Examining Determinants of Customer Bank Selection Using Logistic Regression

Abstract

This study examines the key factors influencing individuals' choice between public and private banks. The analysis focuses on value-added services, perceived risk, reputation, and perceived costs, and applies a binary logistic regression approach to assess their effects on banking preferences. The findings demonstrate that value-added services and institutional reputation significantly increase the likelihood of selecting private banks, while perceived risk and perceived costs significantly reduce this likelihood. These results confirm that consumers place strong emphasis on service enhancements and brand credibility, while remaining highly sensitive to risk exposure and cost considerations. Overall, the study advances empirical understanding of consumer behavior in the banking sector and offers insights relevant to banking strategy, competition, and customer-oriented policy design.

Keywords: Bank Choice, Binary Logistic Regression, Value Added Services, Perceived Risk, Perceived Costs

UNDER PEER REVIEW

Introduction

Binary Logistic Regression Analysis is a statistical technique used to explore the relationship between a binary dependent variable (also referred to as the outcome or response variable) and one or more independent variables (also called predictors or covariates). This method is widely applied across various fields, including healthcare, social sciences, and business, to model and predict the likelihood of an event or outcome based on predictor variables. For instance, **Hosmer et al. (2013) [1]** studied how binary logistic regression can be used to predict the probability of a patient having a specific disease based on independent variables such as age, gender, or health history. The binary logistic regression model is called "logistic" because it uses the log of the odds ratio rather than directly modeling the probability itself. Specifically, the logistic function is used to model the relationship between the dependent variable and the independent variables in terms of odds. The fundamental assumption of binary logistic regression is that the log-odds of the dependent variable are

linearly related to the independent variables, as stated by **Allison (2012) [2]**. The model estimates coefficients for the independent variables, which represent the magnitude and direction of the influence these variables have on the outcome.

Once the coefficients are estimated, they can be used to calculate the predicted probability of the outcome variable, given specific values of the independent variables. One of the key advantages of binary logistic regression is its ability to handle both categorical and continuous independent variables, making it highly flexible for a wide range of applications, as noted by **Menard (2002) [3]**. The researcher essentially answers the question, "What is the probability that a given case falls into one of the two categories of the dependent variable, given the predictors in the model?" A natural question might arise: why not use ordinary least squares (OLS) regression instead of binary logistic regression? While OLS regression assumes a linear relationship between the independent variables and the dependent variable, as well as normally distributed residuals with constant variance, these assumptions are violated when the outcome variable is binary, as explained by **Pituch and Stevens (2016) [4]**. Specifically, the dependent variable in an OLS model would not be restricted to values between 0 and 1, potentially leading to predicted probabilities outside this range, which is nonsensical for binary outcomes. In contrast, binary logistic regression estimates the relationship between the predictors and the outcome variable while accounting for the bounded nature of the probability, ensuring that predicted probabilities fall within the 0 to 1 range. Additionally, logistic regression does not require residuals to be normally distributed or exhibit constant variance, making it a more suitable choice for binary outcomes (Pituch & Stevens, 2016) .

In the context of bank choice modeling, binary logistic regression is particularly useful in analyzing factors influencing customers' preference for public or private banks. For example, variables such as perceived risk, value-added services, and reputation can be modeled to predict the probability of an individual selecting a private or public bank. Given the binary nature of the outcome variable—public vs. private bank choice—binary logistic regression provides a robust framework for understanding and predicting consumer behavior in financial services.

Related Work

2.1 Model Estimation

Binary Logistic Regression (BLR), unlike Ordinary Least Squares (OLS) regression, utilizes Maximum Likelihood Estimation (MLE) to estimate the parameters of the model. MLE is a statistical method that seeks to identify the parameter values that maximize the likelihood of obtaining the observed sample data. This approach operates iteratively, refining the estimates until the values converge to the ones that most likely produced the data under the assumed model. In contrast, OLS regression minimizes the sum of squared residuals to estimate parameters, assuming a linear

UNDER PEER REVIEW

relationship between the dependent and independent variables. One of the key strengths of Maximum Likelihood Estimation in BLR is its flexibility in dealing with the non-linear relationship between the independent variables and the outcome, especially when the dependent variable is binary. However, MLE assumes the availability of sufficiently large sample sizes for reliable estimation. With smaller sample sizes, challenges related to model convergence and parameter estimation may arise, and these can lead to biased or inefficient estimates. Moreover, smaller sample sizes may reduce the power of the statistical tests, making it more difficult to detect significant relationships between the predictors and the outcome.

In such situations where sample size is limited, alternative estimation methods such as Exact Logistic Regression or the Firth procedure can be employed. Both techniques adjust the likelihood estimation to improve performance with smaller samples, particularly by addressing issues like separation in the data (where the outcome is perfectly predicted by certain predictors). However, these methods are not commonly available in most mainstream statistical software packages, which limits their accessibility in routine analyses. For our analysis, the binary logistic regression model will be employed to forecast the likelihood of specific outcomes based on the independent variables, such as avoidance of disclosure and symptom severity. These predictors will be used to model the probability that individuals fall into one of two categories of the outcome variable. Additionally, the results of the logistic regression will be presented in the form of odds ratios for each predictor. The odds ratio quantifies the strength of the association between each independent variable and the likelihood of the outcome occurring. A value greater than 1 suggests that as the predictor increases, the odds of the outcome happening also increase, while a value less than 1 suggests the opposite.

The accompanying table will summarize these odds ratios along with their confidence intervals, providing insight into the precision of the estimates. By examining the statistical significance of each predictor, we will be able to make informed conclusions regarding the relative importance of the independent variables in predicting the outcome. This approach not only offers a detailed interpretation of the individual effects of each predictor but also enables us to assess the overall fit and validity of the model. Through these analyses, the study aims to provide a comprehensive understanding of how avoidance of disclosure and symptom severity influence the likelihood of the outcome, contributing valuable insights to the field.

UNDER PEER REVIEW

2.2 Evaluation of Model fit

In binary logistic regression, assessing the model's fit is essential for determining how well the model predicts outcomes. Several methods are available to evaluate model performance, including goodness-of-fit tests, pseudo R-squared values, and receiver operating characteristic (ROC) curves. These methods assess different aspects of the model's adequacy and predictive ability. Goodness-of-fit tests are used to compare the observed frequencies with the expected frequencies derived from the

model's predictions. One of the most commonly used goodness-of-fit tests is the Hosmer-Lemeshow test. **Hosmer, Hosmer, & Lemeshow (1980) [5]** find that this test divides the data into deciles based on predicted probabilities and compares the observed and expected frequencies within each group. A significant result indicates that the model does not fit the data well, suggesting that the model's predictions are not aligned with the observed outcomes, and the model may need to be revised or improved to better capture the data's structure.

In addition to goodness-of-fit tests, pseudo R-squared values are another commonly used method to measure the extent to which the model explains the variability in the data. **Nagelkerke (1991) [6]** discusses how Nagelkerke's R-squared is an adjusted version of the Cox and Snell R-squared and is specifically designed for logistic regression models. Nagelkerke's R-squared can range from 0 to 1, with values closer to 1 indicating that the model explains a large portion of the variability in the outcome variable. Nagelkerke (1991) suggests that this pseudo R-squared measure is particularly useful when comparing models with different numbers of predictors or when assessing the model fit in more complex logistic regression models. On the other hand, Cox and Snell's R-squared **(Cox & Snell, 1989) [7]** is an alternative pseudo R-squared measure that is derived from the likelihood ratio test. However, it has an upper limit that is less than 1, making it less intuitive than Nagelkerke's version. According to Cox & Snell (1989), this measure is calculated by comparing the likelihood of the null model (without predictors) to the likelihood of the model with the predictors. While it provides valuable insight into the model's explanatory power, it is not as widely used or as easily interpretable as Nagelkerke's R-squared.

Similarly, **McFadden's R-squared (McFadden, 1974) [8]** is another pseudo R-squared measure that is based on the likelihood ratio, but it differs from Cox & Snell's R-squared by using a different formulation. McFadden (1974) argues that this measure tends to be smaller than R-squared values in ordinary least squares regression but still provides a useful gauge of the relative improvement in model fit over a null model. It ranges from 0 to 1, with higher values indicating better model fit. **McFadden (1974) [9]** notes that McFadden's R-squared is particularly useful for comparing models within the context of logistic regression, although its values are generally lower than those produced by R-squared measures in linear regression models. Another important tool for evaluating model fit is the receiver operating characteristic (ROC) curve. The ROC curve, as discussed by Fawcett (2006), plots the trade-off between sensitivity (true positive rate) and specificity (true negative rate) for various classification thresholds. The area under the ROC curve (AUC) provides a summary of the model's discriminative ability. According to Fawcett (2006), the AUC ranges from 0.5 (no better than random guessing) to 1 (perfect classification), with higher values indicating better model performance. The AUC is widely used as a measure of the model's ability to distinguish between the two categories of the outcome variable. It is important to note that no single measure can fully capture the performance of a logistic regression model. Hosmer, Hosmer, & Lemeshow (1980) emphasize that goodness-of-fit tests, pseudo R-squared values, and ROC curves

UNDER PEER REVIEW

each provide unique insights into the model's fit and predictive accuracy. Therefore, a combination of these measures should be used to comprehensively assess the model's performance. This multi-faceted approach ensures that the model is both statistically valid and practically useful.

2.3 Bank Predictions

Zaki et al. (2024) [10] explore the use of predictive analytics and machine learning in direct marketing for bank term deposit subscriptions. The study applies data exploration, visualization, and feature engineering techniques using Kaggle datasets. Various models, including the SGD Classifier, k-nearest neighbor Classifier, and Random Forest Classifier, are evaluated. The Random Forest Classifier outperforms the others, achieving 87.5% accuracy, a negative predictive value (NPV) of 92.9972%, and a positive predictive value (PPV) of 87.8307%. These findings offer valuable insights for optimizing banking marketing strategies. **Rahman et al. (2023) [11]** investigate how customers' perceptions of Islamic banking services influence satisfaction and word of mouth (WOM). The study uses bootstrapping procedures and partial least squares methods to analyze data from 377 respondents in Dhaka city. The findings show a strong relationship between security and customers' perceptions. Ethical responsibility and religious value positively impact perceptions, while benefits have a negative effect. Additionally, customer perceptions mediate the influence of security, ethical responsibility, religious value, and benefits on satisfaction. Satisfaction, in turn, mediates the relationship between customer perceptions and WOM. These insights can help Islamic bank managers enhance customer satisfaction and WOM, ultimately providing a competitive edge in the market.

Tran, Le, and Nguyen (2023) [12] examine customer churn prediction in the banking sector using machine learning models. The study compares the impact of customer segmentation on prediction accuracy and evaluates models such as k-nearest neighbors, logistic regression, decision tree, random forest, and support vector machine. The results show that customer segmentation has minimal impact on accuracy, with the Random Forest model achieving the highest accuracy (97.25%). The findings suggest that the proposed models can be applied to other sectors like education and marketing, with future research focusing on real-time applications and handling imbalanced data. **Singh et al. (2024) [13]** investigate customer churn in the banking industry using machine learning algorithms and develop a data visualization app for customer churn analysis. The study aims to predict which customers are most likely to discontinue using the bank's services. By applying various machine learning models, the researchers compare performance metrics to identify the best predictive approaches. The study also introduces a Data Visualization RShiny app, designed to assist in data science and management by providing insights into customer attrition trends, helping banks retain at-risk customers and enhance customer loyalty.

Rao et al. (2024) [14] introduce IADASYN-FLCatBoost, a new method for handling imbalanced customer churn classification in the banking sector. The method combines an enhanced

UNDER PEER REVIEW

Adaptive Synthetic Sampling (IADASYN) with the Local Outlier Factor (LOF) algorithm to improve data preprocessing and uses a Focal Loss-CatBoost (FLCatBoost) model for better classification. Empirical tests on a Kaggle credit card customer dataset show that IADASYN-FLCatBoost outperforms five other imbalanced classification methods, improving key metrics like Recall, F1 score, G-mean, and AUPRC. The model also demonstrates strong generalizability across various datasets, making it applicable to different industries. **Alizadeh et al. (2023) [15]** develop a customer churn model for the banking industry by integrating both hard and soft data. Hard data, such as transaction records and sensor data, is analyzed using a supervised machine learning approach, specifically a decision tree (DT) model, along with change mining and K-means clustering for data preprocessing. Soft data, which includes interpretative information, is modeled using the Dempster-Shafer theory. The fusion of these data types enhances the accuracy of churn predictions and customer behavior analysis. The results suggest that this approach can significantly improve customer relationship management systems in the banking sector.

Pathak et al. (2024) [16] address the challenge of customer attrition in the banking sector by proposing a dual-phase solution combining churn prediction and personalized recommendations. In the first phase, the authors focus on leveraging historical data to predict churn, enabling banks to implement targeted strategies aimed at reducing customer loss. In the second phase, they utilize Explainable AI (XAI) to offer personalized recommendations tailored to individual customer profiles, ensuring transparency in decision-making. This approach enhances the customer experience and boosts retention, emphasizing the need for customer-centric strategies in a competitive and evolving banking landscape. **Brito et al. (2024) [17]** present a framework to improve churn prediction in retail banking by focusing on data preparation. They use feature engineering based on recency, frequency, and monetary value (RFM), along with oversampling (ADASYN) and undersampling (NEASMISS) to address class imbalance. The XGBoost and Elastic Net models, tested on a large dataset, outperformed other methods in terms of accuracy, precision-recall, and specificity. This framework offers valuable insights for predicting churn and improving customer retention strategies in banking. **Murindanyi et al. (2023) [18]** address the issue of customer churn in retail banking by applying interpretable machine learning (ML) models. They use the Berka database from a Czech bank and Kaggle datasets for feature extraction and churn prediction. Synthetic Minority Over Sampling Techniques (SMOTE) handle class imbalance before training models. The Random Forest algorithm achieved outstanding results, with 99% accuracy and 98.5% recall on the Berka dataset, and 85% accuracy on the Kaggle dataset. For model accountability, they apply Model-Agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP), making the system transparent and trustworthy for the financial sector.

Haddadi et al. (2024) [19] tackle the challenge of customer churn prediction in highly imbalanced datasets, which is common in subscription-based services. The study compares 14 classification methods on three public imbalanced datasets from telecommunications, online retail,

UNDER PEER REVIEW

and banking sectors. It highlights the use of resampling techniques, including SMOTE and ADASYN, alongside a novel two-phase resampling method combining clustering and ensemble techniques with Long Short-Term Memory (LSTM) networks. The results show that the integrated approach consistently outperforms standalone methods, especially in terms of the Area Under the Curve (AUC), offering improved prediction accuracy and addressing class imbalance effectively. **Owolabi et al. (2024) [20]** explore the challenge of customer churn in the U.S. banking and financial services sector, highlighting its impact on profitability and market share. The study compares the performance of logistic regression, random forest, and neural networks for churn prediction, using industry-specific datasets. It incorporates macroeconomic factors to account for external influences on churn behavior and identifies key patterns such as age distribution and dormant accounts. The findings emphasize the importance of using advanced machine learning techniques and comprehensive customer data to prevent churn. The study provides valuable insights for financial institutions to proactively retain at-risk customers and optimize resources, offering a roadmap for future research and practical applications in churn prediction.

Materials and Methods

3.1 Data Collection and Explanation

This study is based on data collected from 341 interviewees, who provided candid assessments regarding their likelihood of choosing a public or private bank. The primary objective of the research is to examine the factors influencing this decision, considering several independent variables. These variables include Gender, which examines how an individual's gender influences their banking choice; Employment, which explores the impact of employment status on this decision; and Technology, which evaluates the role of technology adoption in the selection of a bank. Additionally, the study looks at Interest Rates, which refer to how the rates offered by a bank may affect customer preferences. A key variable in the study is Value Added Services (VAS), referring to supplementary offerings such as financial planning, investment advice, or other benefits provided by the bank. Another important variable is Reputation, which encompasses the bank's overall public perception, customer satisfaction, and service quality. Perceived Costs (PC) are also examined, representing the costs associated with using the bank, such as service fees, account maintenance charges, and other related expenses. The study also considers Attractiveness, referring to the overall appeal or attractiveness of the bank's offerings and brand image. Lastly, the study considers Perceived Risk (PR), which refers to the perceived security risks, including the safety of deposits, the potential for fraud, and other related concerns. This comprehensive set of independent variables allows for a deeper understanding of the factors that drive customers' banking preferences. The dataset used in this study is available for download at bank_prediction_dataset

UNDER PEER REVIEW

3.2 The Model

In this model, the dependent variable, "Preferred Choice of Bank," was coded as 1 for Private Bank and 0 for Public Bank. The Public Bank group was treated as the reference category, while the Private Bank group served as the target category. This coding scheme allows us to model the probability of a customer preferring a Private Bank over a Public Bank based on the predictors in the analysis. The logistic regression model for this binary outcome is specified as:

$$\text{logit(p)} = \ln\left(\frac{\text{P}}{1-\text{P}}\right) \qquad (3-1)$$

$$= \boldsymbol{B_0} + \boldsymbol{B_1} Value\ Added\ Services + \boldsymbol{B_2} Perceived\ Risk + \boldsymbol{B_3} Reputation + \boldsymbol{B_4} Perceived\ Costs$$

$$\pi(x_i) = \frac{e^{\boldsymbol{B_0}+\boldsymbol{B_1} Value\ Added\ Services+\boldsymbol{B_2} Perceived\ Risk+\boldsymbol{B_3} Reputation+\boldsymbol{B_4} Perceived\ Costs}}{1+e^{\boldsymbol{B_0}+\boldsymbol{B_1} Value\ Added\ Services+\boldsymbol{B_2} Perceived\ Risk+\boldsymbol{B_3} Reputation+\boldsymbol{B_4} Perceived\ Costs}}$$

NB: for simplicity, the equation has statistically significant predictors only

Where logit(p) is our dependent variable terminate, B_0 is a constant term and I'ts calculated value from variables in the equation table is **.612,** B_1 is the coefficient of the predictor ***Value Added Services*** and its calculated value from variables in the equation table is **.318**, B_2 is the coefficient of the predictor ***Perceived Risk*** and its calculated value from variables in the equation table is **-.512**, B_3 is the coefficient of the predictor ***Reputation*** and its calculated value from variables in the equation table is **.221** and B_4 is the coefficient of the predictor ***Perceived Costs*** and its calculated value from variables in the equation table is **-.245**.

3.3 Assumption Checking

Before running a binary logistic regression model, certain assumptions need to be satisfied by the data. If these assumptions are met, we can proceed with analyzing the data using the binary logistic regression model. These assumptions include:

a) *The dependent/response variable is binary or* dichotomous.

UNDER PEER REVIEW

Table .1: Dependent Variable Encoding

Original Value	Internal Value
Public	0
Private	1

Logistic regression assumes that the response variable only takes on two possible outcomes. From **Table 1**, we can clearly see that we have two outcomes coded 0 and 1 for '*Public*' and '*Private*' respectfully as our only two outcomes, therefore our model meets this assumption.

b) *The Observations are Independent*
According to this assumption, the observations in the dataset must be unrelated and independent of each other. This means that they should not be correlated or arise from multiple measurements of the same entity. In our specific data, each client's observations are completely independent of one another.

c) *Little or no multicollinearity between the predictor/explanatory* variables.

Table .2: Collinearity Statistics and Coefficients[a]

Model	Collinearity Statistics	
	Tolerance	VIF
Gender	.964	1.038
Employment	.982	1.018
Technology	.546	1.830
Interest Rates	.536	1.867
Value Added Services	.636	1.571
Perceived Risk	.669	1.494
Reputation	.660	1.515
Attractiveness	.673	1.485
Perceived Costs	.685	1.459

a. Dependent Variable: Preferred Choice of Bank

In binary logistic regression, multicollinearity refers to the presence of high correlations between predictor/explanatory variables. "Little or no multicollinearity" means that the variables included in the regression model are not highly correlated with each other. **Table 2** above, shows the Collinearity Statistics and since all our tolerance values are greater than **0.1** meaning our assumption is not violated and also, we can check the VIF values and see that all our predictor values are less than **10** confirming that there is no multicollinearity in our dataset. Multicollinearity can cause problems in binary logistic regression, such as unstable or unreliable estimates of regression coefficients, inflated standard errors, and difficulties in interpreting the coefficients. Therefore, it is important to check for multicollinearity before fitting a binary logistic regression model.

d) *The* sample size is sufficiently large (at least 10-20 observations per independent variable).

Table 3: Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	341	100.0
	Missing Cases	0	.0
	Total	341	100.0
Unselected Cases		0	.0
Total		341	100.0

UNDER PEER REVIEW

Logistic regression assumes that the sample size of the dataset if large enough to draw valid conclusions from the fitted logistic regression model. As a rule of thumb, you should have a minimum of 10 cases with the least frequent outcome for each explanatory variable. Table .3 above shows the total sample size of 341 customers which is a reasonably good sample size number.

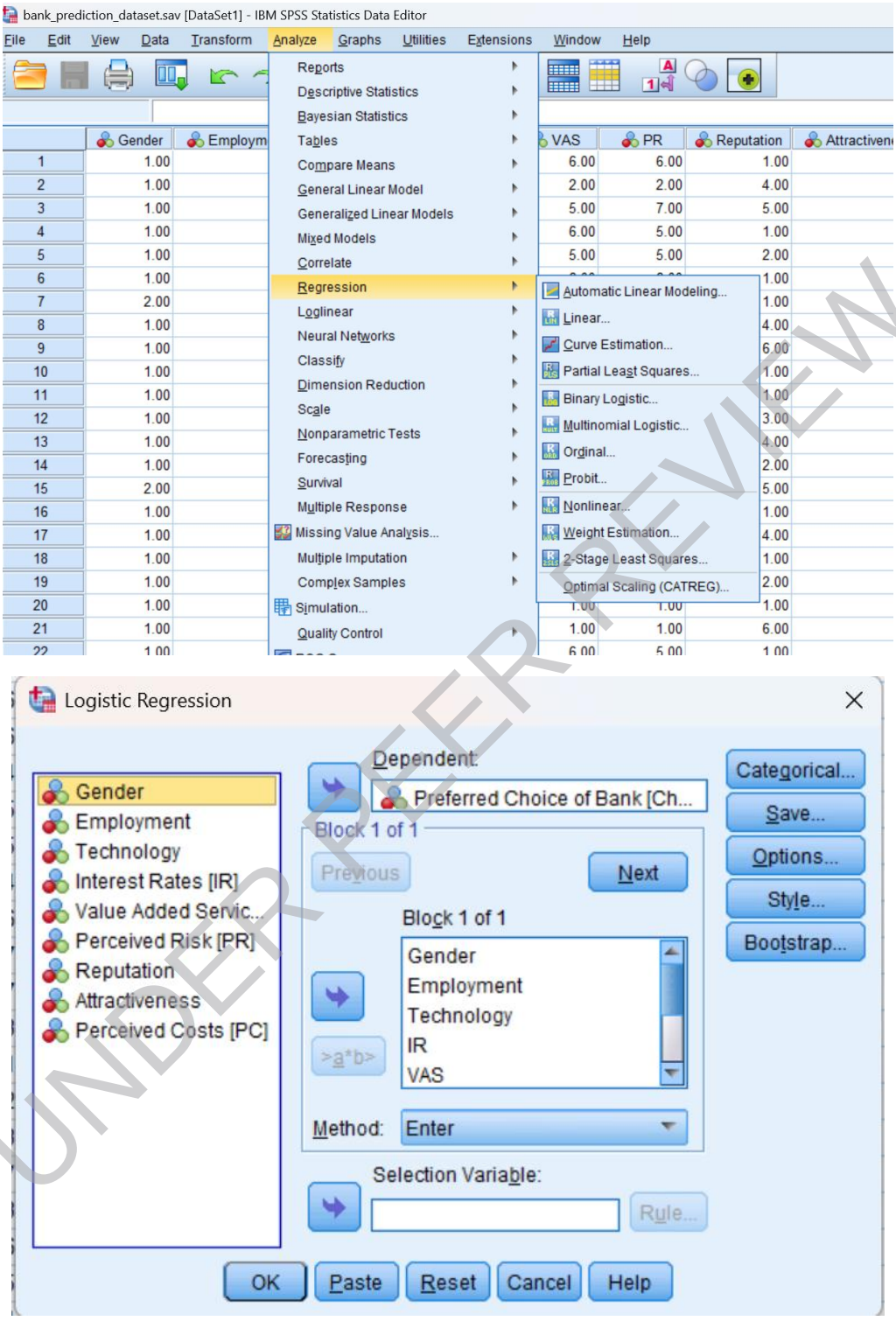
e) *There are no outliers*

The logistic regression model operates under the assumption that the dataset does not contain any significant outliers or influential observations. By employing the Mahalanobis distance method to detect outliers, we observe that our data set lacks any outliers, as indicated by the minimum value being greater than 0.001. Thus, the assumption of the absence of outliers has been satisfied.

Having successfully tested and confirmed that our data fulfills all the necessary assumptions for applying the binary logistic regression model, we are now ready to commence our analysis.

UNDER PEER REVIEW

Evaluation



UNDER PEER REVIEW

Figure 1:
Logistic regression window

The diagram above (see Figure 1) shows the steps on how to execute the binary logistic regression model using SPSS.

Block 0: Beginning Block

Table 4: Classification Table[a,b] (Baseline)

	Observed		Predicted Preferred Choice of Bank		Percentage Correct
			Public	Private	
Step 0	Preferred Choice of Bank	Public	0	90	.0
		Private	0	251	100.0
	Overall Percentage				73.6

Table 5: Variables in the Equation (Baseline)

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.026	.123	69.687	1	.000	2.789

The subsequent part of the output, labeled as Block 0 in Table 4 and Table 5, represents the outcome of the analysis conducted without incorporating any of our independent variables into the model. Consequently, this will serve as a reference point for comparing the model with our predictor variable included. The output is organized into blocks, with Block 0 displaying the outcomes of a null model (i.e., a model with only an intercept term).

Block 1: Method = Enter

Table 6: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	40.772	9	.000
	Block	40.772	9	.000
	Model	40.772	9	.000

Table 7: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	352.829[a]	.113	.165

Following Block 0, we examine Block 1 in the output, which is crucial for interpreting the results, as it reflects the regression model that includes the predictors. The Omnibus Tests of Model Coefficients (see Table 6) provide the results of the likelihood ratio chi-square tests. These tests assess whether the inclusion of the full set of predictors leads to a significant improvement in model fit over the null (intercept-only) model. Essentially, this can be viewed as an omnibus test of the null hypothesis, which posits that the regression coefficients for all predictors in the model are equal to zero (Pituch & Stevens, 2016).The results shown here indicate that the model fits the data significantly better than a null model, $\chi^2(4)=40.772$, $p<.001$.

UNDER PEER REVIEW

The summary of the model presented in Table 7 includes several key metrics: the -2 log likelihood and two 'pseudo-R-square' indices. Although pseudo-R-squares are conceptually similar to the R-square used in OLS regression, they are calculated differently. These indices offer descriptive information and are useful for assessing the overall adequacy of the model. The -2 Log Likelihood, also referred to as the model deviance, is a critical indicator of model fit. A value closer to zero suggests a better fit, reflecting a smaller discrepancy between the model's predictions and the observed data. In contrast, higher values indicate a poorer fit, signaling a greater difference between the model and the data. This discrepancy in fit is observed as a difference between the conditional probabilities of group membership predicted by the model and the actual observed group membership. As mentioned earlier, the likelihood ratio (LR) chi-square value in Table 6 represents the difference in deviances (i.e., -2LL) between the full model, which includes all predictors, and the reduced model, which only contains the intercept. We can calculate the deviance for the intercept-only model using the LR chi-square and the deviance of the full model as follows:

$$Deviance(nullmodel) = 352.829 + 40.772 = 393.601 \qquad (4-1)$$

The Cox & Snell and Nagelkerke R-squares are referred to as "pseudo-R-square" values because they are computed differently from the traditional R-square in OLS regression. In OLS regression, R-square is interpreted as the proportion of variation in the dependent variable (DV) that is explained by the predictors. However, the pseudo-R-square values in logistic regression represent the proportionate improvement or change in model fit relative to the intercept-only model.

The Cox & Snell pseudo-R-square, as shown in Table 8, is computed based on the likelihood ratio, with its upper bound typically being less than 1. This stands in contrast to R-square in OLS regression, which can theoretically approach 1, indicating perfect fit. The Cox & Snell pseudo-R-square is used to gauge how much better the model fits the data compared to a baseline model that only includes the intercept. However, due to its upper limit being less than 1, it is generally interpreted as indicating the relative improvement in model fit, rather than providing a direct measure of variance explained as with R-square in OLS.

Table 8: Model Summary on Cox & Snell R Square Calculation

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	352.829[a]	.113	.165

$$CS = 1 - exp\left(\frac{deviance_{full} - deviance_{null}}{n}\right)$$

$$= 1 - exp\left(\frac{352.829 - 393.601}{341}\right) = 1 - e^{-.1051} = .113 \qquad (4-2)$$

The "exp" in the first two expressions above corresponds to the "e" in the third expression,

UNDER PEER REVIEW

and this notation was used to improve the clarity of the equation. Nagelkerke introduced an adjustment to the Cox & Snell R-square, resulting in an index that ranges from 0 to 1. The following is the calculation for the Nagelkerke pseudo-R-square.

Table 9: Model Summary on Nagelkerke pseudo-R-square

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	352.829[a]	.113	.165

$$Nagelkerke = \frac{CS}{1 - exp\left(-\frac{deviance_{null}}{n}\right)} = \frac{.113}{1 - exp\left(-\frac{393.601}{341}\right)} = .165 \quad (4-3)$$

These versions of the Cox & Snell and Nagelkerke pseudo-R-squares are based on the formulations provided by Field (2018) [21]. The -2*Log likelihood (also referred to as "model deviance" is most useful for comparing competing models, particularly because it is distributed as chi-square.

Table 10: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.691	8	.790

The results from the Hosmer and Lemeshow test, as shown in Table 10, indicate that the model fits the data well. The Chi-square value is 4.691 with 8 degrees of freedom, and the p-value (significance) is 0.790, which is much higher than the conventional threshold of 0.05. This high p-value suggests that there is no significant difference between the observed and expected frequencies in the model's subgroups. As a result, we fail to reject the null hypothesis, meaning that the model's predictions align well with the observed data. In conclusion, the logistic regression model provides a good fit to the data, and its predictions can be considered reliable.

Table 11: Classification Table (Full Model (or Predictive Model))

Observed			Predicted Preferred Choice of Bank		Percentage Correct
			Public	Private	
Step 1	Preferred Choice of Bank	Public	16	74	17.8
		Private	8	243	96.8
	Overall Percentage				76.0

The classification table presented in Table 4 provides an overview of the performance of the full model (or predictive model) in predicting the preferred choice of bank (Public or Private). The

UNDER PEER REVIEW

table compares the observed outcomes (actual choices of the bank) with the predicted outcomes (based on the model), showing the number of correct and incorrect predictions for each category.

- **Public Bank:** Out of the 24 customers who actually preferred a Public Bank, the model correctly predicted 16 (a 17.8% accuracy rate). However, 74 customers who were predicted to prefer a Private Bank were actually Public Bank customers, indicating a significant misclassification rate for Public Bank customers.
- **Private Bank:** The model performed well in predicting the preference for Private Bank customers. Of the 251 customers who actually preferred a Private Bank, the model correctly predicted 243, leading to a high accuracy rate of 96.8%.
- **Overall Accuracy:** The model achieved an overall accuracy of 76%, meaning that, across both categories, it correctly predicted the choice of bank for 76% of the cases. This reflects that the model is generally effective, but it shows a tendency to misclassify Public Bank customers more frequently.

Table 12: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Gender	.079	.385	.043	1	.837	1.083	.509	2.300
	Employment	.309	.367	.708	1	.400	1.362	.663	2.799
	Technology	.137	.116	1.382	1	.240	1.147	.913	1.440
	Interest Rates	.086	.111	.595	1	.440	1.090	.876	1.355
	Value Added Services	.318	.113	7.981	1	.005	1.374	1.102	1.714
	Perceived Risk	-.512	.107	22.994	1	.000	.600	.486	.739
	Reputation	.221	.106	4.328	1	.037	1.248	1.013	1.538
	Attractiveness	-.059	.101	.336	1	.562	.943	.773	1.150
	Perceived Costs	-.245	.092	7.082	1	.008	.782	.653	.937
	Constant	.612	1.154	.281	1	.596	1.844		

The 'Estimate' column in the table presents the regression coefficients, which represent the predicted change in the log odds of being classified into the target group, relative to the reference group, for each one-unit increase in the respective predictor variable. These predictions are made while accounting for the influence of the other predictors in the model. In other words, each regression coefficient indicates the effect of the corresponding predictor on the log odds of the outcome variable. It is important to note that a common misconception is that the regression coefficient directly represents the predicted change in the probability of membership in the target group (i.e., p(Y=1|X's)) per unit increase in the predictor. This interpretation is incorrect. The

UNDER PEER REVIEW

coefficient reflects the change in the log odds, not the probability, for each unit increase in the predictor.

In general, a positive regression coefficient can be interpreted as indicating that the likelihood (loosely speaking) of belonging to the target group increases with an increase in the predictor variable. Conversely, a negative coefficient suggests that the likelihood of being in the target group decreases as the predictor variable increases. If the regression coefficient equals zero, it suggests that changes in the predictor variable have no effect on the likelihood of being in the target group. However, it's important to remember that these interpretations refer to changes in the log odds, not the probability itself.

The Odds Ratio (OR) column represents the multiplicative change in the odds of belonging to the target group for every one-unit increase in a predictor variable. Generally, an odds ratio (OR) greater than 1 indicates that as the predictor variable increases, the odds of being in the target group also increase, meaning a higher likelihood of the outcome. Conversely, an odds ratio (OR) less than 1 suggests that as the predictor increases, the odds of being in the target group decrease, reflecting a lower likelihood of the outcome. If the OR equals 1, it indicates that changes in the predictor have no effect on the odds of being in the target group, meaning no change in the likelihood of the outcome. The 95% confidence interval for the Odds Ratio (OR) provides a range of values within which the true OR is likely to fall with 95% confidence. This interval can also be used to assess whether the observed OR is significantly different from the null hypothesis value of 1.0. If the value of 1.0 lies within the lower and upper bounds of the confidence interval, it suggests that the computed OR is not significantly different from 1.0. This implies that changes in the predictor do not have a statistically significant effect on the odds of being in the target group. Conversely, if the confidence interval does not include 1.0, it indicates that the observed OR is significantly different from 1.0, suggesting a meaningful relationship between the predictor and the outcome.

Table 13: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Gender	.079	.385	.043	1	.837	1.083	.509	2.300
	Employment	.309	.367	.708	1	.400	1.362	.663	2.799
	Technology	.137	.116	1.382	1	.240	1.147	.913	1.440
	Interest Rates	.086	.111	.595	1	.440	1.090	.876	1.355
	Value Added Services	.318	.113	7.981	1	.005	1.374	1.102	1.714
	Perceived Risk	-.512	.107	22.994	1	.000	.600	.486	.739
	Reputation	.221	.106	4.328	1	.037	1.248	1.013	1.538
	Attractiveness	-.059	.101	.336	1	.562	.943	.773	1.150
	Perceived Costs	-.245	.092	7.082	1	.008	.782	.653	.937

UNDER PEER REVIEW

Constant	.612	1.154	.281	1	.596	1.844

In this analysis, we examine the regression results for several predictor variables and their relationship with the likelihood of selecting a Private Bank (coded as 1) versus a Public Bank (coded as 0). Below, we interpret the key findings from the table:

1. Gender (B = 0.079, p = 0.837, Exp(B) = 1.083)

The coefficient for Gender is positive, but the p-value of 0.837 is greater than the significance threshold (0.05), suggesting that Gender is not a statistically significant predictor of the choice of bank. The odds ratio (Exp(B) = 1.083) suggests a small increase in the odds of choosing a Private Bank for one unit increase in the Gender variable but this effect is not statistically significant. Thus, we conclude that Gender does not have a meaningful impact on the bank choice in this model.

2. Employment (B = 0.309, p = 0.400, Exp(B) = 1.362)

The Employment variable has a positive coefficient, indicating that being employed increases the odds of choosing a Private Bank, but the result is not statistically significant (p = 0.400). Therefore, while the model suggests a potential relationship, Employment does not significantly influence the choice of bank in this dataset.

3. Technology (B = 0.137, p = 0.240, Exp(B) = 1.147)

The positive coefficient for Technology suggests that greater involvement with technology is associated with an increased likelihood of choosing a Private Bank. However, the p-value of 0.240 indicates that this relationship is not statistically significant. The odds ratio of 1.147 indicates a modest increase in the odds of choosing a Private Bank as technology usage increases, but this effect is not substantial enough to be conclusive.

4. Interest Rates (B = 0.086, p = 0.440, Exp(B) = 1.090)

Similarly, Interest Rates have a positive coefficient, suggesting that higher interest rates might slightly increase the likelihood of choosing a Private Bank. However, with a p-value of 0.440, this result is not statistically significant. The odds ratio of 1.090 also indicates a minor increase in the odds of choosing a Private Bank, but again, the relationship is not significant.

5. Value Added Services (B = 0.318, p = 0.005, Exp(B) = 1.374)

The coefficient for Value Added Services is positive and statistically significant (p = 0.005). This indicates that for every one-unit increase in the provision of value-added services, the odds of selecting a Private Bank increase by 37.4%. This suggests that the availability of additional services, such as financial planning or investment advice, significantly influences customers' choices toward Private Banks.

UNDER PEER REVIEW

6. **Perceived Risk (B = -0.512, p = 0.000, Exp(B) = 0.600)**

Perceived Risk has a negative coefficient and a highly significant p-value (p = 0.000), indicating that as perceived risk increases, the odds of choosing a Private Bank decrease. Specifically, for every one-unit increase in perceived risk, the odds of choosing a Private Bank decrease by 40% (Exp(B) = 0.600). This suggests that customers who perceive higher risks associated with banking may be more likely to choose Public Banks.

7. **Reputation (B = 0.221, p = 0.037, Exp(B) = 1.248)**

The Reputation variable has a positive coefficient and is statistically significant (p = 0.037). This implies that a better reputation of the bank increases the likelihood of selecting a Private Bank. The odds ratio of 1.248 suggests that for each unit increase in the perceived reputation, the odds of choosing a Private Bank increase by 24.8%. Therefore, reputation is an important factor in influencing customers' banking choices.

8. **Attractiveness (B = -0.059, p = 0.562, Exp(B) = 0.943)**

The coefficient for Attractiveness is negative, suggesting that greater attractiveness might decrease the likelihood of choosing a Private Bank. However, the p-value of 0.562 indicates that this result is not statistically significant. The odds ratio of 0.943 also indicates a very slight decrease in the likelihood of choosing a Private Bank as attractiveness increases, but the effect is negligible and non-significant.

9. **Perceived Costs (B = -0.245, p = 0.008, Exp(B) = 0.782)**

Perceived Costs has a negative and statistically significant coefficient (p = 0.008). This suggests that as perceived costs associated with a Private Bank increase, the odds of choosing a Private Bank decrease. Specifically, for every unit increase in perceived costs, the odds of choosing a Private Bank decrease by 21.8% (Exp(B) = 0.782). This finding underscores the importance of cost considerations in customers' banking decisions.

10. **Constant (B = 0.612, p = 0.596)**

The constant term, or intercept, represents the predicted log-odds of selecting a Private Bank when all predictor variables are zero. With a p-value of 0.596, this value is not statistically significant and does not contribute meaningfully to explaining the bank choice when the other variables are included in the model

Conclusion

Binary Logistic Regression was used to predict the choice of bank (Public or Private) based on independent variables that include Value Added Services, Perceived Risks, Reputation, and Perceived Costs. A preliminary analysis suggested that the assumption of multicollinearity was met

UNDER PEER REVIEW

(**tolerance: p>0.1**). The model was statistically significant**, x^2 (4, 9=341) =40.772, p<.001** indicating that the model fits the data significantly better than a null model and can model the likelihood of choosing Private bank correctly. The model explained between **11.2%** (Cox and Snell R square) and **16.5 %** Nagelkerke R square) of the variance in the dependent variable and correctly classified **76%** of the cases.

In conclusion, the analysis reveals several key factors that significantly influence the likelihood of choosing a Private Bank over a Public Bank. Value Added Services and Reputation positively impact the probability of selecting a Private Bank, highlighting the importance of additional services and a strong public image in attracting customers. Conversely, Perceived Risk and Perceived Costs are significant deterrents, with higher perceived risks and costs decreasing the likelihood of choosing a Private Bank. Notably, other factors such as Gender, Employment, Technology, Interest Rates, and Attractiveness did not show statistically significant effects, suggesting that they may not be as influential in customers' decisions between public and private banking. Overall, the findings suggest that private banks may benefit from emphasizing value-added services and cultivating a positive reputation while addressing concerns related to risk and cost to improve customer preference.

References

[1] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons.

[2] Allison, P. D. (2012). Logistic regression using SAS: theory and application. SAS Institute.

[3] Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.

[4] Pituch, K.A., & Stevens, J.A. (2016). Applied multivariate statistics for the social sciences (6th ed). *New York*: Routledge.

[5] Hosmer, D.W., Hosmer, T. and Lemeshow, S. (1980) A Goodness-of-Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics*, 10, 1043-1069. https://doi.org/10.1080/03610928008827941

[6] Nagelkerke, N.J.D. (1991) A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78, 691-692. https://doi.org/10.1093/biomet/78.3.691

[7] Cox, D. R., & Snell, E. J. (1989). Analysis of binary data (2nd ed.). Chapman and Hall.

[8] McFadden, D. (1974) Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P., Ed., Economic Theory and Mathematical Economics, *Academic Press, New York, NY*, 105-142.

[9] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

UNDER PEER REVIEW

[10] Zaki, A. M., Khodadadi, N., Lim, W. H., & Towfek, S. K. (2024). Predictive analytics and machine learning in direct marketing for anticipating bank term deposit subscriptions. *American Journal of Business and Operations Research*, 11(1), 79-88.

[11] Rahman, M. K., Hoque, M. N., Yusuf, S. N. S., Yusoff, M. N. H. B., & Begum, F. (2023). Do customers' perceptions of Islamic banking services predict satisfaction and word of mouth? Evidence from Islamic banks in Bangladesh. *PLoS One*, 18(1), e0280108.

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

[13] Singh, P. P., Anik, F. I., Senapati, R., Sinha, A., Sakib, N., & 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), 7-16.

[14] Rao, C., Xu, Y., Xiao, X., Hu, F., & Goh, M. (2024). Imbalanced customer churn classification using a new multi-strategy collaborative processing method. *Expert Systems with Applications*, 247, 123251.

[15] Alizadeh, M., Zadeh, D. S., Moshiri, B., & Montazeri, A. (2023). Development of a customer churn model for banking industry based on hard and soft data fusion. *IEEE Access*, 11, 29759-29768.

[16] Pathak, P., Chandgadkar, V., Solanki, A., Shrivastava, A., Pulgam, N., & Maktum, T. (2024, March). Customer Churn Prediction and Personalised Recommendations in Banking. *In International Conference on Artificial Intelligence and Smart Energy* (pp. 409-421). Cham: Springer Nature Switzerland.

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

[18] Murindanyi, S., Mugalu, B. W., Nakatumba-Nabende, J., & 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.

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

[20] Owolabi, O. S., Uche, P. C., Adeniken, N. T., Efijemue, O., Attakorah, S., Emi-Johnson, O. G., & Hinneh, E. (2024). Comparative Analysis of Machine Learning Models for Customer Churn Prediction in the US Banking and Financial Services: Economic Impact and Industry-Specific Insights. *Journal of Data Analysis and Information Processing*, 12(3), 388-418.

[21] Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed). Los Angeles: Sage.

UNDER PEER REVIEW