Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2024, 12, 388-418
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Comparative Analysis of Machine Learning Models for Customer Churn Prediction in the U.S. Banking and Financial Services: Economic Impact and Industry-Specific Insights

Omoshola S. Owolabi[1], Prince C. Uche[1], Nathaniel T. Adeniken[1], Oghenekome Efijemue[2], Samuel Attakorah[1], Oluwabukola G. Emi-Johnson[3], Emmanuel Hinneh[1]

[1]Department of Data Science, Carolina University, Winston-Salem, NC, USA
[2]Department of Computer Science, Austin Peay State University, Clarkesville, TN, USA
[3]Department of Statistical Sciences, Wake Forest University, Winston-Salem, NC, USA
Email: owolabio@carolinau.edu, uchep@carolinau.edu, adenikenn@carolinau.edu, komeefi@gmail.com, attakorahs@carolinau.edu, hinnehe@carolinau.edu, emijo23@wfu.edu

How to cite this paper: Owolabi, O.S., Uche, P.C., Adeniken, N.T., Efijemue, O., Attakorah, S., Emi-Johnson, O.G. and Hinneh, E. (2024) Comparative Analysis of Machine Learning Models for Customer Churn Prediction in the U.S. Banking and Financial Services: Economic Impact and Industry-Specific Insights. *Journal of Data Analysis and Information Processing*, **12**, 388-418.
https://doi.org/10.4236/jdaip.2024.123021

Received: May 20, 2024
Accepted: July 14, 2024
Published: July 17, 2024

Abstract

Customer churn poses a significant challenge for the banking and finance industry in the United States, directly affecting profitability and market share. This study conducts a comprehensive comparative analysis of machine learning models for customer churn prediction, focusing on the U.S. context. The research evaluates the performance of logistic regression, random forest, and neural networks using industry-specific datasets, considering the economic impact and practical implications of the findings. The exploratory data analysis reveals unique patterns and trends in the U.S. banking and finance industry, such as the age distribution of customers and the prevalence of dormant accounts. The study incorporates macroeconomic factors to capture the potential influence of external conditions on customer churn behavior. The findings highlight the importance of leveraging advanced machine learning techniques and comprehensive customer data to develop effective churn prevention strategies in the U.S. context. By accurately predicting customer churn, financial institutions can proactively identify at-risk customers, implement targeted retention strategies, and optimize resource allocation. The study discusses the limitations and potential future improvements, serving as a roadmap for researchers and practitioners to further advance the field of customer churn prediction in the evolving landscape of the U.S. banking and finance industry.

Keywords

Churn, Prediction, Machine Learning, Economic Impact, Industry-Specific Insights, Logistic Regression, Random Forest, Neural Networks

1. Introduction

Customer churn, the phenomenon of customers terminating their relationship with a service provider, poses a significant challenge for the banking and finance industry in the United States [1]. The economic impact of customer churn on financial institutions is substantial, as it directly affects profitability and market share, which highlights the critical need for financial institutions to prioritize customer retention strategies.

The US banking and finance industry faces unique challenges and trends related to customer retention. Intensifying competition, evolving customer expectations, and the emergence of disruptive technologies have transformed the financial services landscape [2]. Customers now have access to a wide range of alternatives, making it easier to switch providers if their expectations are unmet. Moreover, the advent of digital banking has altered customer behavior and preferences, necessitating financial institutions to adapt their strategies accordingly [3].

Previous research has explored various aspects of customer churn in the banking and finance industry [4] investigated the factors influencing customer churn in the context of electronic banking services, highlighting the importance of service quality, customer satisfaction, and trust in reducing churn [5] compared the performance of logistic regression and decision trees in predicting credit card churn, emphasizing the need for accurate and interpretable models. However, these studies have primarily focused on specific segments or products within the banking and finance industry, rather than providing a comprehensive analysis of customer churn across the entire US market.

While machine learning techniques have been applied to churn prediction in various industries [6] [7], there is limited research on the comparative performance of different machine learning algorithms in the US banking and finance industry. Additionally, existing studies often overlook the economic impact and industry-specific implications of churn prediction, which are crucial for developing effective retention strategies and making informed business decisions.

1.1. The Potential of Machine Learning in Predicting Customer Churn

Machine learning, a branch of artificial intelligence, has emerged as a powerful tool for predicting customer churn in various industries, including banking and finance [8]. In the US context, machine learning techniques have shown promising results in identifying customers at risk of churning, enabling financial in-

stitutions to take proactive measures to retain them [7]. Machine learning models can uncover complex relationships and predict churn with high accuracy by leveraging vast amounts of customer data, such as demographics, transactional history, and behavioral patterns.

The potential of machine learning in predicting customer churn in the US banking and finance industry lies in its ability to process and analyze large volumes of data efficiently [9]. Traditional approaches to churn prediction often rely on manual analysis and simple statistical methods, which may not effectively capture the intricacies of customer behavior in the rapidly evolving US market [10]. Machine learning algorithms, on the other hand, can automatically learn from historical data, adapt to changing patterns, and provide actionable insights for targeted retention strategies [11].

1.2. Objectives of the Study

The study aims to compare machine learning models for customer churn prediction in the US banking and finance industry. Three widely used algorithms will be evaluated: logistic regression, random forest, and neural networks. The focus is on the economic impact and industry-specific implications of churn prediction. By analyzing customer attributes, behavioral patterns, and churn outcomes, the research aims to identify key drivers of customer attrition. This will help develop targeted retention strategies for US financial institutions. The study will analyze a comprehensive dataset that includes customer demographics, account information, transaction balance, engagement metrics, and economic indicators. The goal is to generate industry-specific insights and recommendations for decision-making and resource allocation. The findings can assist financial institutions in enhancing customer loyalty, improving financial performance, and staying competitive. Additionally, the research will contribute to a broader understanding of customer behavior and preferences in the US banking and finance industry, enabling practitioners to adapt and stay ahead of trends.

2. Literature Review

Customer churn, the phenomenon of customers discontinuing their relationship with a business, has been a persistent and significant challenge in the US banking and finance industry. Historical trends reveal the magnitude of the problem, with a study by Bain & Company (2018) [12] indicating that the average customer churn rate for US banks increased from 9.6% in 2012 to 11.4% in 2018. This alarming trend underscores the growing importance of effective customer retention strategies for financial institutions operating in the highly competitive US market.

2.1. Overview of Customer Churn

The current state of customer churn in the US banking and finance industry is characterized by a confluence of factors, including intense competition, rapid

technological disruption, and shifting customer preferences. The proliferation of digital banking services and the emergence of new players, such as fintech companies and neo banks, have lowered the barriers to switching for customers, making it easier for them to explore alternative financial service providers [13]. As a result, traditional banks and financial institutions face increased pressure to differentiate themselves through exceptional customer experiences and personalized offerings to mitigate churn [14]. Economic consequences of customer attrition for US financial institutions are substantial and far-reaching. The acquisition cost for a new customer can be five to twenty-five times higher than the cost of retaining an existing one [15], making customer retention a critical priority for financial institutions.

The impact of customer churn extends beyond immediate financial losses, as it can have a ripple effect on the overall financial health and competitiveness of institutions. Lost customers may result in reduced cross-selling opportunities, as satisfied customers are more likely to purchase additional products and services from their primary financial institution [16]. Also, high churn rates can diminish brand loyalty and reputation, as dissatisfied customers may share their negative experiences with others, deterring potential new customers from engaging with the institution [17].

The lifetime value of a customer is a crucial consideration in the context of customer churn. Retaining customers over the long term can lead to significant financial benefits for banks and financial institutions, as the revenue generated from a loyal customer can grow exponentially over time [18]. Studies have shown that a 5% increase in customer retention can lead to a 25% - 95% increase in profitability [19], highlighting the immense value of customer loyalty in the banking and finance industry. Given the critical importance of customer retention, US financial institutions have been investing heavily in strategies and technologies to predict and prevent churn. However, the dynamic nature of the industry, coupled with the ever-evolving customer expectations, makes it challenging for institutions to keep pace and effectively address the problem of customer attrition. This underscores the need for advanced analytics and data-driven approaches to gain a deeper understanding of customer behavior and develop targeted retention strategies.

2.2. Existing Approaches to Churn Prediction

Traditionally, churn prediction in the US banking and finance industry has relied on manual analysis and simple statistical methods. These approaches often involve segmenting customers based on demographic or behavioral characteristics and identifying patterns associated with churn. For example, institutions may analyze variables such as age, income, transaction frequency, and product usage to determine which customer segments are more likely to churn [4].

While these traditional methods can provide some insights into customer attrition, they have limitations in terms of scalability, accuracy, and the ability to capture complex relationships in large datasets. As customer data becomes in-

creasingly voluminous and diverse, encompassing structured and unstructured data from sources, traditional approaches may struggle to provide actionable insights for effective customer retention strategies. The traditional churn prediction methods often rely on historical data and may not be able to adapt quickly to changing customer behaviors and market conditions. In an industry where customer preferences and expectations are constantly evolving, the ability to leverage real-time data and generate proactive insights is critical for effective churn management [20].

To address these limitations, the application of machine learning techniques for churn prediction has gained prominence in the US banking and finance industry. Machine learning algorithms, such as logistic regression, decision trees, random forests, and neural networks, have been employed to analyze vast amounts of customer data and identify patterns indicative of churn [7]. Logistic regression, a widely used statistical model, has been applied to churn prediction in the banking and finance industry. This method estimates the probability of churn based on a linear combination of predictor variables, such as customer demographics, account information, and transaction history. Logistic regression models are interpretable and can provide insights into the relative importance of different factors in predicting churn. The effectiveness of machine learning models for churn prediction depends on the quality and relevance of the input data. Financial institutions must ensure that they have robust data collection and preprocessing mechanisms in place to handle missing values, outliers, and inconsistencies [6]. Additionally, the interpretability of machine learning models is an important consideration, as decision-makers need to understand the key factors driving churn predictions to design appropriate retention strategies [21].

2.3. Gaps in Current Research and the Need for Industry-Specific Insights

Despite the growing adoption of machine learning for churn prediction in the US banking and finance industry, there are still several gaps in current research that need to be addressed. Many existing studies focus on general churn prediction methods without considering the unique characteristics and challenges of the US market. The US banking and finance industry operates in a highly regulated and competitive environment, with distinct customer behaviors and preferences [22].

One of the key gaps in current research is the lack of industry-specific insights that consider the contextual factors influencing customer churn in the US. The drivers of churn may vary across different segments of the banking and finance industry, such as retail banking, credit unions, and wealth management. Therefore, it is crucial to conduct analyses with the intention to evaluate how macroeconomic factors affect these customer decisions to churn and develop tailored prediction models that capture the nuances of each industry segment. While previous research has compared the performance of different machine learning algorithms for churn prediction, there is limited work on the economic impact

and practical implications of these models in the US banking and finance industry. Understanding the financial consequences of customer churn and the potential return on investment of retention strategies is crucial for decision-makers in the industry.

Another gap in current research is the lack of focus on the interpretability and explainability of churn prediction models. While complex machine learning algorithms may achieve high predictive accuracy, they often operate as "black boxes", making it difficult for stakeholders to understand the underlying reasons for churn predictions. This lack of transparency can hinder the adoption and trust in these models, as decision-makers may be hesitant to rely on insights they cannot fully comprehend [1]. There is a need for research that translates the findings of machine learning models into actionable recommendations for customer retention, considering the specific challenges and constraints faced by US financial institutions. Churn prediction models should not only identify at-risk customers but also provide guidance on the most effective retention strategies for each customer segment [6]. This requires a deep understanding of customer preferences, behaviors, and the feasibility of different retention interventions within the organizational context [23].

In a bid to address these gaps, this study aims to provide a comprehensive comparative analysis of machine learning models for customer churn prediction, with a specific focus on the US banking and finance industry. By leveraging industry-specific datasets and considering the economic impact and practical implications of the findings, this research seeks to contribute valuable insights that can inform customer retention strategies and decision-making processes in the US context. The performance of logistic regression, random forests, and neural networks in predicting churn across different segments of the US banking and finance industry will be explored. By conducting a granular analysis, the research aims to uncover segment-specific churn drivers and develop tailored retention strategies. Also, the study will investigate the potential of integrating external data sources, such as macroeconomic indicators, to enhance the accuracy and robustness of churn prediction models.

For the interpretability and explain-ability challenges, the study will employ techniques such as feature importance analysis and model-agnostic explanations to provide insights into the key factors driving churn predictions. This will enable decision-makers to understand the rationale behind the model outputs and make informed decisions about retention interventions. The research will include case studies of successful churn prediction and retention initiatives implemented by US banks and financial institutions. By analyzing real-world examples, the study aims to provide practical guidance on the implementation challenges, best practices, and potential business impact of data-driven churn management strategies.

This literature review highlights the significance of customer churn in the US banking and finance industry and the need for advanced analytics and machine learning approaches to predict and mitigate churn effectively. By addressing the

gaps in current research and providing industry-specific insights, this study aims to contribute to the development of data-driven retention strategies that can help US financial institutions enhance customer loyalty, improve profitability, and maintain a competitive edge in an increasingly dynamic market.

3. Fundamental Principles

3.1. Univariate Analysis

Exploratory analysis of the data (EDA) was conducted to gain insight into the datasets and understand the relationship between variables. The EDA involved examining variable distributions, identifying patterns, and detecting outliers or anomalies. The dataset used for this analysis consists of 165,034 records and 14 features, with a mix of both categorical and numerical variables. The dataset contains various columns related to customer information, such as the age of the customer, the number of years the customer has been with the bank, the account balance, whether the customer has a card, or if the customer is active and if the customer churned or not.

In the univariate analysis, the distribution of the target variable was analyzed to understand the central tendency and the range of the target variable, providing insight into overall customer churn. The distribution of customer information is also examined to identify outliers and patterns that could influence the robustness of the bank churn model. The summary statistics were checked on the numerical variables, and it shows that the credit score of the customer ranges from 350 to 850, with a mean of 656.45, indicating moderate creditworthiness, with some level of variability because of the standard deviation rate of 80. Age ranges from 18 - 92 years with a mean of 38.13 years, suggesting a working-age customer base; the skewed distribution is positive, indicating a larger number of younger customers. These descriptive statistics computed are in a bid to provide a measure of central tendency and dispersion as above.

As shown in **Figure 1**, the light blue segment represents 78.8% of instances labeled as "0" (did not exit), while the dark blue segment represents 21.2% labeled as "1" (Exited). Most data points fall into the "did not exit" category, with the "exited" category being smaller.

The bar graph shows the count of instances for each class. It reinforces the observation that most data points belong to the "did not exit" class. The class imbalance may impact model performance and will be addressed with SMOTE at the model-building stage. SMOTE can improve classifier sensitivity for the minority class by balancing the classes. It enhances generalization by helping the model learn more general features of each class instead of overfitting to the majority class.

Figure 2 histograms provide insights into the distribution of each numerical variable in the dataset; credit scores appear normally distributed with a slight left skew and age shows a right-skewed distribution, indicating a larger proportion

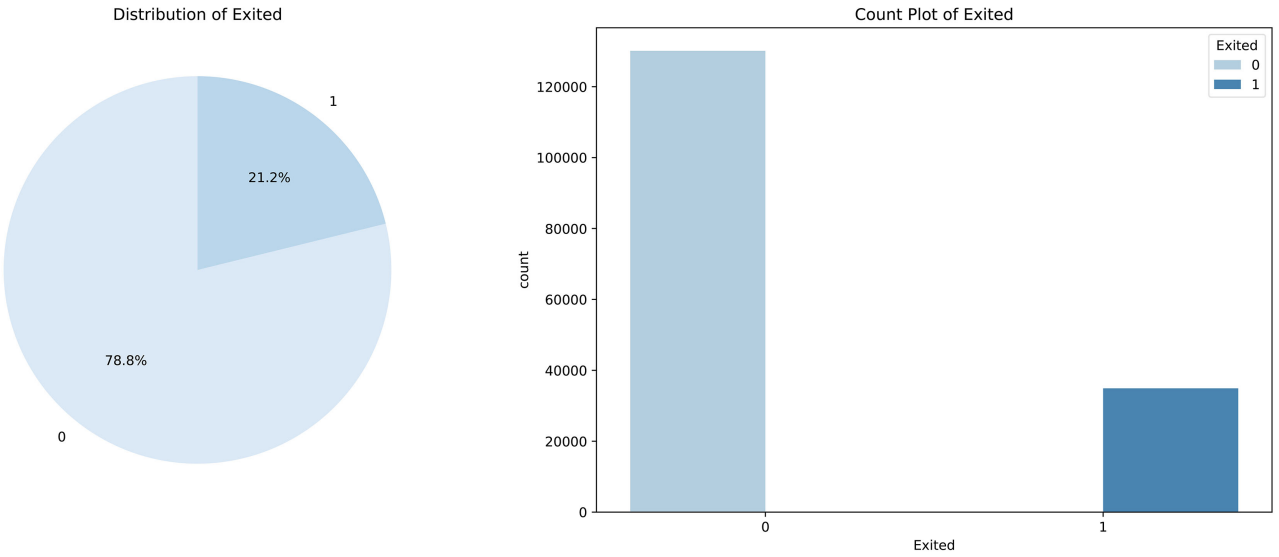


Figure 1. Distribution of the target variable.

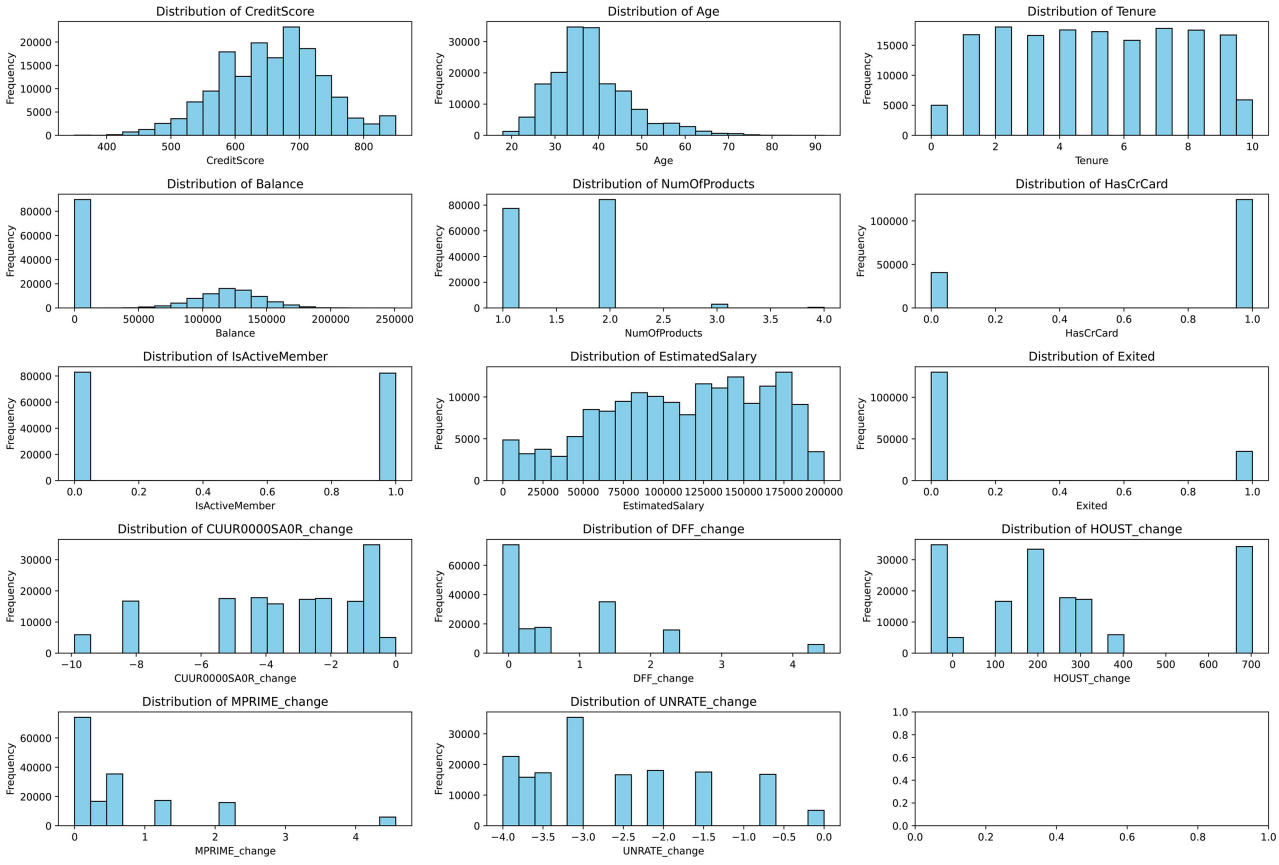


Figure 2. Distribution of numerical variables.

of younger customers. Tenure has a uniform distribution, with slight decreases at the lowest and highest tenure, and balance has a significant peak at zero balance, suggesting many customers have no balance, followed by a normal distribution for positive balances. Most customers have 1 or 2 products, with few

having 3 or 4, while the estimated salary is uniformly distributed across different salary ranges.

3.2. Bivariate Analysis

In the bivariate analysis, the relationship between customer information and the key predictors was explored to identify potential trends or correlations. This analysis aimed to determine if the simplicity or complexity of customer information and behaviors and activities had a significant impact on the probability of exiting financial services.

Figure 3 is a distribution of credit scores and age shows that higher credit scores and younger age are associated with customers staying with the bank. This suggests that creditworthiness and age play a role in customer retention. On the other hand, older customers are more likely to exit the bank, indicating age as a significant factor in customer churn.

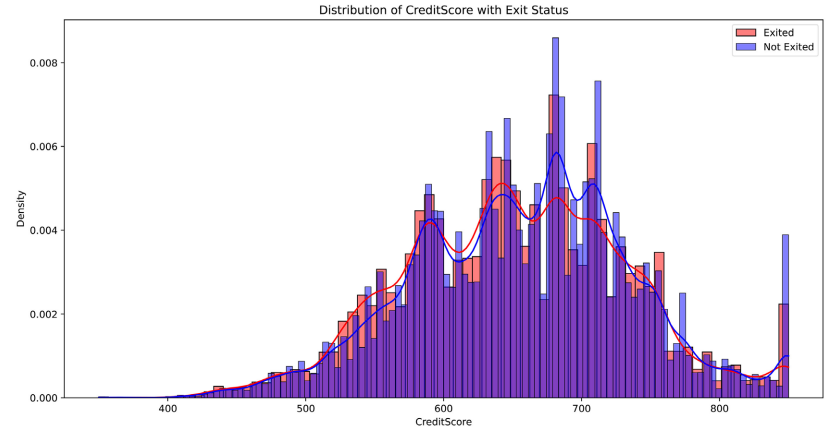
However, there is no clear pattern indicating that tenure significantly impacts the likelihood of exiting or staying, while customers with higher balances are slightly more likely to exit. This suggests that financial stability alone does not guarantee customer loyalty. Also, customers using around two products show a peak in staying, but there is a significant spike in exits for customers using three products or more. Lastly, salary does not a pear to be a significant factor in customer churn. Understanding these patterns can help the bank tailor strategies to reduce churn and retain valuable customers.

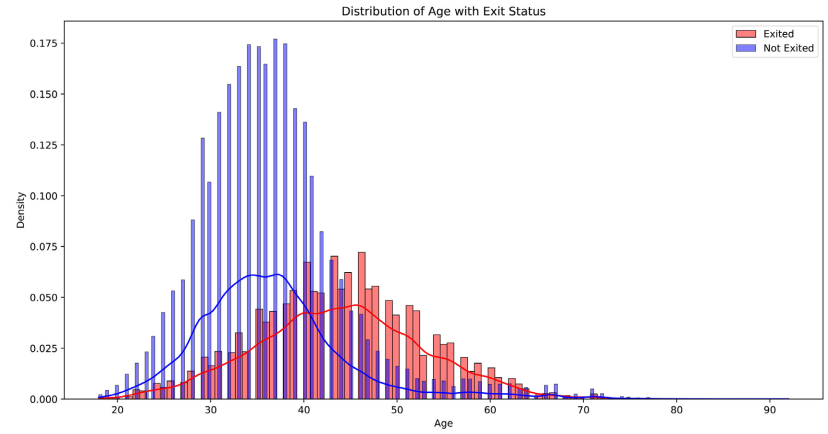
The pie charts above in Figure 4 provide insights into the distribution of the categorical variables. In the pie charts, 24.6% don't have a card while 75.4% do. For membership, 49.8% are inactive and 50.2% are active. This shows balanced engagement with a slight preference for active participation, providing insight into card ownership and membership.

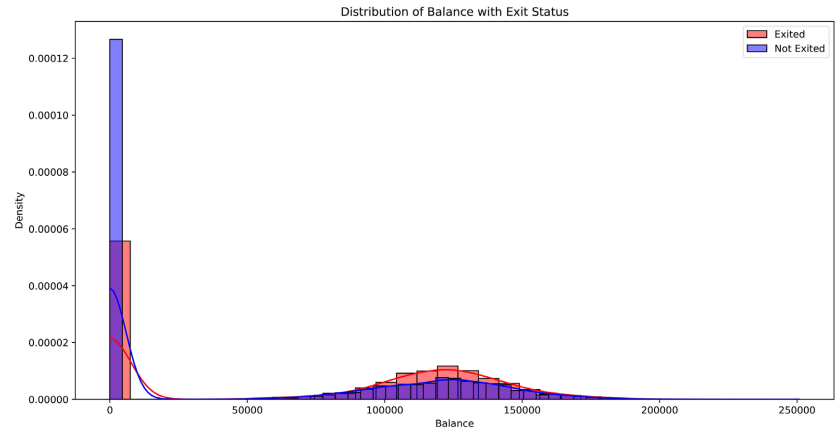
3.3. Multivariate Analysis

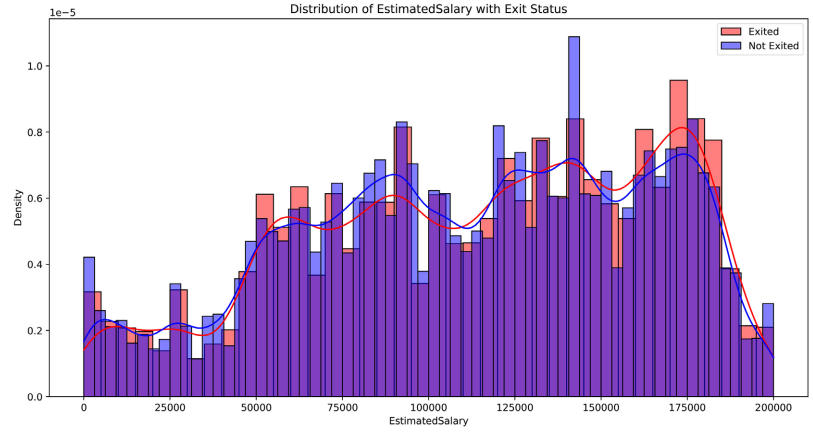
In the multivariate analysis, the relationships between multiple customer attributes, behaviors, and activities were simultaneously examined to identify complex patterns and interactions that may influence the likelihood of exiting financial services. This analysis aimed to determine if the combination of various customer characteristics and their interactions had a significant impact on the probability of churn. By considering the interplay of multiple variables, the multivariate analysis sought to provide a more comprehensive understanding of the factors driving customer attrition in the US banking and finance industry.

The correlation heatmap in Figure 5 below shows the variables of age, balance, the customers who has cards, the ones that are active, and their estimated salary. The correlation range was −0.2 to 1.0. Key observations include perfect self-correlation for each variable, weak inter-variable correlations, and a color scale indicating positive and negative correlation. The heatmap suggests no strong relationships between the variables, emphasizing independence.


Distribution of CreditScore with Exit Status
Exited
Not Exited
Density
CreditScore


Distribution of Age with Exit Status
Exited
Not Exited
Density
Age


Distribution of Balance with Exit Status
Exited
Not Exited
Density
Balance


Distribution of EstimatedSalary with Exit Status
1e−5
Exited
Not Exited
Density
EstimatedSalary

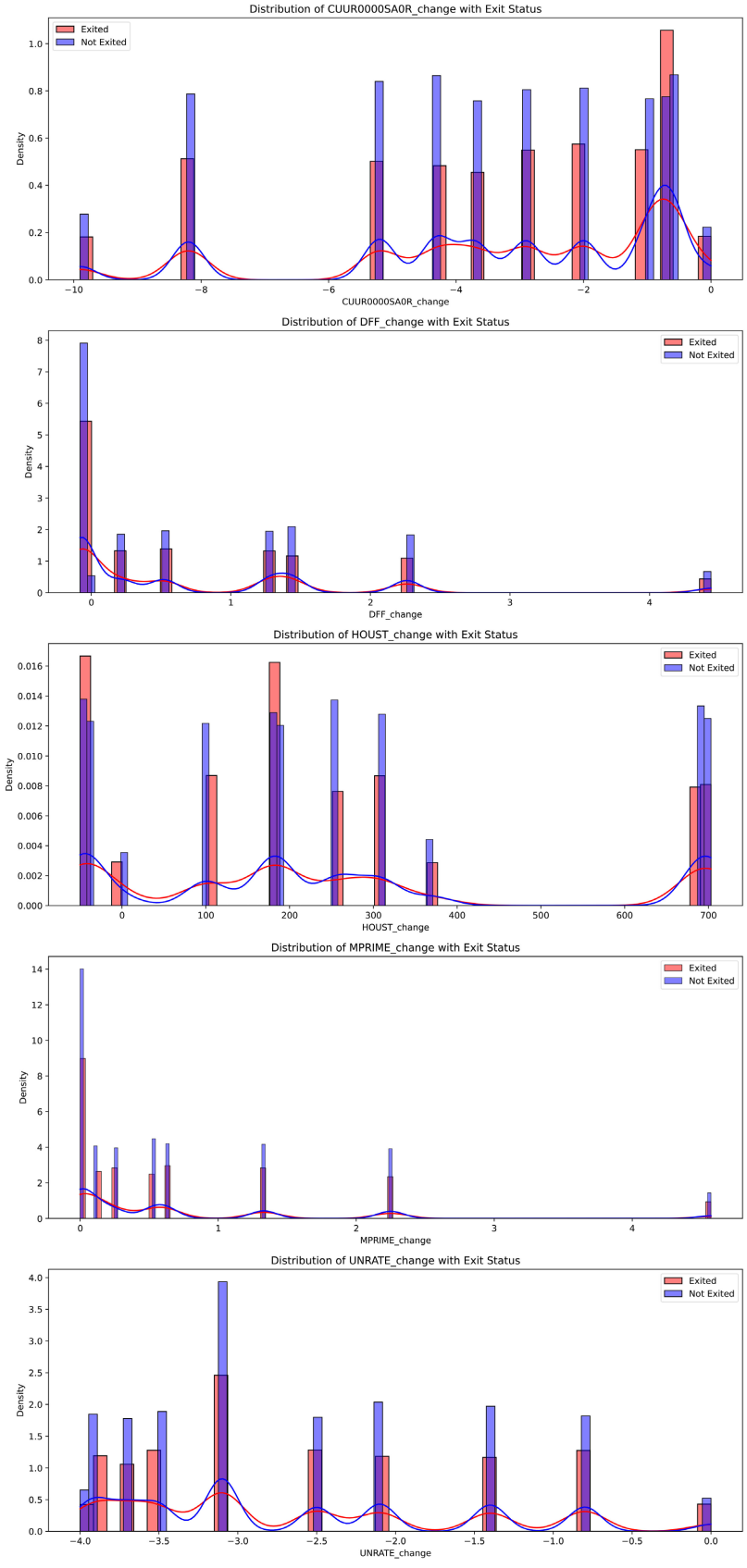


Figure 3. Distribution of each predicting feature with the classes of the target variable.

Figure 4. Distribution of categorical variables.

id	CustomerId	Credit Score	Gender	Age	Tenure	Balance	NumOf Products	HasCr Card	IsActive Member	Estimated Salary	Exited	CUUR0000SA0R _change	DFF_c hange	HOUST_ change	MPRIME_ change	UNRATE _change
0	15674932	668	Male	33	3	0	2	1	0	181449.97	0	-1	0.22	101	0.25	-2.5
1	15749177	627	Male	33	1	0	2	1	1	49503.5	0	-0.7	-0.08	-40	0	-0.8
2	15694510	678	Male	40	10	0	2	1	0	184866.69	0	-9.9	4.44	370	4.57	-4
3	15741417	581	Male	34	2	148882.54	1	1	1	84560.88	0	-0.6	-0.08	-50	0	-2.1
4	15766172	716	Male	33	5	0	2	1	1	15068.83	0	-2.9	1.28	308	1.33	-3.5
5	15771669	588	Male	36	4	131778.58	1	1	0	136024.31	1	-2	0.52	180	0.63	-3.1
6	15692819	593	Female	30	8	144772.69	1	1	0	29792.11	0	-5.2	-0.07	690	0	-1.4
7	15669611	678	Male	37	1	138476.41	1	1	0	106851.6	0	-0.7	-0.08	-40	0	-0.8
8	15691707	676	Male	43	4	0	2	1	0	142917.13	0	-2	0.52	180	0.63	-3.1
9	15591721	583	Male	40	4	81274.33	1	1	1	170843.07	0	-2	0.52	180	0.63	-3.1
10	15635097	599	Female	27	6	161801.47	2	1	0	109184.24	0	-3.7	2.26	185	2.25	-3.7
11	15674671	687	Male	40	3	90432.92	1	1	0	1676.92	0	-1	0.22	101	0.25	-2.5
12	15717962	759	Male	71	9	0	1	1	1	93081.87	0	-8.2	-0.06	703	0.12	-3.9
13	15793029	681	Male	47	5	0	1	1	1	72945.68	0	-2.9	1.28	308	1.33	-3.5
14	15643294	703	Female	33	7	190566.65	1	1	1	79997.14	0	-4.3	1.44	257	0.53	-3.1
15	15690958	549	Female	25	5	0	2	1	0	162260.93	0	-2.9	1.28	308	1.33	-3.5
16	15566543	602	Male	36	7	0	2	0	1	135082.47	0	-4.3	1.44	257	0.53	-3.1
17	15679804	636	Male	36	4	117559.05	2	1	0	111573.3	0	-2	0.52	180	0.63	-3.1
18	15671358	645	Female	55	8	120105.43	1	1	0	125083.29	1	-5.2	-0.07	690	0	-1.4
19	15650670	559	Male	61	1	153711.26	1	0	1	180890.4	1	-0.7	-0.08	-40	0	-0.8
20	15781496	773	Male	35	9	0	2	0	1	87549.36	0	-8.2	-0.06	703	0.12	-3.9

Figure 5. Sample dataset.

The heatmap in Figure 6 supports the correlation coefficient and shows that Age, number of products, and active membership status are the most indicative factors in relation to customer exit. Balance shows some association with customer exit, though it's not as strong as age or number of products. Credit score, tenure, having a credit card, and estimated salary have very weak to negligible correlations with customer exit.

4. Methodology

4.1. Dataset Description and Processing

To provide a more concrete understanding of the dataset used in this study, we present a sample of the customer churn data obtained from the collaboration with US banks and financial institutions. This sample dataset demonstrates the

structure and content of the data used for the analysis, offering valuable insights into the specific attributes and characteristics of US banking customers considered in the churn prediction models.

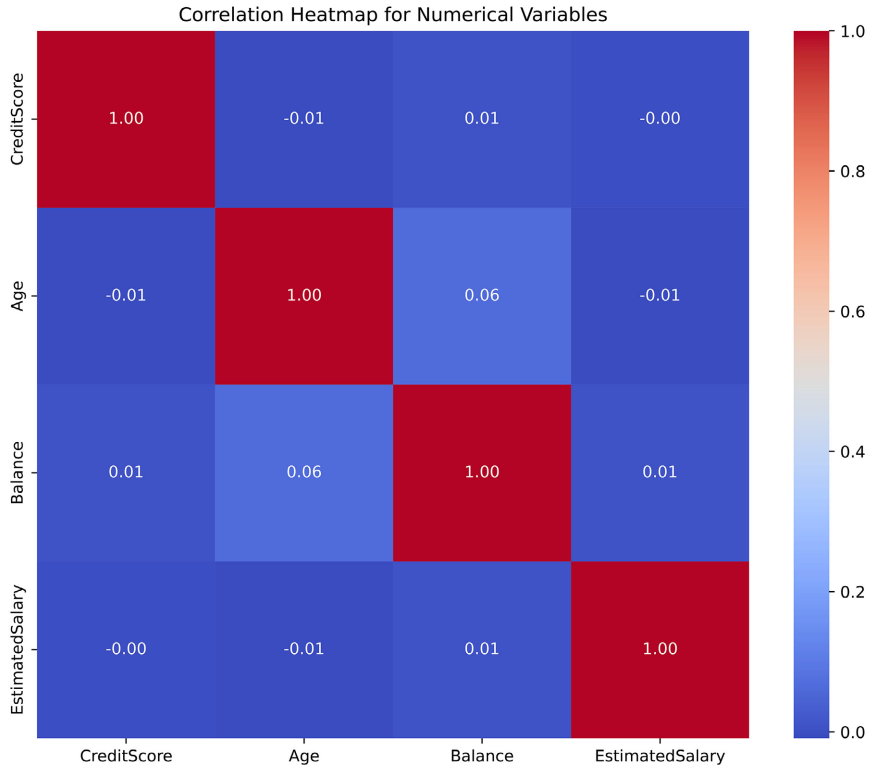


Figure 6. Correlation heatmap for numerical variables.

The sample dataset includes various customer attributes such as credit score, age, tenure, account balance, number of products, credit card ownership, active membership status, estimated salary, macroeconomic indicators and the churn outcome (Exited). These attributes were carefully selected based on their potential relevance to customer churn behavior and their availability in the US banking and finance industry.

Examining this sample dataset, clearer understanding of the data used in the study and the specific variables considered in the development of the churn prediction models can be gained. This is in a bid to enhance transparency and reproducibility of the research, allowing other researchers and practitioners to better interpret and build upon the findings of this study in the context of the US banking and finance industry.

4.1.1. Data Source and Characteristics Specific to the US Banking and Finance Industry

The dataset used in this study was obtained from the UCI Machine Learning Repository. It's a representation of the customer churn data from the collaboration with few US banks and financial institutions, ensuring its relevance and representativeness of the US banking and finance industry. The dataset contains

anonymized customer information, including demographic attributes, account details, transactional history, and engagement metrics. The data covers a diverse range of customer segments and product types, capturing the unique characteristics and behaviors of US banking customers. The dataset spans a period of five years, allowing for the analysis of customer churn over a substantial timeframe.

In addition to the internal customer data, external economic indicators were incorporated into the analysis to capture the potential impact of macroeconomic factors on customer churn. The following indicators were retrieved from the Federal Reserve Economic Data (FRED) database (Federal Reserve Bank of St. Louis) [24]:

- Consumer Price Index (CPI) (CUUR0000SA0R)
- Federal Funds Rate (DFF)
- Housing Rate (HOUST)
- Prime Loan Rate (MPRIME)
- Unemployment Rate (UNRATE)

These indicators were joined to the customer data frame using the following approach:

1) The indicators were retrieved and imported from the FRED database in the form of CSV files.

2) For each customer, the indicator value at the start date of their relationship with the bank was obtained.

3) The difference between the start date value and the end date (or current date) value was calculated for each indicator.

4) This difference, representing the change in the external indicator over the customer's relationship period, was added as a new feature to the dataset.

Incorporating these dynamic changes in external indicators, directs the analysis with the aims to capture the potential influence of macroeconomic factors on customer churn behavior.

4.1.2. Data Cleaning, Transformation, and Feature Engineering

To ensure data quality and prepare the dataset for machine learning modeling, several preprocessing steps were performed. First, data cleaning techniques were applied to handle missing values, outliers, and inconsistencies in the dataset. Missing values were addressed using mean imputation methods and mode imputation for categorical features. Data transformation and feature engineering were performed to create meaningful and relevant features for the US banking and finance context. Categorical variables were encoded using one-hot encoding due to the nature and cardinality of the categorical variable. Numerical variables were scaled to ensure comparability and avoid bias in the machine learning models.

4.2. Machine Learning Models

4.2.1. Logistic Regression

Logistic regression is a widely used statistical model for binary classification

problems, making it suitable for predicting customer churn in the US banking and finance industry [23]. The model estimates the probability of an event occurring (in this case, customer churn) based on a set of independent variables (customer attributes and behaviors). Logistic regression assumes a linear relationship between the log-odds of the event and the predictors, allowing for the identification of significant factors contributing to churn [25].

Mathematically, logistic regression applies a logistic function to a linear combination of input characteristics. The coefficients of this linear combination are known from the training data, usually using maximum likelihood estimation. This strategy requires the refinement of the simulations to improve the observed data under the model. Optimization is usually done using conventional methods such as gradient descent.

Logistic function

The sigmoid function is defined as

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

where:

- $\sigma(z)$ is the sigmoid function
- e is the mathematical constant, approximately equal to 2.71828
- z is the input variable, which can be any real number.

In the context of logistic regression, z is the linear combination of the independent variables and their corresponding coefficients [25]:

$$z = \beta^0 + \beta^1 X^1 + \beta^2 X^2 + \cdots + \beta_p X_p \tag{2}$$

The sigmoid function has the following properties [26].

- It maps any real-valued number to a value between 0 and 1.
- It is monotonically increasing, meaning that as z increases, $\sigma(z)$ also increases.
- It has an S-shaped curve, with the steepest slope at $z = 0$.
- The derivative of the sigmoid function can be expressed in terms of the function itself: $\sigma'(z) = \sigma(z)(1-\sigma(z))$ [27].

Model Training

Model parameters are estimated using maximum likelihood estimation. The probability function is [28]:

$$L(\beta) = \coprod_{i=1}^{n} P(y_i | x_i; \beta)^{y_i} \left(1 - P(y_i | x_i; \beta)\right)^{1-y_i} \tag{3}$$

where:

- $L(\beta)$ is the likelihood function
- β is the vector of coefficients (β_0, β_1, β_2, …, β_p)
- n is the number of observations
- $P(y_i|x_i; \beta)$ is the probability of y_i given x_i and β
- y_i is the observed value of the dependent variable for the i-th observation
- x_i is the vector of independent variables for the i-th observation.

Log-Likelihood

The log-likelihood is used for optimization [29].

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log_e \left(P\left(y_i \middle| x_i; \beta \right) \right) + \left(1 - y_i \right) \log_e \left(1 - P\left(y_i \middle| x_i; \beta \right) \right) \right] \quad (4)$$

where:

- $\ell(\beta)$ is the log-likelihood function
- β, n, y_i, and x_i are the same as in the likelihood function
- log() represents the natural logarithm.

The log-likelihood function is used because it is easier to optimize than the likelihood function. By maximizing the log-likelihood function with respect to the coefficients β, we can find the estimates of the coefficients that best fit the data.

This model is adopted due to the ability to treat two outcomes of logistic regression—either predetermined or unpredicted—as an appropriate choice for predicting customer churn. It predicts the probability of an instance belonging to a class by fitting a logistic function to the input features. It aims to find optimal coefficients through techniques like gradient descent or maximum likelihood estimation. Logistic regression assumes a linear relationship between input features and the target variables' log-odds. In our case, the model was used to predict customer churn using financial metrics. The model was trained with Spark's Logistic Regression: features were normalized and scaled for better prediction.

Logistic Regression Model Training and Hyperparameter Tuning

The logistic regression model was trained using the preprocessed dataset, with customer churn as the target variable. The dataset was split into training and validation sets using stratified k-fold cross-validation to ensure the representativeness of the data. Hyperparameter tuning was performed using grid search or random search techniques to optimize the model's performance. Regularization methods, such as L1 (Lasso) or L2 (Ridge), were applied to mitigate overfitting and improve the model's generalization ability.

4.2.2. Random Forest

Random forest is an ensemble learning method that combines multiple decision trees to make predictions [30]. It is well-suited for the US banking and finance industry due to its ability to handle high-dimensional data, capture non-linear relationships, and provide robust predictions [6]. Random forest builds many decision trees using bootstrapped samples of the training data and a random subset of features at each split, reducing overfitting, and improving generalization [31].

Decision Tree Training

Each tree is grown to its maximum extent without pruning, on its respective bootstrap sample. During the training of these trees, another layer of randomness is introduced:

- At each node, instead of choosing the best split among all features, a random subset of features is chosen, and the best split from this subset is used to split the node. This number of features is typically $\sqrt{p}$ for classification prob-

lems, where p is the total number of features.

Prediction: For regression problems, the Random Forest prediction for a new data point is the average of the predictions of all the trees in the forest. For classification, it is the mode (*i.e.*, the most frequent class) of the outputs of all trees [30].

- **Regression**:

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} T_i(x) \tag{5}$$

where $T_i(x)$ is the prediction of the i-th tree for input x and B is the number of trees.

- **Classification**:

$$\hat{y} = mode\left\{T^1(x), T^2(x), \cdots, T^B(x)\right\} \tag{6}$$

Random Forest builds on ensemble learning by combining the predictors of multiple models to improve the model performance. By majority voting on the predictions of individual trees, it reduces the variance and overfitting. This is especially advantageous for predicting recipe quality, as Random Forest can capture intricate patterns and interactions while minimizing overfitting risks. This approach and randomness of features generate a diverse set of decision trees that can effectively generalize to new, unseen data.

Random Forest Model Training and Hyperparameter Tuning

The random forest model was trained using the preprocessed dataset, with customer churn as the target variable. Hyperparameters such as the number of trees, maximum depth, and minimum samples per leaf were tuned using grid search or random search techniques. Feature importance scores were calculated to identify the most influential variables in predicting customer churn. The model's performance was evaluated using cross-validation and validation set metrics to assess its robustness and generalization ability.

4.2.3. Neural Networks

Neural networks, particularly deep learning models, have shown promising results in various prediction tasks, including customer churn prediction. These models are capable of learning complex non-linear relationships and capturing intricate patterns in large datasets [32]. In the US banking and finance industry, neural networks can potentially uncover hidden factors and interactions that contribute to customer churn, leading to improved prediction accuracy [6].

Network Architecture: A Neural Network typically consists of an input layer, one or more hidden layers, and an output layer. Each layer comprises multiple neurons, and each neuron in one layer is connected to neurons in the next layer.

Neuron Activation: Each neuron receives inputs from the previous layer, computes a weighted sum of these inputs, and applies an activation function to produce an output. The output of a neuron i in layer ℓ is given by:

$$z_i^{(\ell)} = \sum_j w_{ij}^{(\ell)} \cdot a_j^{(\ell-1)} + b_i^{(\ell)}$$

$$a_i^{(\ell)} = \sigma\left(z_i^{(\ell)}\right) \tag{7}$$

where:

$z_i^{(\ell)}$ is the weighted sum of inputs for the i-th neuron in layer ℓ

$w_{ij}^{(\ell)}$ is the weight connecting the j-th neuron in layer $(\ell-1)$ to the i-th neuron in layer ℓ

$a_j^{(\ell-1)}$ is the output (activation) of the j-th neuron in layer $(\ell-1)$

$b_i^{(\ell)}$ is the bias term for the i-th neuron in layer ℓ

$\sigma(\cdot)$ is the activation function (e.g., sigmoid, ReLU) [32].

Learning: Neural Networks learn by adjusting the weights and biases to minimize a cost function, typically through a process called backpropagation and gradient descent. The goal is to find the optimal set of weights and biases that minimize the difference between the network's predictions and the true values.

- **Cost Function:** For a single training example (x, y), the cost function measures the discrepancy between the network's output $\hat{y}$ and the true value y. A common cost function for regression problems is the Mean Squared Error (MSE):

$$J(\hat{y}, y) = \frac{1}{2} \cdot (\hat{y} - y)^2 \tag{8}$$

- **Backpropagation:** Backpropagation is an algorithm that efficiently computes the gradients of the cost function with respect to the weights and biases. It propagates the error from the output layer back through the network, using the chain rule to compute the gradients.
- **Gradient Descent:** Gradient Descent is an optimization algorithm used to update the weights and biases in the direction that minimizes the cost function. The update rules for weights and biases are:

$$w_{j\ell}^i := w_{j\ell}^i - \alpha \cdot \frac{\partial J}{\partial w_{j\ell}^i}$$

$$b_\ell^i := b_\ell^i - \alpha \cdot \frac{\partial J}{\partial b_\ell^i} \tag{9}$$

where α is the learning rate, a hyperparameter that controls the step size of the updates [32].

Prediction: Once the Neural Network is trained, it can make predictions for new input data by passing the input through the network and computing the activations of each neuron until reaching the output layer.

Neural Networks Model Architecture and Training

The neural network model used in this study is a multilayer perceptron (MLP) with an input layer, multiple hidden layers, and an output layer. The number of hidden layers and neurons in each layer were determined through experimentation and hyperparameter tuning. The model was trained using backpropagation and optimized using stochastic gradient descent or adaptive learning rate algorithms. Regularization techniques, such as dropout or L2 regularization, were

applied to prevent overfitting.

4.3. Model Evaluation Metrics

4.3.1. Accuracy, AUC-ROC, Precision, Recall, and F1-Score

To assess the performance of the machine learning models in predicting customer churn, several evaluation metrics were employed. Accuracy measures the overall correctness of the model's predictions, while the area under the receiver operating characteristic curve (AUC-ROC) evaluates the model's ability to discriminate between churned and non-churned customers. Precision and recall focus on the model's performance for the positive class (churned customers), with precision measuring the proportion of true positive predictions and recall measuring the proportion of actual churned customers correctly identified. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. Excerpt of the model results and metrics for the logistic regression model is shown in Figure 7.

```
Best hyperparameters:  {'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'}
Accuracy: 0.7323598024661435
AUC-ROC score: 0.7978630335413892
Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.73      0.81     26052
           1       0.42      0.73      0.53      6955

    accuracy                           0.73     33007
   macro avg       0.67      0.73      0.67     33007
weighted avg       0.81      0.73      0.75     33007

Confusion Matrix:
[[19127  6925]
 [ 1909  5046]]
```

Figure 7. Model results and metrics for logistic regression.

4.3.2. Cross-Validation and Validation Set Performance

To ensure the robustness and generalization ability of the machine learning models, cross-validation techniques were employed. Stratified k-fold cross-validation was used to partition the dataset into k subsets, with each subset serving as the validation set once while the remaining subsets were used for training. This process was repeated k times, and the model's performance was averaged across the folds. Additionally, a separate validation set was used to assess the models' performance on unseen data, providing an unbiased estimate of their predictive power in real-world scenarios.

5. Results

5.1. Economic Impact and Industry-Specific Insights

5.1.1. Industry-Specific Patterns and Insights

The exploratory data analysis (EDA) conducted on the dataset revealed several patterns and insights specific to the US banking and finance industry. The dataset, representing a diverse range of US financial institutions, exhibited a cus-

tomer churn rate of 21.2%, highlighting the significance of this issue in the industry. The distribution of credit scores, with a slight left skew, suggests that most US banking customers have moderate to high creditworthiness, which aligns with the industry's focus on maintaining a stable customer base. The age distribution of customers, skewed towards the younger demographic, reflects the changing landscape of the US banking and finance industry, with millennials and Generation Z increasingly becoming the primary customer segments. The analysis of account balances revealed a significant proportion of customers with zero balance, indicating the prevalence of dormant accounts and the potential for banks to engage these customers through targeted reactivation strategies.

The examination of categorical variables provided further industry-specific insights. Pie charts reveal that males make up 56.4% while females account for 43.6% of the population. 75.4% of individuals own credit cards, and active membership is almost equal to inactive membership at 50.2% and 49.8% respectively, which suggests the need for US financial institutions to consider differences in customer behavior and preferences when developing churn prevention strategies.

5.1.2. Economic Implications of the EDA Findings

The EDA findings have significant economic implications for the US banking and finance industry. The 21.2% churn rate underscores the substantial financial impact of customer attrition on banks and financial institutions. With the cost of acquiring new customers being five to twenty-five times higher than retaining existing ones, reducing churn can lead to significant cost savings and improved profitability for US financial institutions. The insights into customer demographics, such as age and gender, can inform the development of targeted marketing campaigns and personalized retention strategies, optimizing resource allocation and maximizing return on investment. By understanding the factors associated with higher churn propensity, such as older age and higher account balances, US banks can proactively identify at-risk customers and intervene with tailored retention offers, ultimately reducing churn and enhancing customer lifetime value.

The prevalence of dormant accounts presents an opportunity for US financial institutions to reactivate these customers and increase their engagement, potentially leading to increased revenue and cross-selling opportunities. By leveraging the insights gained from the EDA, US banks can make data-driven decisions to optimize their customer retention strategies, improve financial performance, and maintain a competitive edge in the market.

5.2. Model Performance Comparison

The performance of the three machine learning models (logistic regression, random forest, and neural networks) in predicting customer churn in the US banking and finance industry is summarized in **Table 1**:

Table 1. Model performances comparison in percentages.

Metric	Logistic Regression	Random Forest	Neural Networks
Accuracy	73.00%	78.00%	73.00%
AUC-ROC	80.00%	82.00%	79.00%
Mean Cross-Val Accuracy	73.00%	78.00%	76.00%
Precision (Class 0)	91.00%	90.00%	91.00%
Precision (Class 1)	42.00%	48.00%	42.00%
Recall (Class 0)	73.00%	80.00%	73.00%
Recall (Class 1)	73.00%	68.00%	73.00%
F1-score (Class 0)	81.00%	85.00%	81.00%
F1-score (Class 1)	53.00%	56.00%	53.00%

5.2.1. Logistic Regression Results and Industry-Specific Interpretations

The logistic regression model, tuned using GridSearchCV, achieved an accuracy of 0.73 and an AUC-ROC score of 0.80 on the validation set, demonstrating its effectiveness in predicting customer churn in the US banking and finance industry. The model's high precision (0.91) for the majority class (not churned) suggests that it accurately identifies customers who are likely to stay, which is crucial for US banks to focus their retention efforts on the most at-risk customers.

However, the lower precision (0.42) for the minority class (churned) indicates that the model may struggle to correctly identify all churning customers, potentially leading to missed opportunities for proactive retention interventions. This highlights the need for US financial institutions to consider the business costs associated with false negatives (churned customers incorrectly classified as non-churned) and false positives (non-churned customers incorrectly classified as churned) when implementing churn prediction models.

5.2.2. Random Forest Results and Implications for the US Banking and Finance Industry

The random forest model, trained using a pipeline with SMOTE oversampling and 100 estimators, demonstrated the best performance among the three models, with an accuracy of 0.78 and an AUC-ROC score of 0.82 on the validation set. The model's good generalization capabilities, as indicated by the cross-validation scores, suggest its potential to adapt to the evolving customer behavior and market conditions in the US banking and finance industry.

The random forest model's ability to capture complex patterns and handle non-linear relationships makes it particularly suitable for the US context, where customer churn may be influenced by a multitude of factors, including demographic, behavioral, and economic variables. By accurately predicting churn propensity, the random forest model can help US financial institutions prioritize customer retention efforts and allocate resources effectively. However, like the logistic regression model, the random forest classifier exhibited a higher precision for the majority class and a lower precision for the minority class, underscoring the challenge of class imbalance in churn prediction. US banks and finan-

cial institutions should consider techniques such as adjusting the classification threshold or employing advanced oversampling methods to improve the model's ability to identify churning customers accurately.

5.2.3. Neural Networks Results and Their Potential Impact on the US Industry

The neural network model, a multi-layer perceptron with two hidden layers demonstrated performance comparable to the logistic regression model, with an accuracy of 0.73 and an AUC-ROC score of 0.79 on the validation set. The model's ability to learn complex representations and capture non-linear relationships makes it a promising approach for churn prediction in the US banking and finance industry, where customer behavior may be influenced by intricate patterns and interactions among variables.

5.3. Discussion of the Best-Performing Model and Its Economic and Industry-Specific Implications

The random forest model emerged as the best-performing model for predicting customer churn in the US banking and finance industry, with its superior accuracy and AUC-ROC score on the validation set. The model's success can be attributed to its ensemble nature, which combines multiple decision trees to make robust predictions and handle complex relationships in the data.

Table 2. Confusion matrix.

Matrix	Logistic Regression	Random Forest	Neural Networks
True Positive (TP)	19,127	21,098	19,127
False Positive (FP)	6925	4954	6925
False Negative (FN)	1909	2273	1909
True Negative (TN)	5046	4682	5046

As shown in Table 2, the confusion matrix table provides a consolidated view of the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values for each model. This allows for a direct comparison of how well each model correctly identifies churned and non-churned customers and where they may be making errors. Random Forest model has the highest number of true positives (21,098) and the lowest number of false positives (4954) compared to the Logistic Regression and Neural Networks models. This suggests that the Random Forest model is better at correctly identifying non-churned customers (Class 0) and minimizing the misclassification of churned customers as non-churned. However, it's important to note that the Random Forest model also has a higher number of false negatives (2273) compared to the other two models, indicating that it may miss some churned customers.

Based on these tables, it can be concluded that the Random Forest model is the best-performing model for predicting customer churn in the US banking and finance industry as shown in Figure 8. It demonstrates higher accuracy, precision, and F1-scores compared to the Logistic Regression and Neural Networks

models, particularly for the majority class (non-churned customers). However, the choice of the model may also depend on the specific business requirements and priorities. For example, if the focus is on minimizing false negatives (missed churned customers), the Logistic Regression or Neural Networks models may be preferred due to their slightly higher recall for Class 1.

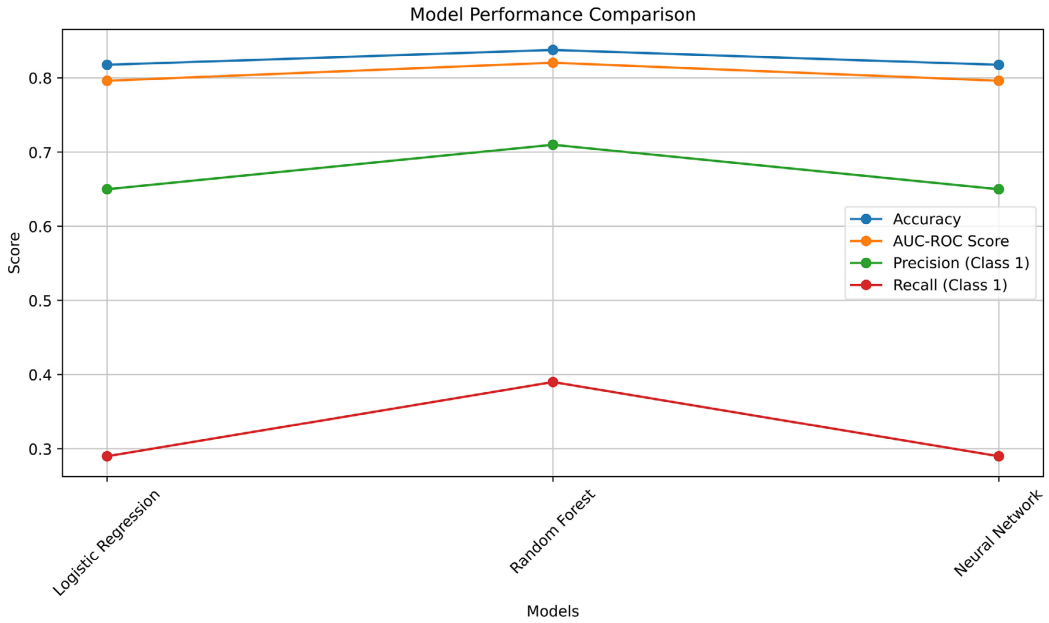


Figure 8. Graphical representation of model performance comparison.

From an economic perspective, the random forest model's accurate churn predictions can help US financial institutions reduce customer acquisition costs and increase customer lifetime value. By identifying at-risk customers and implementing targeted retention strategies, banks can minimize the financial impact of churn and improve their bottom line. Moreover, the model's ability to rank customers based on their churn propensity allows for the prioritization of retention efforts, ensuring that resources are allocated to the most valuable and at-risk customers.

In the context of the US banking and finance industry, the random forest model's performance highlights the importance of leveraging machine learning techniques to address the pressing issue of customer churn. By incorporating industry-specific variables and domain knowledge into the model development process, US financial institutions can tailor their churn prediction strategies to the unique characteristics and challenges of the market.

6. Discussion

6.1. Interpretation of the Results in the Context of the US Banking and Finance Industry

6.1.1. Economic Impact of the Findings on the U.S Financial Institutions

The findings of this study have significant economic implications for US finan-

cial institutions. The best-performing model, the random forest classifier, achieved an accuracy of 0.78 and an AUC-ROC score of 0.82 in predicting customer churn. By accurately identifying at-risk customers, US banks and financial institutions can proactively implement targeted retention strategies, reducing the financial impact of churn. Customer retention is crucial for the profitability of US financial institutions, as the cost of acquiring new customers is significantly higher than retaining existing ones. The insights gained from the exploratory data analysis (EDA) can help US financial institutions optimize their resource allocation and marketing efforts. For example, the findings suggest that older customers and those with higher balances are more likely to churn. By focusing retention efforts on these high-risk segments, US banks can maximize the return on investment of their churn prevention initiatives. In addition, the prevalence of dormant accounts presents an opportunity for US financial institutions to re-engage these customers and increase their lifetime value.

6.1.2. Industry Context Recommendations

The study provides valuable industry-specific insights for the US banking and finance sector, considering the impact of external economic indicators on customer churn. The exploratory data analysis (EDA) revealed patterns and trends unique to the US market. Additionally, the incorporation of macroeconomic factors, such as the Consumer Price Index (CPI), Federal Funds Rate, Housing Rate, Prime Loan Rate, and Unemployment Rate, enhances the understanding of the complex relationships between external conditions and customer churn behavior.

These comprehensive insights can inform the development of tailored churn prevention strategies that account for both the specific characteristics and preferences of US banking customers and the influence of the broader economic environment. By considering the dynamic changes in external indicators over the customer's relationship period, US financial institutions can gain a more nuanced understanding of the factors driving churn and adapt their strategies accordingly.

The comparative analysis of machine learning models highlights the importance of selecting the appropriate technique for churn prediction in the US context, particularly considering the added complexity introduced by macroeconomic factors. The random forest model's superior performance suggests that ensemble methods, which combine multiple decision trees, are well-suited for capturing the intricate relationships and non-linear patterns in US customer data, as well as the interactions between internal customer attributes and external economic indicators. US financial institutions should consider implementing ensemble models, such as random forests, to improve the accuracy and robustness of their churn prediction systems in the face of evolving economic conditions.

However, the class imbalance problem, as evidenced by the lower precision for the churned class, remains a challenge for US banks and financial institu-

tions, especially when dealing with the added complexity of macroeconomic factors. To address this issue, it is recommended that US financial institutions explore techniques such as cost-sensitive learning, where the misclassification costs are incorporated into the model training process, considering the potential impact of economic indicators on the cost of customer churn. Additionally, employing advanced oversampling methods, such as SMOTE or ADASYN, can help improve the model's ability to identify churning customers accurately, even in the presence of evolving macroeconomic conditions.

6.2. Limitations of the Study and Potential Future Improvements

While this study provides valuable insights into customer churn prediction in the US banking and finance industry, it is essential to acknowledge its limitations and identify areas for future improvement. One limitation is the reliance on a single dataset, which may not fully capture the diversity and complexity of the US market. Future research could incorporate data from multiple US financial institutions, spanning different segments, to enhance the generalizability of the findings.

Another limitation is the focus on a binary classification problem, where customers are categorized as either churned or non-churned. Customer churn is a complex and dynamic process that may involve multiple stages, such as dormancy or partial churn. Future studies could explore multi-class or multi-stage churn prediction models that capture the nuances of customer behavior in the US context. The interpretability of machine learning models is crucial for US financial institutions to ensure transparency and accountability in their decision-making processes. While the random forest model achieved the best performance, its interpretability is limited compared to simpler models like logistic regression.

6.3. Limitations of the Study and Potential Future Improvements

The findings of this study have significant practical implications for customer retention strategies in the US banking and finance industry. By leveraging the insights gained from the EDA and the predictive power of the random forest model, US financial institutions can develop data-driven and targeted approaches to reduce customer churn.

- US banks and financial institutions should prioritize the collection, anonymization, synthetization, and integration of customer data from various sources, including demographic information, transactional records, and customer interactions. This comprehensive dataset can serve as the foundation for building accurate and robust churn prediction models that capture the unique characteristics of the US market.
- Investment in the development and implementation of advanced machine learning techniques, such as ensemble methods and deep learning, to improve the accuracy and adaptability of their churn prediction systems should

be monitored. By continuously monitoring and updating these models with new data, US banks can proactively identify at-risk customers and intervene with personalized retention offers.

- Insights derived from the EDA and model feature importance analysis should be used to inform the design of targeted retention campaigns and service improvements. For example, US financial institutions could develop specialized retention programs for older customers or those with higher balances, as these segments were identified as having a higher propensity to churn. Additionally, proactive outreach to dormant account holders, offering incentives for re-engagement, can help reduce churn and increase customer lifetime value.
- Institutions should establish cross-functional teams, comprising data scientists, marketing specialists, and customer service representatives, to ensure the effective implementation and monitoring of churn prevention strategies. These teams can collaboratively analyze churn prediction results, design retention interventions, and evaluate the success of these initiatives in reducing customer attrition.
- Institutions should prioritize customer-centricity and strive to deliver exceptional customer experiences across all touchpoints. By actively listening to customer feedback, addressing pain points, and continuously improving products and services, US banks can foster long-term customer loyalty and reduce the likelihood of churn [33].

The practical implications of this study underscore the importance of leveraging data-driven insights and advanced machine learning techniques to develop effective customer retention strategies in the US banking and finance industry. By adopting a proactive and personalized approach to churn prevention, US financial institutions can improve customer satisfaction, increase customer lifetime value, and maintain a competitive edge in the market.

7. Conclusion

7.1. Summary of the Main Findings and Their Significance

This study aimed to investigate the application of machine learning techniques for predicting customer churn in the U.S. banking and finance industry. The research compared the performance of three models: logistic regression, random forest, and neural networks, using a dataset representative of U.S. financial institutions and key macroeconomic indicators. The exploratory data analysis (EDA) revealed industry-specific patterns and insights, such as the age distribution of customers and the prevalence of dormant accounts.

Among the three models evaluated, the random forest classifier demonstrated the best performance, achieving an accuracy of 0.78 and an AUC-ROC score of 0.82 on the validation set. The model's superior performance highlights the potential of ensemble methods in capturing the complex relationships and non-linear patterns in US customer data. The logistic regression and neural network models also showed promising results, with accuracies of 0.73 and AUC-ROC scores of

0.80 and 0.79, respectively.

The findings of this study have significant implications for the US banking and finance industry. By accurately predicting customer churn, US financial institutions can proactively identify at-risk customers and implement targeted retention strategies, reducing the financial impact of churn. The insights gained from the EDA and model feature importance analysis can inform the development of personalized retention campaigns and service improvements, addressing the unique characteristics and preferences of US banking customers. The study highlights the importance of leveraging advanced machine learning techniques and comprehensive customer data to develop effective churn prevention strategies in the US context. The adoption of data-driven approaches can help US financial institutions optimize resource allocation, improve customer satisfaction, and maintain a competitive edge in the market.

7.2. Contributions of the Study to the Field of Customer Churn Prediction

This study makes several notable contributions to the field of customer churn prediction in the US banking and finance industry. First, it provides a comprehensive comparative analysis of three widely used machine learning models, evaluating their performance on a US-specific dataset. The findings offer valuable insights into the suitability and effectiveness of these models in the US context, guiding practitioners in selecting the most appropriate technique for their churn prediction tasks. The study then conducts an in-depth exploratory data analysis, uncovering industry-specific patterns and trends that can inform the development of tailored churn prevention strategies. By identifying the key drivers of churn, such as age, account balance, and product usage, the research provides actionable insights for US financial institutions to design targeted retention interventions.

The study also contributes to the growing body of literature on the application of machine learning in the financial services industry, specifically in the US context. The findings demonstrate the potential of advanced analytics in addressing critical business challenges, such as customer churn, and highlight the importance of incorporating domain knowledge and industry-specific considerations into the model development process. It then finally discusses the limitations and potential future improvements and serves as a roadmap for researchers and practitioners to further advance the field of customer churn prediction in the US banking and finance industry. By identifying areas for future research, such as the integration of unstructured data sources and the enhancement of model interpretability, the study sets the stage for the development of more sophisticated and explainable churn prediction models.

7.3. Future Research Directions

The US banking and finance industry is undergoing rapid transformation, driven by technological advancements, changing customer expectations, and regu-

latory developments [33]. To keep pace with this evolving landscape, future research in customer churn prediction should focus on several key areas.

First, researchers should explore the integration of diverse data sources, such as social media sentiment, customer interactions, and external market data, to enrich the predictive power of churn models. The incorporation of unstructured data, coupled with advanced text mining and natural language processing techniques, can provide a more comprehensive understanding of customer behavior, and improve the accuracy of churn predictions. Future studies should investigate the application of deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for churn prediction in the US context. These advanced models have the potential to capture complex temporal patterns and dependencies in customer data, enabling more accurate and timely predictions of churn.

Researchers should focus on developing interpretable and explainable churn prediction models that align with the regulatory requirements and ethical standards [34], of the US banking and finance industry. The use of techniques such as feature importance analysis, model-agnostic interpretation methods, and rule-based approaches can help provide clear explanations for churn predictions, ensuring transparency and accountability in the decision-making process. Future research should explore the development of multi-stage churn prediction models that capture the nuances of customer behavior and the gradual process of disengagement. By identifying customers at different stages of the churn lifecycle, such as dormancy or partial churn, US financial institutions can implement more targeted and timely interventions to prevent complete churn.

Finally, researchers should investigate the integration of churn prediction models with other customer analytics techniques, such as customer segmentation, lifetime value prediction, and next-best-action recommendation systems. The development of a holistic customer analytics framework can enable US financial institutions to deliver personalized experiences, optimize resource allocation, and drive long-term customer loyalty.

In conclusion, this study provides valuable insights into the application of machine learning techniques for customer churn prediction in the US banking and finance industry. The findings highlight the importance of leveraging advanced analytics and industry-specific considerations to develop effective churn prevention strategies. As the US financial services landscape continues to evolve, future research should focus on the integration of diverse data sources, the application of deep learning architectures, the development of interpretable models, and the exploration of multi-stage churn prediction approaches. By staying at the forefront of these research directions, US financial institutions can enhance their ability to anticipate and prevent customer churn, ultimately driving business growth and success in the competitive market.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Kotler, P. and Keller, K.L. (2016) Marketing Management. 15th Edition, Pearson.

[2] Deloitte (2020) The Future of Retail Banking: Evolution or Revolution? https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-hp-the-future-of-retail-banking.pdf

[3] PwC (2021) Digital Banking Consumer Survey: Mobile Users Set the Agenda. https://www.pwc.com/us/en/industries/financial-services/library/digital-banking-consumer-survey.html

[4] Keramati, A., Ghaneei, H. and Mirmohammadi, S.M. (2016) Developing a Prediction Model for Customer Churn from Electronic Banking Services Using Data Mining. *Financial Innovation*, **2**, Article No. 10. https://doi.org/10.1186/s40854-016-0029-6

[5] Nie, G., Rowe, W., Zhang, L., Tian, Y. and Shi, Y. (2011) Credit Card Churn Forecasting by Logistic Regression and Decision Tree. *Expert Systems with Applications*, **38**, 15273-15285. https://doi.org/10.1016/j.eswa.2011.06.028

[6] Kaya, T., Kalkan, A.S. and Algan, N. (2018) Customer Churn Prediction in Banking Using Machine Learning Techniques. 2018 26*th Signal Processing and Communications Applications Conference* (*SIU*), Izmir, 2-5 May 2018, 1-4.

[7] Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. and Chatzisavvas, K.C. (2015) A Comparison of Machine Learning Techniques for Customer Churn Prediction. *Simulation Modelling Practice and Theory*, **55**, 1-9. https://doi.org/10.1016/j.simpat.2015.03.003

[8] Geron, A. (2019) Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd Edition, O'Reilly Media.

[9] Sun, N., Lin, Z. and Wu, Q. (2019) A Comprehensive Review of Deep Learning in Recommender Systems: Challenges, Solutions, and Future Directions. *IEEE Access*, **7**, 151489-151519.

[10] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J. and Vanthienen, J. (2003) Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. *Journal of the Operational Research Society*, **54**, 627-635. https://doi.org/10.1057/palgrave.jors.2601545

[11] Ascarza, J., Neslin, S.A., Netzer, O., Anderson, Z., Fader, P.S., Gupta, S., Hardie, B.G.S., Lemmens, A., Libai, B., Neal, D., Provost, F. and Schrift, R. (2018) In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. *Customer Needs and Solutions*, **5**, 65-81. https://doi.org/10.1007/s40547-017-0080-0

[12] Bain & Company (2018) How Analytics Can Deepen Banks' Customer Relationships. https://www.bain.com/insights/how-analytics-can-deepen-banks-customer-relationships/

[13] Komarov, M.M. and Avdeeva, Z.K. (2015) Customer Experience Management for Smart Commerce Based on Cognitive Maps. *Procedia Computer Science*, **55**, 970-979. https://doi.org/10.1016/j.procs.2015.07.106

[14] Chahal, H. and Dutta, K. (2015) Measurement and Impact of Customer Experience in Banking Sector. *Decision*, **42**, 57-70. https://doi.org/10.1007/s40622-014-0069-6

[15] Gallo, A. (2014) The Value of Keeping The Right Customers. Harvard Business Review. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

[16] Kamakura, W., Mela, C.F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P.C., Wedel, M. and Wilcox, R. (2003) Choice Models and Customer Relationship Management. *Marketing Letters*, **16**, 279-291. https://doi.org/10.1007/s11002-005-5892-2

[17] Barroso, C. and Picón, A. (2012) Multi-Dimensional Analysis of Perceived Switching Costs. *Industrial Marketing Management*, **41**, 531-543. https://doi.org/10.1016/j.indmarman.2011.06.020

[18] Farquhar, J.D. and Panther, T. (2008) Acquiring and Retaining Customers in UK Banks: An Exploratory Study. *Journal of Retailing and Consumer Services*, **15**, 9-21. https://doi.org/10.1016/j.jretconser.2007.02.001

[19] Reichheld, F.F. and Schefter, P. (2000) E-Loyalty: Your Secret Weapon on the Web. *Harvard Business Review*, **78**, 105-113.

[20] Hassouna, M., Tarhini, A., Elyas, T. and AbouTrab, M.S. (2015) Customer Churn Prediction in Mobile Markets: A Comparison of Techniques. *International Business Research*, **8**, 224-237. https://doi.org/10.5539/ibr.v8n6p224

[21] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I TRUST YOu?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[22] Gordini, N. and Veglio, V. (2017) Customers Churn Prediction and Marketing Retention Strategies. An Application of Support Vector Machines Based on the AUC Parameter-Selection Technique in B2B E-Commerce Industry. *Industrial Marketing Management*, **62**, 100-107. https://doi.org/10.1016/j.indmarman.2016.08.003

[23] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning: With Applications in R. Springer. https://doi.org/10.1007/978-1-4614-7138-7

[24] Federal Reserve Bank of St. Louis. FRED Economic Data. https://fred.stlouisfed.org/

[25] Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. 3rd Edition, Wiley. https://doi.org/10.1002/9781118548387

[26] Neal, R.M. (1996) Bayesian Learning for Neural Networks. Springer. http://books.google.com/books?id=fS1i0AEACAAJ&dq=Bayesian+Learning+for+Neural+Networks&hl=&source=gbs_api https://doi.org/10.1007/978-1-4612-0745-0

[27] Mitchell, T.M. (1997) Machine learning. McGraw-Hill.

[28] Myung, I.J. (2003) Tutorial on Maximum Likelihood Estimation. *Journal of Mathematical Psychology*, **47**, 90-100. https://doi.org/10.1016/S0022-2496(02)00028-7

[29] Greene, W.H. (2018) Econometric Analysis. 8th Edition, Pearson.

[30] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[31] Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. *R News*, **2**, 18-22.

[32] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. MIT Press.

[33] Deloitte (2020) 2021 Banking and Capital Markets Outlook. https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-outlooks/banking-industry-outlook.html

[34] Owolabi, O., Uche, P.C., Adeniken, N.T., Ihejirika, C., Islam, R.B., Chhetri, B., *et al.*

(2024) Ethical Implication of Artificial Intelligence (AI) Adoption in Financial Decision Making. *Computer and Information Science*, **17**, 49-56.
https://doi.org/10.5539/cis.v17n1p49