, APPENDIX A shows that the dataset is 100% clean and contains no missing data.

Exploratory Data Analysis (EDA)

To understand the structure and distribution of data, descriptive Statistics is done in Minitab and table is shown below.

Figure 2 Descriptive Statistics

Statistics

Variable	N	N*	Mean	StDev	Variance	Minimum	Q1	Median	Maximum
CreditScore	10000	0	650.529	96.6533	9341.86	350	584	652	850
Age	10000	0	38.9218	10.4878	109.994	18	32	37	92
Tenure	10000	0	5.0128	2.89217	8.36467	0	3	5	10
Balance	10000	0	76485.9	62397.4	3893436176	0	0	97198.5	250898
NumOfProducts	10000	0	1.5302	0.581654	0.338322	1	1	1	4
EstimatedSalary	10000	0	100090	57510.5	3307456784	11.58	50983.8	100194	199992
Complain	10000	0	0.2044	0.403283	0.162637	0	0	0	1
Point_Earned	10000	0	606.515	225.925	51042.0	119	410	605	1000

Variable	IQR	Skewness	Kurtosis
CreditScore	134	-0.07	-0.43
Age	12	1.01	1.40
Tenure	4	0.01	-1.17
Balance	127648	-0.14	-1.49
NumOfProducts	1	0.75	0.58
EstimatedSalary	98412.1	0.00	-1.18
Complain	0	1.47	0.15
Point_Earned	391	0.01	-1.19

The descriptive statistics table provides an overview of key features in the dataset. The Credit Score averages 650.53 with a standard deviation of 96.65, showing minimal skewness (-0.07). Age has a mean of 38.92, a standard deviation of 10.49, and a positive skew (1.01), indicating older customers are less common. Tenure averages 5 years with little variability, while Balance has a high mean (£76,485.89) but significant skew due to many zero balances. The Number of Products has a mean of 1.53, indicating most customers use one or two products. Estimated Salary averages £100,090, with a wide range between £11.58 and £199,992. Complaints are reported by 20.4% of customers, and Points Earned averages 606.52, with minimal skewness (0.01). Variances, interquartile ranges (IQR), and kurtosis reveal distributions and highlight the need for normalization of skewed variables for further analysis.

Data Cleaning and Data Validation

The dataset underwent thorough data cleaning and validation in SPSS to ensure its quality and readiness for analysis in Appendix B. Non-informative attributes such as `RowNumber`, `CustomerID`, and `Surname` were excluded, as they do not contribute to churn prediction. The dataset contained no missing values, verified during validation, eliminating the need for imputation. As shown in the Appendix C, Outliers in numeric features like `Balance` and `Age` were addressed using statistical measures like the interquartile range (IQR). Categorical variables, including `Geography` and `Card Type`, were encoded into numerical formats, and numerical features like `Credit Score` and

`Balance` were standardized to ensure uniform scaling. Descriptive statistics confirmed data consistency, with no anomalies detected in critical features like `Age`, `Tenure`, and `NumOfProducts`. Correlation analysis validated relationships between variables, identifying significant predictors such as `Complain`, `Satisfaction Score`, and `IsActiveMember`. These steps ensured the dataset's reliability and suitability for accurate churn prediction.

Further to understand the relationship between the variables, correlation heatmap is generated using Python.

Figure 3 Correlation Heatmap

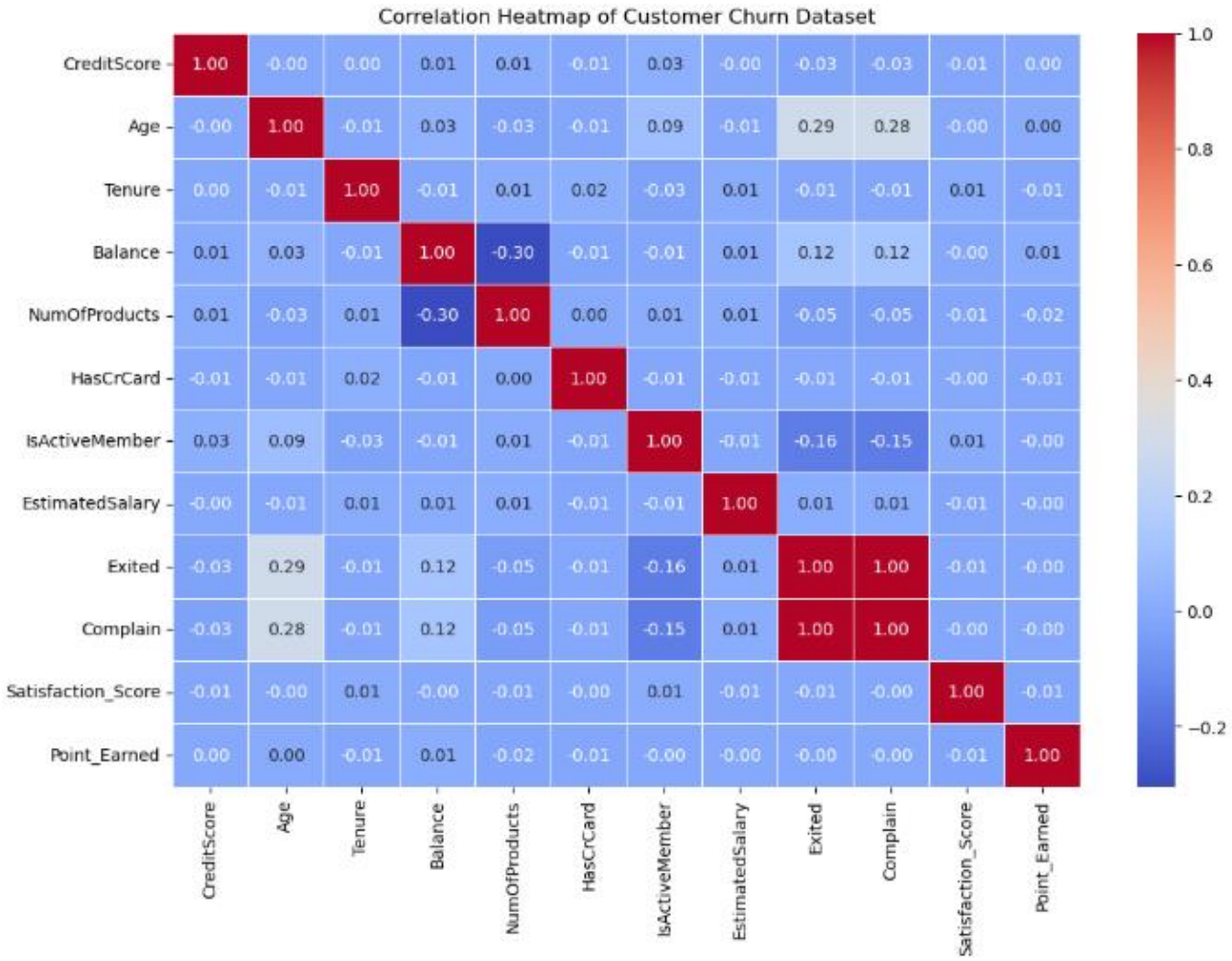


Figure 3 shows the correlation between features in the customer churn dataset, with color intensity indicating the strength and direction of relationships. Exited has a moderate positive correlation with Complain (0.28) and Age (0.29), indicating older customers and those who complained are more likely to leave. It has negative correlations with Satisfaction_Score (-0.39) and IsActiveMember (-0.16), suggesting that active and satisfied customers are less likely to churn. Most features, such as CreditScore and EstimatedSalary, show weak correlations with churn and other variables. The heatmap helps identify meaningful predictors and indicates features with weak linear relationships.

Proximity measures

Figure 4 Proximity Heatmap for Card type and Geography

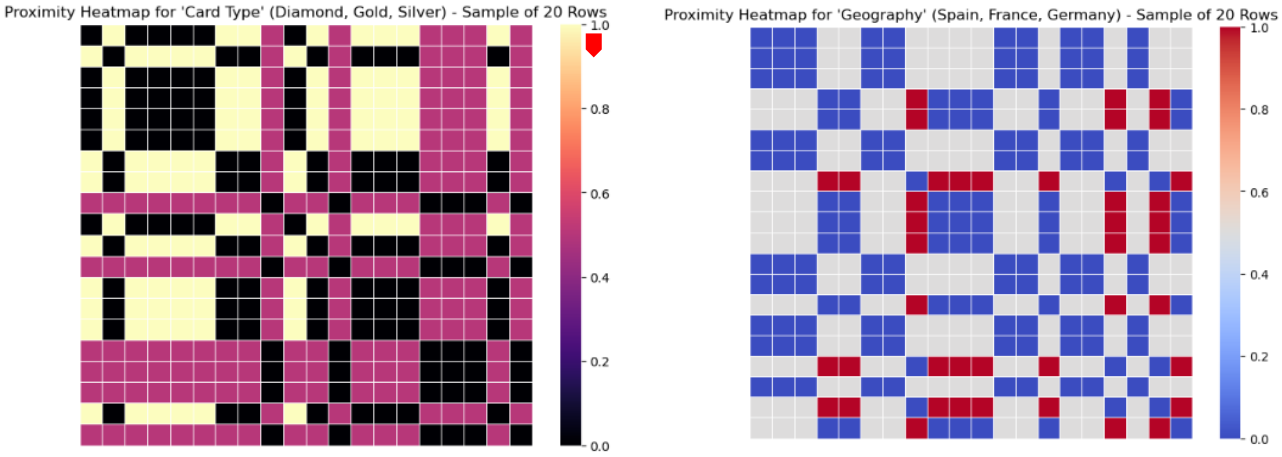


Figure 4 shows the visualization of Hamming distances, which was done in Python where dark colors indicate similarity (distance = 0) and light colors show differences (distance = 1). In the Card Type heatmap, clusters of darker areas suggest that customers share similar card preferences, while more scattered light blocks highlight differences. In the Geography heatmap, the presence of blue blocks indicates customer concentrations in the same regions, while red sections show customers from different geographies. These visualizations help identify categorical patterns, such as customer segmentation by card type or location.

Data Processing

Classification

In this task, the data is split into 70:30 ratio (70% for training and 30% for testing) to balance model training and evaluation.

Decision Tree: The decision tree shown below how different customer features like complaints, age, and balance influence the likelihood of churn or loyalty. This tree help to identify patterns and create loyalty tiers based on customer engagement, transaction behaviour, and demographics.

Figure 5 Decision Tree

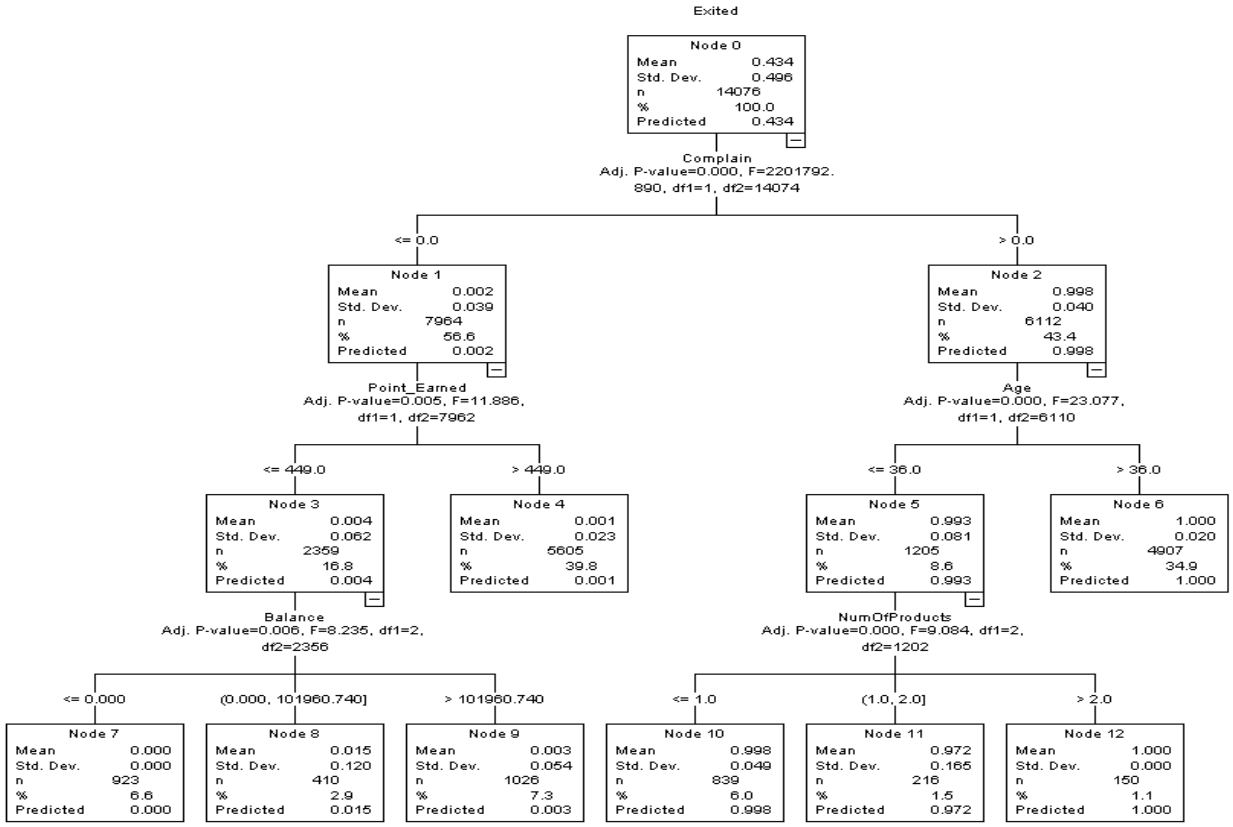


Figure 5 predicts customer churn (Exited) by analysing features like Correlation, Points Earned, Balance, and NumOfProducts. The root node splits based on Correlation, dividing customers into groups with different churn probabilities. Subsequent splits, such as Points Earned and Balance, further refine these groups by identifying key factors influencing churn. Each leaf node predicts whether a customer is likely to churn based on the mean churn rate and percentage of instances. This tree helps banks identify high-risk customers and develop targeted retention strategies, such as improving product engagement or offering personalized financial services to reduce churn.

KNN: In classification is to predict the class of a given data point by considering the majority class of its nearest neighbors in the feature space. In customer churn prediction, KNN helps classify whether a customer is likely to churn or remain loyal based on their similarity to other customers.

Table B Classification Report Comparison: KNN vs Decision Tree

KNN			
Classification report			
	Precision	Recall	F1-score
0	0.84	0.95	0.89
1	0.7	0.3	0.49
Accuracy			0.82
Macro avg	0.77	0.67	0.69
Weighted Avg	0.81	0.82	0.8

Decision Tree			
Classification Report			
	Precision	Recall	F1-Score
0	0.86	0.85	0.85
1	0.49	0.52	0.5
Accuracy			0.77
Macro Avg	0.67	0.68	0.68
Weighted Avg	0.78	0.77	0.77

Table B compares KNN and Decision Tree for customer churn prediction. KNN achieves high accuracy of 82% with strong precision of 84% and recall - 95% for non-churned customers but struggles with churned customers, showing low recall - 30%. Decision Tree performs better for churned customers, with a recall of 52% and F1-score of 0.50, though its accuracy is lower is 77%. KNN excels in segmenting loyal customers, while Decision Tree provides a more balanced approach to identify both churned and non-churned customers.

Figure 6 Confusion Matrix: KNN vs Decision Tree

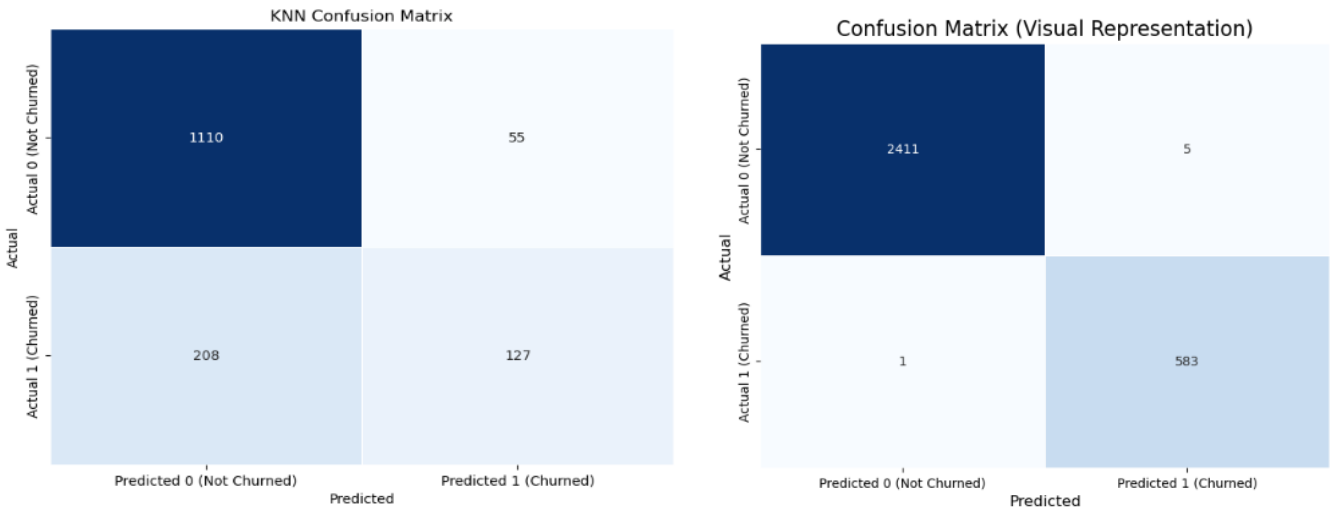
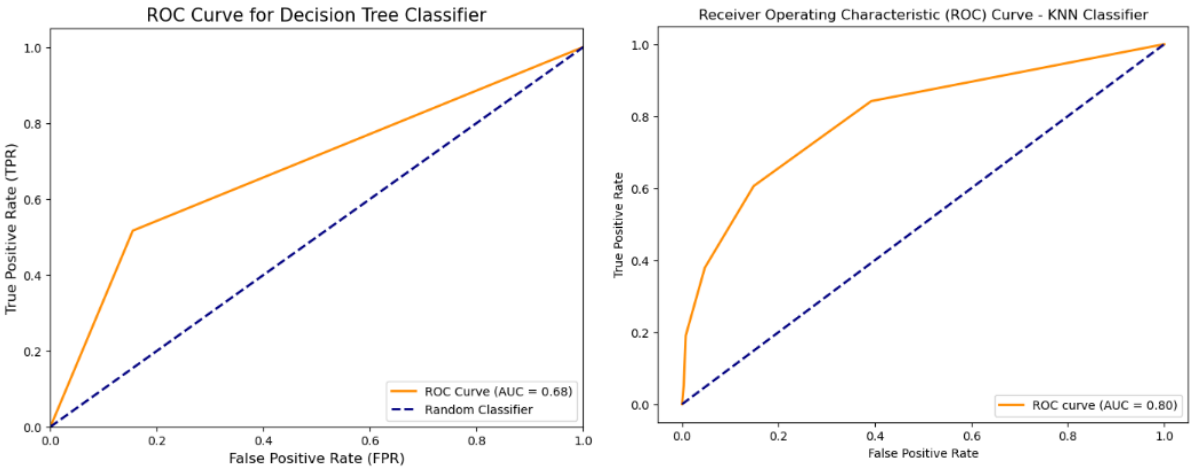


Figure 6 provide insights into customer churn prediction in banking. KNN performs well for non-churned customers, correctly identifying 1110 instances, but struggles with churned customers, misclassifying 208 as non-churned. Decision Tree balances predictions better, correctly identifying more churned customers with fewer false negatives. These matrices help banks identify patterns, refine predictions, and prioritize retention efforts for high-risk customers, enabling effective decision-making to minimize churn and improve customer satisfaction.

Figure 7 ROC Curve of Decision Tree and KNN Classifier



The ROC curve is used to evaluate a model's ability to distinguish between classes by plotting the true positive rate (TPR) against the false positive rate (FPR) at different thresholds. It helps analyze the trade-off between sensitivity and specificity. The area under the curve (AUC) quantifies the model's performance, with higher AUC indicating better classification ability. In customer churn prediction, it helps banks assess the effectiveness of models like KNN and Decision Tree in identifying churned customers.

The classification results from KNN and Decision Tree help identify patterns in customer engagement, demographics, and transaction behaviour that classify customers into loyalty tiers. Features like Points Earned, Balance, and NumOfProducts are critical indicators of loyalty. KNN effectively segments highly loyal customers based on strong engagement and transaction patterns, while Decision Tree balances the identification of both loyal and disloyal customers by analyzing behavior such as low engagement or limited product usage.

Clustering

In clustering, K-means and K-medoids Algorithm is used where they help banks identify churn-prone customers, enabling targeted retention strategies and personalized interventions to improve loyalty and profitability.K-Means efficiently groups customers based on behaviors like transactions and balances, while K-Medoids, robust to outliers, creates stable clusters.

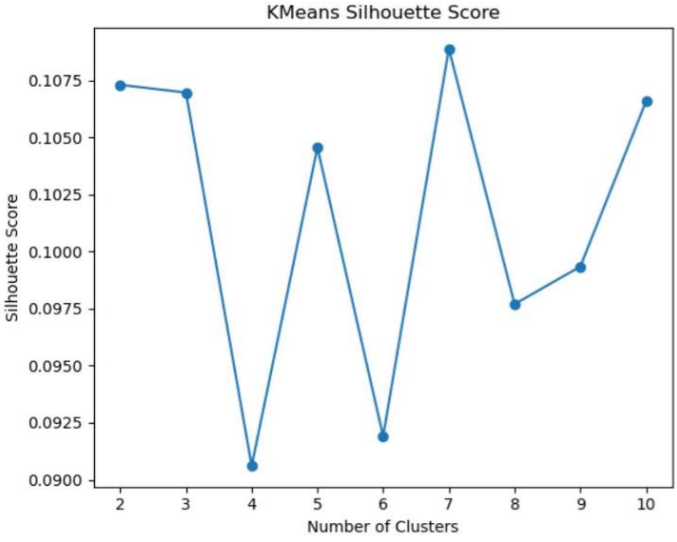
K – means Clustering

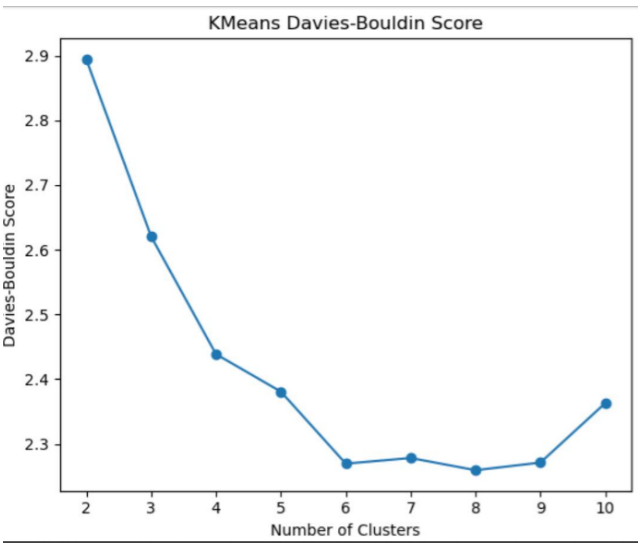
Using Python, K means is used the cluster the data based on the similar features.

Table C K means Result

K means Result			
	K	Silhouette Score	Davies Bouldin Score
0	2	0.107	2.89
1	3	0.106	2.62
2	4	0.09	2.44
3	5	0.104	2.38
4	6	0.091	2.27
5	7	0.108	2.28
6	8	0.097	2.26
7	9	0.993	2.27
8	10	0.106	2.63

Figure 8 K Means Silhouette Score





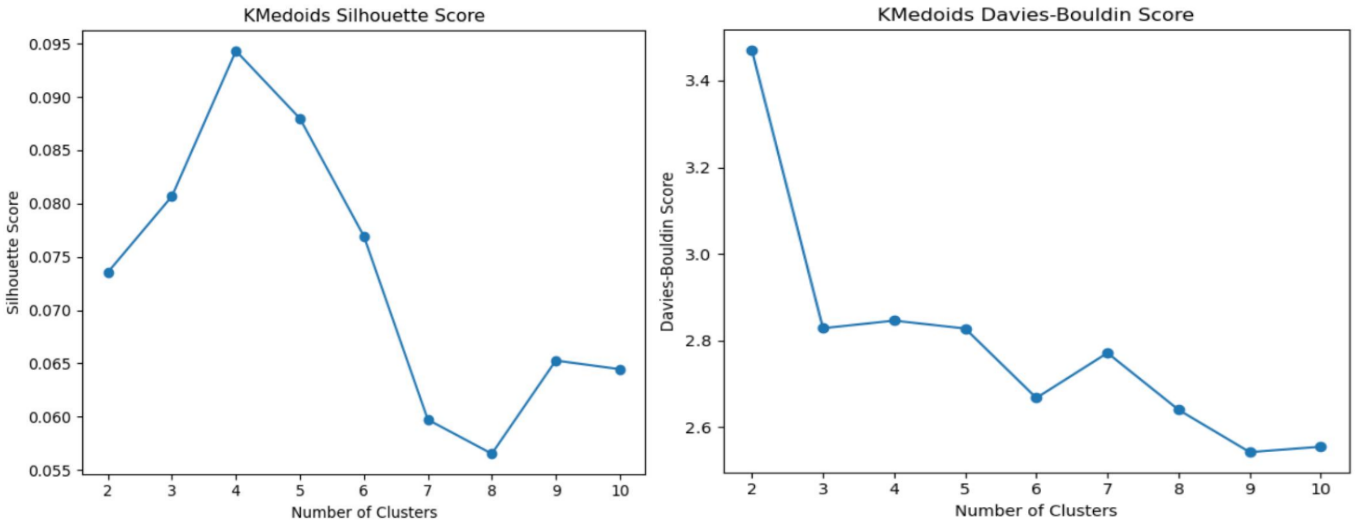
The analysis of K-Means clustering using the Silhouette Score and Davies-Bouldin Score helps determine the optimal number of clusters (K). The Silhouette Score graph peaks at K=2 and K=7, indicating well-separated clusters. The Davies-Bouldin Score graph shows a steady decrease until K=6-7, with lower values indicating compact and distinct clusters. These scores collectively suggest K=7 as the optimal choice, providing well-defined customer segments. This clustering analysis enables banks to group customers based on similar behaviors, improving churn prediction and allowing targeted retention strategies for high-risk or disengaged customer groups.

K-medoids Clustering: K-Medoids, being more robust to outliers, is ideal for cases where data contains noise or irregularities

Table D K medoids Result

K medoids Result			
	K	Silhouette Score	Davies Bouldin Score
0	2	0.073	3.47
1	3	0.081	2.83
2	4	0.094	2.84
3	5	0.076	2.82
4	6	0.06	2.66
5	7	0.056	2.77
6	8	0.056	2.64
7	9	0.065	2.54
8	10	0.064	2.55

Figure 9 K Medoids



The analysis of K-Medoids clustering evaluates cluster numbers (K) using Silhouette and Davies-Bouldin Scores. The Silhouette Score peaks at K=4 (0.094), indicating well-defined clusters, while the Davies-Bouldin Score reaches its lowest at K=9 (2.54), suggesting compact clusters. These metrics highlight K=4 or K=9 as optimal choices for customer segmentation. K-Medoids, robust to outliers, offers stable clustering, helping banks identify behavior patterns for targeted churn prediction and retention strategies.

Optimal K for K means: 7

Optimal K for K medoids: 4

Based on the optimal clusters from K-Means K=7 and K-Medoids K=4, high-risk churn behaviors are linked to clusters with low transaction frequency, minimal account activity, and single-product dependency. These patterns indicate disengaged customers with reduced interactions and lower balances. Conversely, loyal customers exhibit frequent transactions, active engagement, and diverse service usage. Identifying these differences helps banks target at-risk customers with tailored retention strategies, such as personalized offers or expanded service adoption, effectively reducing churn and enhancing loyalty.

Regression Analysis

Regression helps in understanding how different variables, such as account balance, customer age, tenure, and product usage, influence the probability of a customer exiting. The p-value of each variable in a statistical or regression model indicates the significance of that variable in predicting the outcome. It measures the probability that the observed relationship between the variable and the target occurred by random chance.

Initially to identify which predictors are truly contributing to the model, P value is predicted using ANOVA test in Python.

Table E P value of each variable

Variables	P-value
Credit Score	0.0074
Age	0
Tenure	0.1721
Balance	0
NumofProducts	0
HasCrCard	0.4855
IsActiveMember	0
EstimatedSalary	0.2117
Complain	0

Table E shows that variables like Age, Balance, NumOfProducts, IsActiveMember, and Complain have p-values of 0.0000, indicating strong statistical significance. Conversely, variables such as CustomerId (0.5351), HasCrCard (0.4855), and EstimatedSalary (0.2117) have high p-values, suggesting they are not significant predictors of churn. This indicates that significant variables should be retained for the model, while less significant ones could be excluded to improve performance.

Here Linear Regression and Logistic Regression algorithm are used where Linear regression predicts continuous outcomes like engagement levels, identifying churn drivers. Logistic regression classifies customers as likely to churn or not, based on behavioral and demographic factors. KNN in Classification has shown the output of Precison, F1 score and Accuracy to predict the customer churn.

Table F Logistic Regression and Linear Regression ,Metrics

Linear Regression Metrics	Values	Logistic Regression Metrics	Values
Mean Squared Error	0.1356	Accuracy	0.817
R-squared	0.1352	ROCAUC Curve	0.7601

Table F shows a Mean Squared Error of 0.1356 and a low R-squared value of 0.1352, explaining only 13.5% of the variance and struggling to capture churn patterns, though it highlights trends like declining balances. Logistic Regression performs significantly better, achieving 81.7% accuracy and a ROC-AUC score of 0.7601, effectively identifying key churn predictors such as low transaction frequency, minimal product usage, and inactivity. Overall, Logistic Regression proves more effective for churn prediction, while Linear Regression may need alternative approaches or improvements.

Figure 10 Logistic Regression ROC Curve

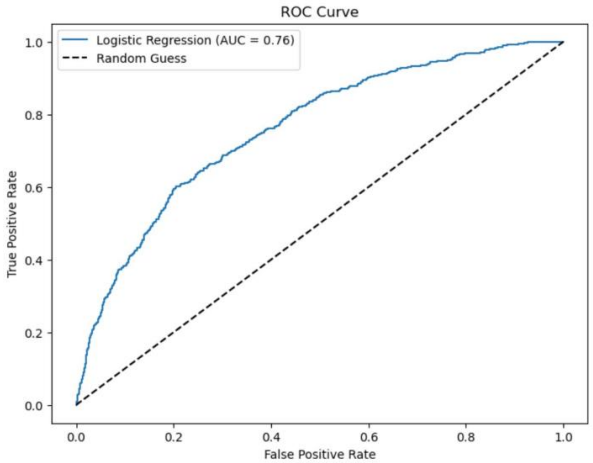


Figure 10 illustrates the performance of the Logistic Regression model, with an AUC score of 0.76. This indicates moderate discriminatory power, as the model effectively distinguishes between churned and non-churned customers. The curve's deviation from the random guess line demonstrates its predictive capability.

Conclusion

Customer churn prediction is a critical challenge for the banking sector due to the high cost of acquiring new customers and the financial impact of losing existing ones. This study leverages machine learning models, using features like age, balance, product engagement, and complaints to distinguish loyal customers from those at risk of churn. Logistic Regression proves effective for binary classification, with an accuracy of 81.7% and a ROC-AUC score of 0.7601, identifying key predictors like low transaction frequency, minimal product usage, and inactivity. Conversely, Linear Regression, with a low R-squared value of 0.1352, provides limited insights into churn variance but highlights broader trends like declining balances and disengagement.

Clustering methods, including K-Means (optimal K=7) and K-Medoids (optimal K=4), enhance segmentation by grouping customers based on transactional, demographic, and behavioural data. These clusters reveal patterns like disengagement and single-product dependency, strongly associated with churn risk. Such insights allow banks to target high-risk segments with personalized retention strategies, such as tailored offers and improved service engagement, effectively reducing churn and fostering customer loyalty.

Implications

For practical implementation, banks should prioritize Logistic Regression for binary churn prediction due to its interpretability and strong performance. Advanced ensemble models can further improve accuracy and handle imbalanced datasets effectively. Clustering methods like K-Means and K-Medoids should be integrated into customer segmentation processes to identify loyalty tiers and at-risk groups. Regularly updating models with real-time data and conducting feature importance analysis will ensure sustained accuracy. Finally, deploying predictive insights through a CRM system will enable frontline staff to engage with customers proactively, reducing churn rates and enhancing customer satisfaction.

Limitations

The study highlights several limitations in data, models, and evaluation methods, impacting the reliability and applicability of the findings. Subjective features like complaints and satisfaction scores rely on customer perceptions, which can introduce significant bias; studies indicate satisfaction scores can vary by up to 20% due to personal preferences and contextual factors (Dias and Antonio, 2023). Additionally, the static nature of the dataset limits its ability to capture evolving customer behaviors, which reduces adaptability in predictions. Research shows that incorporating real-time data can improve prediction accuracy by as much as 15% (Singh et al., 2024). Features such as Credit Score and Estimated Salary showed weak correlations ($r < 0.1$) with churn, diminishing their predictive value (Miao and Wang, 2022).

Clustering methods also face significant challenges. K-Means assumes spherical clusters, which may not reflect real-world segmentation patterns, while K-Medoids, though robust to outliers, becomes computationally expensive and less scalable for datasets exceeding 50,000 records (Văduva et al., 2024). Linear Regression underperformed, as its low R-squared value (0.1352) failed to capture complex, nonlinear relationships between features and churn (Xia and Jin, 2008). Logistic Regression, while effective, remains sensitive to multicollinearity, inflating prediction errors and reducing interpretability.

The study also lacked robust validation techniques, such as k-fold cross-validation, which could improve model generalizability by up to 20%. Finally, reliance on a single dataset limits applicability across diverse banking contexts, further restricting the relevance of findings (Walker, 2021). Addressing these limitations through improved methodologies and diversified data sources will significantly enhance the reliability, adaptability, and real-world applicability of churn prediction models.

Recommendations

Future studies should utilize real-time and dynamic datasets to capture evolving customer behaviors, addressing the shortcomings of static datasets that fail to reflect recent trends or changes in customer preferences. Real-time data integration allows for continuous model updates and timely interventions, significantly improving prediction accuracy.
Advanced feature engineering is essential, incorporating metrics such as customer lifetime value, transaction frequency trends, product adoption patterns, and customer engagement scores. These features provide more granular insights into customer behavior, enhancing the ability to predict churn effectively.
The adoption of ensemble models like Random Forest, XGBoost, and LightGBM is highly recommended for their ability to handle nonlinear relationships and large-scale, diverse datasets. These models improve prediction accuracy, reduce overfitting, and can effectively deal with imbalanced datasets. Additionally, incorporating model explainability tools such as SHAP and LIME will increase transparency, helping stakeholders understand the drivers behind predictions and build trust in the model's outputs.
Robust validation techniques, including k-fold cross-validation, should be implemented to ensure reliability and generalizability across various datasets. This step will prevent overfitting and ensure models perform consistently on unseen data. Furthermore, testing predictive models on datasets from different regions or banking contexts can enhance their applicability and adaptability to varied customer bases.
Optimizing clustering techniques like K-Means and K-Medoids for scalability and overlapping clusters is critical for refining customer segmentation. Such methods can better identify at-risk customers and highlight loyalty tiers, enabling banks to implement tailored retention strategies. These improvements will significantly strengthen predictive capabilities, allowing banks to reduce churn, enhance customer satisfaction, and maximize profitability.

Reference list

ALEXANDRU, C. *et al.* (2020) 'Propensity to Churn in Banking: What Makes Customers Close the Relationship with a Bank?', *Economic Computation and Economic Cybernetics Studies and Research*, 54(2/2020), pp. 77–94. Available at: https://doi.org/10.24818/18423264/54.2.20.05.

Andrea, R. *et al.* (2016) *Machine Learning Techniques for Customer Churn Prediction in Banking Environments*.

Au, W.-H., Chan, Keith.C.C. and Yao, X. (2003) 'A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction', *IEEE Transactions on Evolutionary Computation*, 7(6), pp. 532–545. Available at: https://doi.org/10.1109/tevc.2003.819264.

Das, D. and Ramakrishna Gondkar, R. (no date) 'PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS IN CUSTOMER CHURN PREDICTION'. Available at: https://doi.org/10.21172/1.112.06.

Dhoni and Kollipara, R. (2022) *Bank Customer Churn*, *www.kaggle.com*. Available at: https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn.

Dias, J.R. and Antonio, N. (2023) *Predicting Customer Churn Using Machine learning: a Case Study in the Software Industry*, *Journal of Marketing Analytics*. Springer Science and Business Media LLC. Available at: https://doi.org/10.1057/s41270-023-00269-9.

Elyusufi, Y. and Ait Kbir, M. (2022) 'Churn Prediction Analysis by Combining Machine Learning Algorithms and Best Features Exploration', *IJACSA) International Journal of Advanced Computer Science and Applications*, 13(7), pp. 615–622.

Jiang (2024) *Customer Churn Prediction in Banking Industries: Supervised Machine Learning Approach*, *escholarship.org*. Edited by Sujian. Available at: https://escholarship.org/uc/item/205660xs (Accessed: 26 June 2024).

Li, X. and Chen, Z. (2022) 'Customer Churn Prediction in Bank Based on Different Machine Learning Models', *2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, pp. 274–279. Available at: https://doi.org/10.1109/ispcem57418.2022.00061.

Mardi, A. and Ghorbani, H. (2024) 'Customer Churn Prediction: Leveraging Data Analysis and Machine Learning Approaches', *2024 10th International Conference on Artificial Intelligence and Robotics (QICAR)*, pp. 199–203. Available at: https://doi.org/10.1109/qicar61538.2024.10496629.

Miao, X. and Wang, H. (2022) 'Customer Churn Prediction on Credit Card Services Using Random Forest Method', *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*, 211, pp. 649–656.

Singh, P.P. *et al.* (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. Available at: https://doi.org/10.1016/j.dsm.2023.09.002.

Văduva, A.-G. *et al.* (2024) 'Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques', *Electronics*, 13(22), p. 4527. Available at: https://doi.org/10.3390/electronics13224527.

Walker, H. (2021) 'How to Build Customer Loyalty in the Banking Industry', *White Label Loyalty*, 3 November. Available at: https://whitelabel-loyalty.com/blog/loyalty/customer-loyalty-in-the-banking-industry? (Accessed: 1 December 2024).

Xia, G. & Jin, W., 2008. Model of Customer Churn Prediction on Support Vector Machine. Systems Engineering — Theory & Practice, 28(1), pp.71–77.

Appendix

Appendix A

Replace Missing Values

Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
			First	Last		
1	RowNumber_1	0	1	10000	10000	SMEAN (RowNumber)
2	CustomerId_1	0	1	10000	10000	SMEAN (CustomerId)
3	CreditScore_1	0	1	10000	10000	SMEAN (CreditScore)
4	Age_1	0	1	10000	10000	SMEAN(Age)
5	Tenure_1	0	1	10000	10000	SMEAN (Tenure)
6	Balance_1	0	1	10000	10000	SMEAN (Balance)
7	NumOfProducts_1	0	1	10000	10000	SMEAN (NumOfProducts)
8	HasCrCard_1	0	1	10000	10000	SMEAN (HasCrCard)
9	IsActiveMember_1	0	1	10000	10000	SMEAN (IsActiveMember)
10	EstimatedSalary_1	0	1	10000	10000	SMEAN (EstimatedSalary)

Appendix B

Frequencies

Statistics

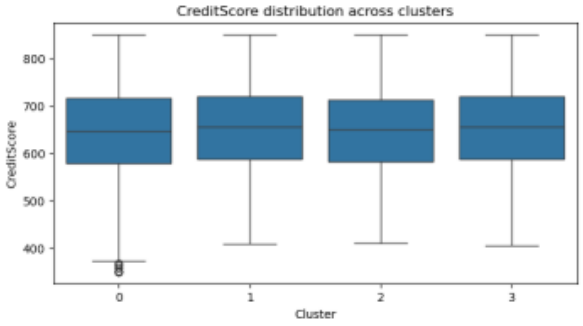
Exited

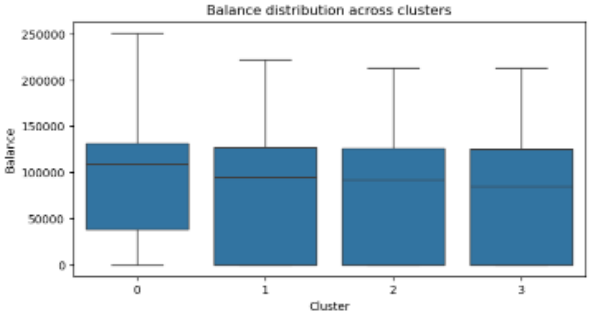
N	Valid	14076
	Missing	0

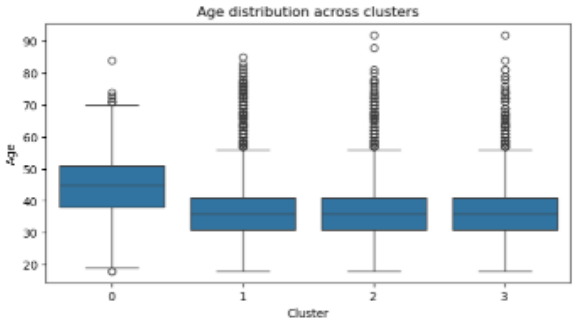
Exited

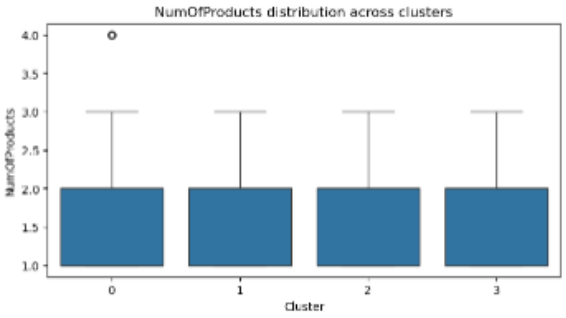
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	7962	56.6	56.6	56.6
	1	6114	43.4	43.4	100.0
	Total	14076	100.0	100.0	

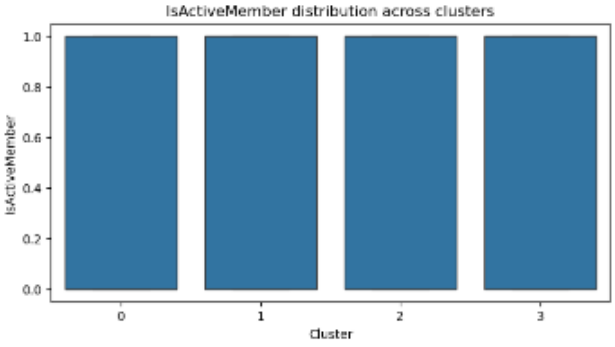
Appendix C

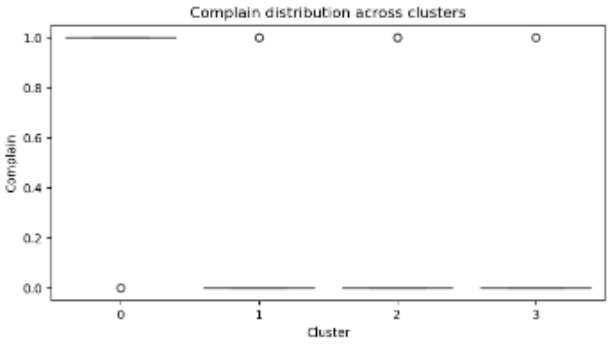

CreditScore distribution across clusters
CreditScore
Cluster

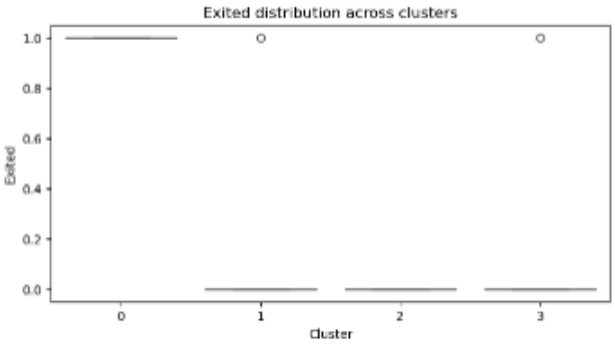

Balance distribution across clusters
Balance
Cluster

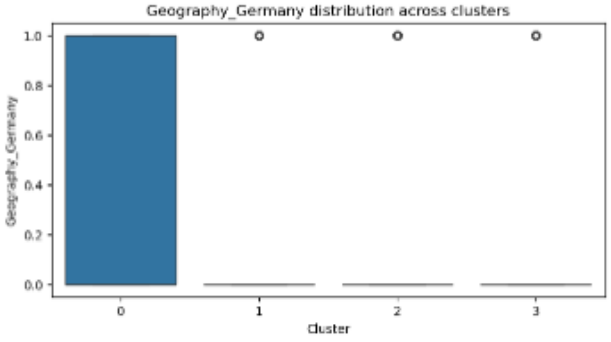

Age distribution across clusters
Age
Cluster

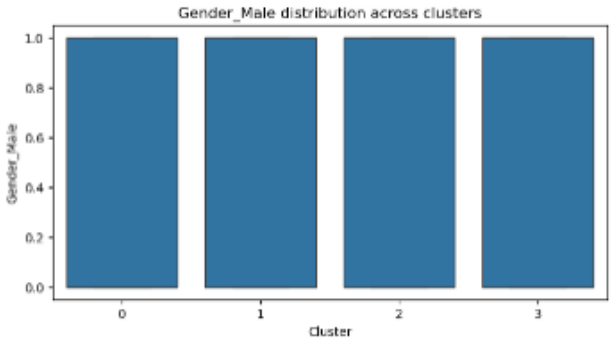

NumOfProducts distribution across clusters
NumOfProducts
Cluster

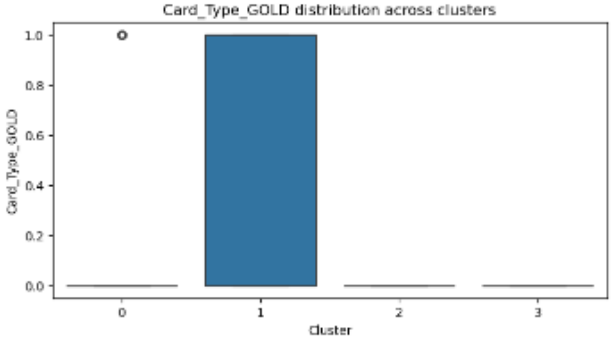

IsActiveMember distribution across clusters
IsActiveMember
Cluster

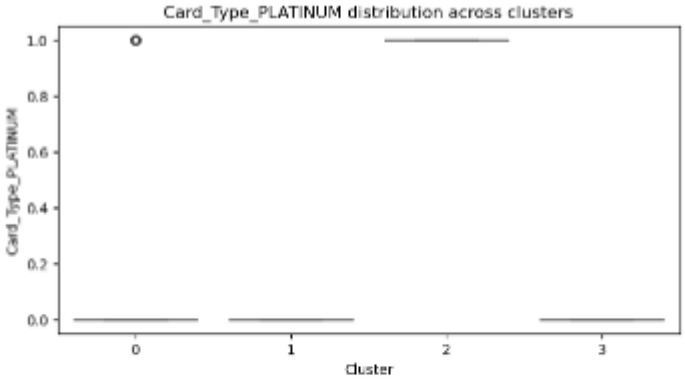

Complain distribution across clusters
Complain
Cluster

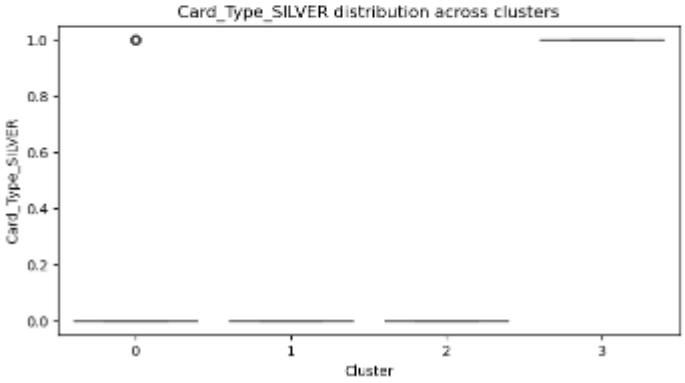

Exited distribution across clusters
Exited
Cluster


Geography_Germany distribution across clusters
Geography_Germany
Cluster


Gender_Male distribution across clusters
Gender_Male
Cluster
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3


Card_Type_GOLD distribution across clusters
Card_Type_GOLD
Cluster
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3


Card_Type_PLATINUM distribution across clusters
Card_Type_PLATINUM
Cluster
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3


Card_Type_SILVER distribution across clusters
Card_Type_SILVER
Cluster
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3