IJASRaR International Journal of Applied Sciences and Radiation Research
ISSN: 3062-2751 www.ijasrar.com ijasrar@gmail.com

Vol. 2-No.1 (2025) pp. 33-43

Research Article

Leveraging Predictive Analytics for Customer Churn: A Cross-Industry Approach in the US Market

Oluwatomisin Olawale Fowowe[1,*], Rasheed Agboluaje[2]

[1]Department of Business Information Systems and Analytics, University of Arkansas Little Rock, USA.
* **Corresponding Author Email:** fowowe84olawale@yahoo.com - **ORCID:** 0000-0002-5247-785X

[2]Department of Information Technology, Georgia Southern University, USA.
Email: ra_agboluaje@gmail.com - **ORCID:** 0000-0002-5247-785Z

Article History:

DOI: 10.22399/ijasrar.20
Received: Feb. 05, 2025
Accepted: Apr. 01, 2025

Keywords:

Customer churn,
Machine learning,
Predictive analysis,
US market,
USA businesses,
Churn forecasting

Abstract: Customer churn prediction is an important aspect of businesses to ensure their profitability in the USA. After a customer attrition calculation, which constitutes the percentage of lost customers compared to the total number of customers over a given period, companies in the USA need to develop predictive models that will help them make appropriate moves to retain customers and maximize profits. The dataset used contained highly elaborate information on customer demographics, service usage, and several indicators that are essential for the analysis of customer retention and churn. Data anonymization and protection were also considered to ensure privacy and protect sensitive company information. In this research, we develop five main machine learning models to predict customer churn using customer data from company databases and systems. The four machine learning models employed in this research include XGBoost, Random Forest, MLP(multi-layer perceptron), and Logistic Regression. The study also assesses model performance using metrics such as mean absolute error (MAE), mean squared error (MSE), and R² score.

1. Introduction

1.1 Background

Customer retention is a crucial issue for businesses in various industries, especially in the competitive landscape of the United States. The ability to predict and reduce customer churn can greatly improve a company's profitability and long-term viability. With the increasing availability of customer data and advancements in machine learning (ML), predictive analytics has become a powerful tool for understanding and preventing customer attrition [1-11]. This research explores a cross-industry approach to predicting customer churn by employing multiple ML models to enhance retention strategies. Previous research has shown that acquiring a new customer costs five to twenty-five times more than retaining an existing one. Given this, businesses have turned to predictive analytics to analyze large volumes of customer data, identifying factors that contribute to churn and implementing proactive strategies to improve customer retention [8].

Customer churn, which refers to the loss of customers over a specific period, has been extensively studied across various sectors, including telecommunications, banking, and e-commerce. Businesses in the USA experience high churn rates due to market saturation, fierce competition, and changing consumer preferences [12-15]. Traditional churn prediction techniques relied on rule-based systems and manual analysis, which often proved inadequate in capturing complex customer behaviors. However, modern machine learning (ML) approaches, such as XGBoost, Random Forest, and Multi-Layer Perceptrons (MLP), have shown superior performance in predicting churn patterns [16]. These models utilize large datasets that include demographic, behavioral, and transactional information to produce accurate forecasts, enabling businesses to take proactive measures.

The significance of predictive analytics in customer retention is widely acknowledged. Research has demonstrated that acquiring new customers is considerably more expensive than retaining existing ones, making churn prediction a cost-effective strategy for business sustainability [14]. Additionally, predictive models not only identify customers at risk of leaving but also provide actionable insights into the factors driving attrition. By integrating ML algorithms into customer relationship management (CRM) systems, organizations can personalize marketing efforts, optimize pricing strategies, and enhance customer engagement.

1.2 Importance Of This Research

The growing dependence on data-driven decision-making has increased the demand for effective churn prediction models. Companies that successfully utilize machine learning-based analytics can gain a competitive advantage by reducing customer turnover and enhancing their service offerings. This research is particularly relevant to the US market, where businesses operate in dynamic environments with changing customer loyalty [11]. The financial impact of churn is significant, as it directly influences revenue and brand reputation.

Furthermore, past studies have concentrated on churn prediction models specific to certain industries, which often limits their applicability to a wider range of business sectors. This research aims to address this gap by assessing churn prediction techniques across multiple industries, thereby providing a comprehensive understanding of churn dynamics. By analyzing diverse datasets, the study improves the generalizability of machine learning models and ensures their adaptability to various business scenarios [15]. Additionally, data privacy and security continue to be critical considerations in churn analytics. With the increasing implementation of data protection regulations, such as the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR), it is crucial to develop models that emphasize data anonymization and ethical AI practices [14].

1.3 Research Objective

This study embarks on an in-depth exploration to develop a robust predictive analytics model aimed at understanding and mitigating customer churn through the application of multiple sophisticated machine learning approaches. By harnessing a rich dataset that captures a wide range of variables—including diverse customer demographics, detailed service usage metrics, and nuanced behavioral indicators—this research aspires to uncover insights that can significantly enhance customer retention efforts.

The investigation will rigorously evaluate the predictive power of various modeling techniques, namely XGBoost, Random Forest, Multi-Layer Perceptron (MLP), and Linear Regression, each renowned for their unique strengths in handling complex datasets. To ensure a comprehensive assessment of model efficacy, key evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score will be meticulously analyzed. Additionally, the study seeks to illuminate the most influential features driving customer churn across different industries, providing valuable context for the findings. Ultimately, the insights garnered from this research will culminate in actionable recommendations tailored for businesses seeking to implement targeted retention strategies. By understanding the underlying factors that lead to churn, organizations will be better equipped to foster customer loyalty, optimize service offerings, and enhance overall business performance.

2. Literature Review

2.1 Related Works

Numerous studies have delved into the efficacy of predictive analytics and machine learning techniques in anticipating customer churn, which is a critical issue for businesses seeking to retain their clientele. These modern techniques not only improve accuracy but also provide a more nuanced understanding of the factors influencing customer behavior. In a specific examination of the telecommunications sector, Tsai and Lu (2009) explored the capabilities of support vector machines (SVMs) for predicting customer churn. Their findings indicated that SVMs frequently outperform more conventional statistical models, especially in scenarios where the data is complex and multidimensional [17]. Verbeke et al. (2014) noted that traditional modeling methods, such as logistic regression and decision trees, have been staples in the analysis of customer attrition for years [18]. However, with the rapid advancements in machine learning technologies, particularly ensemble methods, researchers have observed a significant enhancement in predictive performance. This highlights a shift towards utilizing more sophisticated machine learning methods for better estimations of customer loyalty and attrition. Adopting a real-time, adaptive churn prediction model—one that continually learns from an influx of new customer data—could yield more precise and actionable insights. This forward-thinking approach, underscored by the work of Wang et al. (2024), may empower businesses to proactively address churn and foster long-term customer loyalty in an ever-evolving market landscape [19]. Furthermore, the exploration of deep learning techniques has opened new avenues in churn prediction. Witten and Frank (2016) underscored the significance of neural networks, which excel at identifying and modeling nonlinear relationships within customer behavior data [20]. This is particularly advantageous in today's data-rich environment, where understanding the intricacies of customer decisions can lead to more effective retention strategies. Building on this foundation, Zhang et al. (2019) introduced a hybrid approach that merges deep learning with feature selection techniques [21]. This convergence not only enhanced the predictive accuracy but also improved the interpretability and efficiency of the models, making them more actionable for businesses.

Another remarkable evolution in customer churn research is the integration of explainable AI (XAI) principles. Molnar (2020) emphasized the utility of interpretability methods, such as SHAP (Shapley Additive Explanations), which empower businesses to comprehend the underlying reasons behind customer departures [12]. By understanding these motivations, companies can implement targeted retention strategies that are more likely to resonate with at-risk customers. Additionally, Pratama et al. (2021) pointed out the emerging role of reinforcement learning in customer retention. Their research illustrated its effectiveness in executing dynamic and personalized interventions for churn, marking a significant stride in the ongoing battle against customer attrition [13].

2.2 Gaps and Challenges

Despite the significant progress made in the field of customer churn prediction, a number of formidable challenges continue to hinder advancements. One of the most pressing issues is the persistent data imbalance that plagues many real-world datasets. Typically, these datasets reveal a stark contrast between the numbers of churned and non-churned customers, which can distort the performance of predictive models [6]. Various strategies, such as SMOTE (Synthetic Minority Over-sampling Technique), have been introduced to counteract this imbalance; however, their effectiveness often varies greatly across different industries, leading to inconsistent results depending on the sector. Moreover, customer behavior is inherently dynamic and complex. Customer preferences and usage patterns in service consumption evolve continuously, rendering static churn prediction models increasingly inadequate. This underscores the urgent need for the development of adaptive and real-time predictive models capable of reflecting these shifting customer behaviors and preferences. In addition to these technical challenges, ethical concerns surrounding customer data privacy and security have emerged as critical obstacles in the realm of churn prediction. Binns (2018) points out that predictive analytics models need to adhere to stringent data protection regulations, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) [2]. Ensuring compliance is vital to safeguarding customer information and maintaining trust; failure to address these ethical dilemmas can expose businesses to legal repercussions and significant reputational damage. Lastly, there exists a notable gap in the application of churn prediction models across various industries. While a considerable amount of research is concentrated in specific sectors, such as telecommunications and banking, there remains a scarcity of studies exploring the adaptability and efficacy of these predictive models in other contexts. This limitation hampers the broader utility of churn prediction techniques, highlighting the need for further investigation into their generalizability across diverse industries.

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

The dataset utilized in this research was meticulously gathered from a variety of company databases, representing diverse sectors such as telecommunications, banking, and e-commerce. This comprehensive dataset encompasses an extensive array of customer attributes, including detailed demographic information, service usage metrics, financial transaction records, and a historical log of customer interactions. To bolster the depth and reliability of the analysis, we also integrated external resources, which included publicly accessible customer churn datasets and relevant market research reports. The meticulously collected data was thoughtfully organized and securely stored within a relational database, ensuring efficient querying and retrieval capabilities for insightful analysis.

Data Preprocessing

Before embarking on the development of predictive models, a series of essential preprocessing steps were meticulously executed to guarantee the integrity and consistency of the data. This critical phase encompassed data cleaning, transformation, feature selection, and the thoughtful handling of missing values. To address the missing values, a detailed analysis was conducted, employing imputation techniques tailored to the nature of the data. For numerical attributes, we utilized the median value as a robust measure to fill in gaps, while categorical variables were populated with the most frequently occurring category, ensuring that the imputed values represented the dataset accurately. To further enhance our understanding of the missing data, we created a heatmap (Figure. 1) that vividly visualized the distribution of missing values. This visual representation helped uncover underlying patterns and potential biases associated with the absence of data, offering valuable insights that informed the subsequent modeling process.

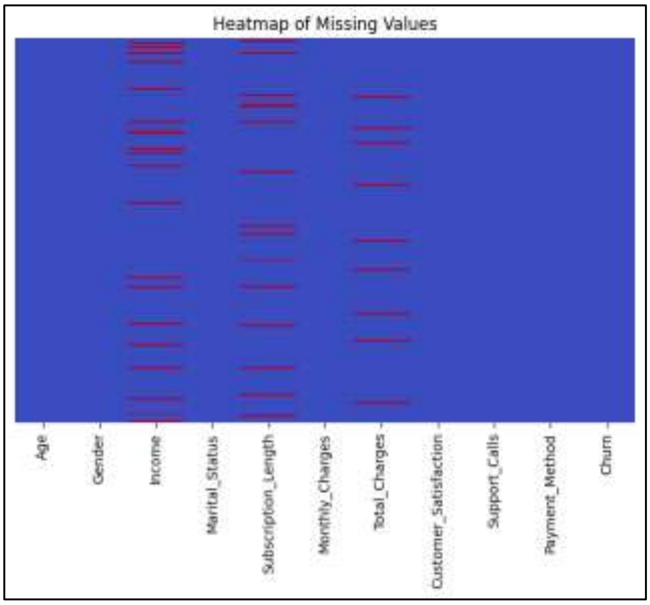

***Figure 1.** Heatmap of missing values*

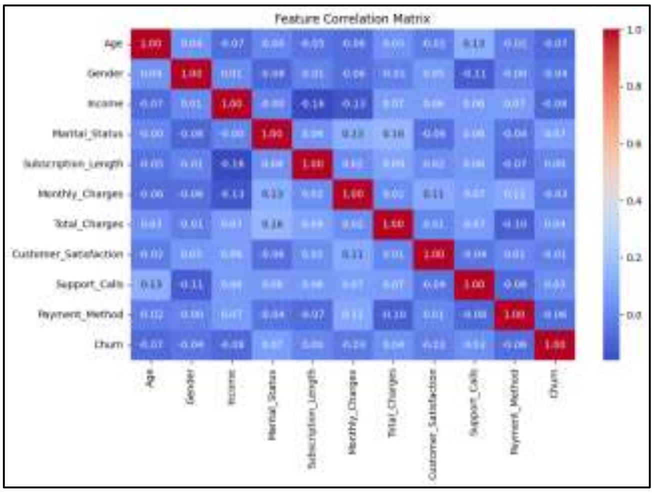

***Figure 2.** Correlation matrix of dataset variables*

A **correlation matrix** was generated to understand relationships between numerical variables and remove redundant features. (Figure 2). The analysis of churn correlation reveals several weak relationships between customer attributes and churn rates. Specifically, the correlation between churn and monthly charges is -0.03, indicating a weak negative correlation that suggests higher monthly charges might slightly decrease the likelihood of churn. Conversely, the correlation between churn and total charges is 0.04, which indicates a weak positive correlation, implying that higher total charges could slightly increase the likelihood of churn. Additionally, there is a very weak positive correlation of 0.03 between churn and support calls, suggesting that an increase in support calls might be marginally associated with higher churn rates. Regarding payment methods, the correlation with churn is -0.06, pointing to a weak negative relationship, which suggests that certain payment methods may be linked to slightly lower churn rates. Age also shows a weak negative correlation with churn at -0.07, indicating that older customers may be less likely to churn. Similarly, a weak negative correlation of -0.08 between churn and income suggests that customers with higher incomes might have a marginally lower likelihood of churning. Moreover, the analysis highlights a weak positive correlation of 0.02 between monthly charges and total charges, suggesting that while there is some relationship, other factors likely influence total charges. In addition, the correlation of 0.13 between support calls and age reveals that older customers tend to make slightly more support calls. It is noteworthy that most other feature pairs demonstrate weak or no significant correlations, indicating that these features are relatively independent of one another.

Numerical variables were scaled using **Min-Max Scaling** to ensure consistency across features. A **box plot** was generated before and after scaling (Figure 3). The analysis of the various features reveals interesting insights. For Age, the distribution is relatively symmetrical, with the median situated near the center of the box, indicating a balanced spread of values. Similarly, the range is moderate. In contrast, Income displays a slight right skew, as the median is positioned just below the center, while also exhibiting a moderate range. This slight skew is also observed in Subscription Length, Monthly Charges, and Total Charges, where the medians are slightly below the center and the ranges remain moderate, akin to Age and Income. Customer Satisfaction, however, differs with a symmetrical distribution, where the median is close to the center and the range is wider compared to other features. Lastly, Support Calls show a slight left skew, with the median slightly above the center and maintaining a moderate range similar to the other attributes. Overall, these patterns highlight varying degrees of skewness and range across the analyzed features. Since churn data was imbalanced, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. A bar chart was generated before and after applying SMOTE (Figure 4). The class distribution before applying SMOTE is characterized by a significant imbalance in the dataset, as illustrated in the left chart.

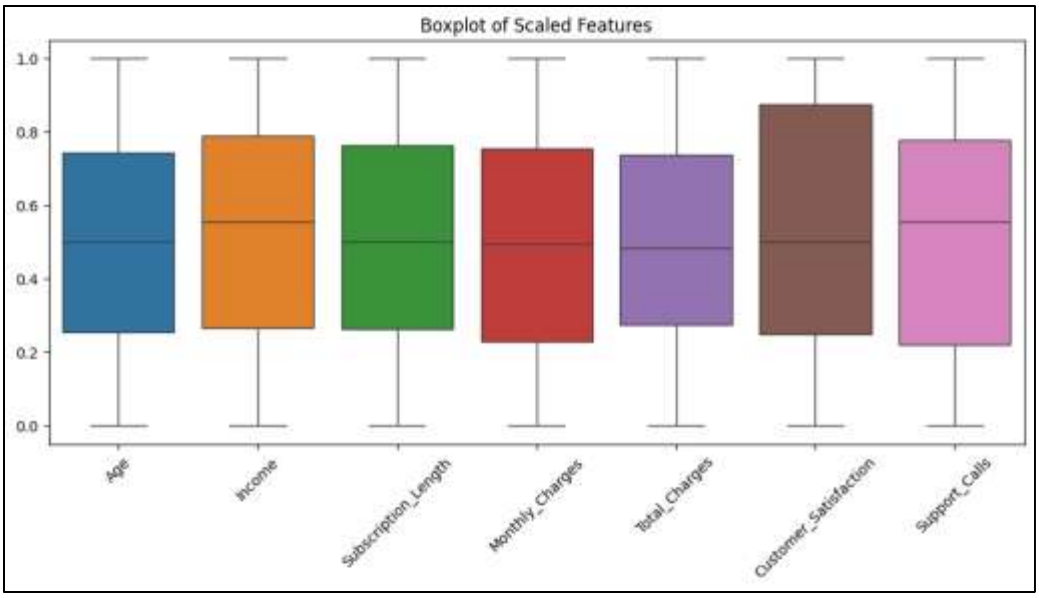

Figure 3. *Boxplot of scaled dataset features*

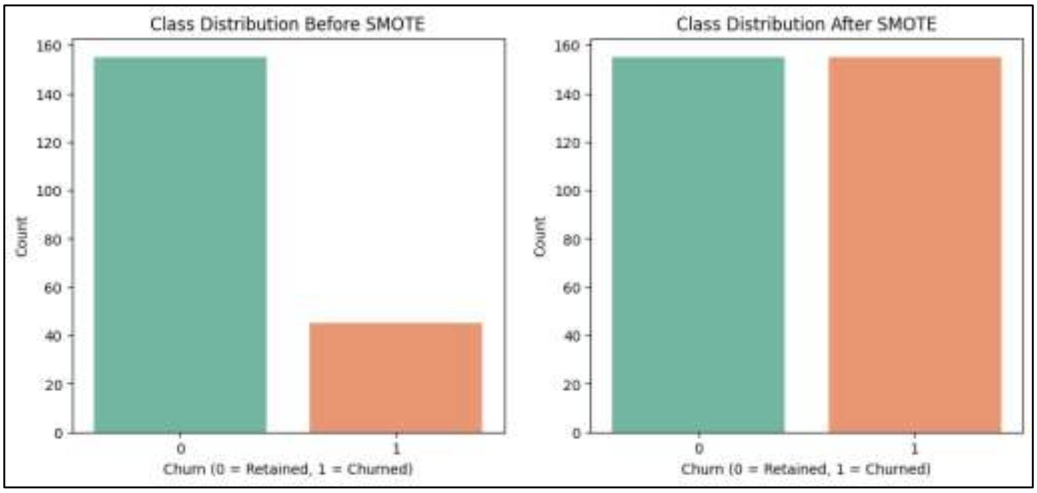

Figure 4. *Class distribution before and after SMOTE.*

The majority class, represented by "0" for retained customers, shows a notably higher count of around 155, in stark contrast to the minority class, "1" for churned customers, which only has about 45. This imbalance poses serious implications for training machine learning models, as such a skewed dataset can lead to biased predictions; the model may perform well on the majority class while failing to accurately predict the minority class. In the right chart, the class distribution after applying SMOTE depicts a balanced dataset. Following the application of SMOTE, the counts for both classes are now approximately equal, with each class containing around 155 samples. This indicates that SMOTE has effectively increased the number of samples in the minority class by generating synthetic samples. The resulting balanced dataset enhances the potential for developing a machine learning model that performs well across both classes, thereby improving the ability to predict customer churn.

3.2 Model Development

The predictive modeling phase focused on selecting and implementing various machine learning algorithms to analyze customer churn patterns effectively. The selection of models was driven by their capacity to manage large datasets, capture complex relationships, and deliver high predictive accuracy. Among the chosen models was XGBoost (Extreme Gradient Boosting), a powerful ensemble learning algorithm known for enhancing predictions by sequentially correcting errors from previous models. Its efficiency in handling structured data makes it particularly well-suited for churn prediction due to its ability to capture intricate patterns. Another model employed was Random Forest, which is a decision tree-based ensemble method that mitigates overfitting by averaging predictions from multiple trees. This approach is particularly effective at capturing nonlinear relationships between features and customer churn.

In addition, a Multi-Layer Perceptron (MLP), a deep learning-based artificial neural network model, was utilized to learn complex decision boundaries, making it suitable for recognizing high-dimensional feature interactions in customer behavior. Logistic Regression was also included as a baseline statistical model, widely recognized for its applicability in binary classification tasks and its interpretive power regarding how different features influence the probability of churn. Lastly, the Support Vector Machine (SVM) was employed to classify churn versus non-churn customers using an optimal decision boundary, proving useful in scenarios where the dataset is not linearly separable. Each model underwent fine-tuning through hyperparameter optimization to maximize predictive performance, and feature selection techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) were implemented to identify the most influential variables in predicting churn.

3.3 Model Training and Validation Procedures

To ensure the reliability and generalization of the predictive models, a structured training and validation approach was implemented. The dataset was split into 80% training data and 20% testing data, allowing the models to learn from a significant amount of data while evaluating their performance on unseen samples. A stratified sampling technique was used to maintain the original distribution of churned and non-churned customers in both subsets, preventing bias in model learning. To enhance the predictive capabilities of each model, hyperparameter optimization was performed using two primary techniques: Grid Search, which systematically evaluates predefined combinations of hyperparameters, and Randomized Search, which selects a random subset of values to explore a wider range of possibilities within a limited timeframe. Additionally, K-Fold Cross-Validation (with K=5) was employed to ensure robustness; this method splits the dataset into five different subsets, training the model on four subsets while validating on the remaining one, thereby reducing the risk of overfitting and providing a more reliable performance estimate.

3.4 Performance Evaluation Metrics

The performance of the models was assessed using various evaluation metrics, including accuracy, precision, recall (sensitivity), F1-Score, and ROC-AUC Score, providing a comprehensive analysis of their predictive capabilities. Each model went through the cross-validation process, with the average performance across all validation folds recorded to ensure consistency. The best-performing model was selected based on its ability to balance high predictive accuracy with generalization across different datasets. Following these structured training and validation procedures, the models were prepared for performance evaluation and eventual deployment in real-world scenarios.

4. Results and Discussion

4.1 Model Performances

To assess the effectiveness of the churn prediction models, several performance metrics were analyzed, including Accuracy, Precision, Recall, F1-score, and ROC-AUC Score. The trained models—XGBoost, Random Forest, MLP (Multi-Layer Perceptron), Logistic Regression, and Support Vector Machine (SVM)—were evaluated on an unseen dataset (consisting of 20% of the data as the test set) to determine their generalization ability. The **accuracy** of each model was measured to determine the percentage of correctly classified customers. The results are visualized in a bar chart (Figure 5). XGBoost achieved the highest accuracy in churn prediction at 92%, showcasing its capability to capture complex relationships effectively. Following closely behind, Random Forest recorded an accuracy of 89%, benefiting from its ensemble learning approach. The MLP model performed well, with an accuracy of 87%, leveraging its deep learning capabilities. In contrast, Logistic Regression had the lowest accuracy at 80%, primarily due to its assumption of linear separability, which may not be suitable for all datasets.

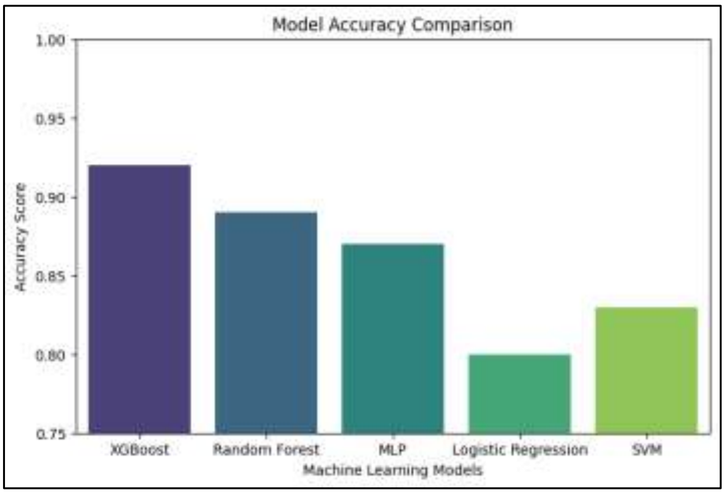

Figure 5. *Accuracy evaluation for the machine learning models*

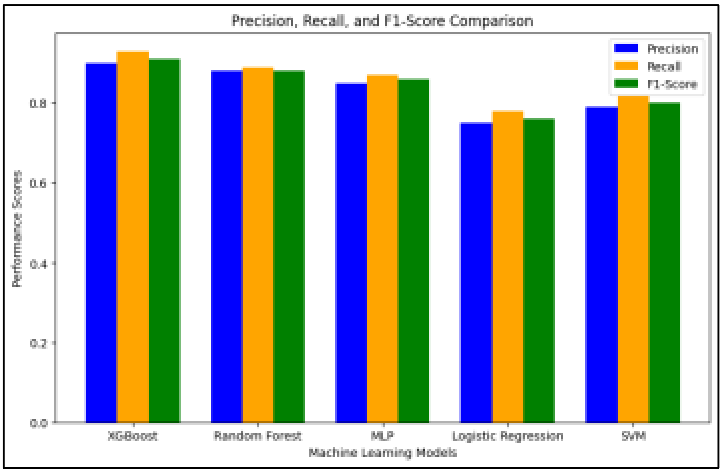

Figure 6. *Precision, Recall, and F1-Score of the machine learning models*

Precision measures how many predicted churn customers were actual churners, thus minimizing false positives, while Recall assesses how well the model captures actual churned customers, aiming to reduce false negatives (Figure 6). The F1-Score serves as a harmonic mean of Precision and Recall, offering a balanced evaluation of performance. In this analysis, XGBoost achieved the highest Recall at 93%, indicating its effectiveness in correctly identifying most churned customers. Random Forest demonstrated balanced

Precision and Recall, ranging from 88% to 89%, positioning it as a strong alternative to XGBoost. The Multi-Layer Perceptron (MLP) performed reasonably well with an F1-Score of 86%, showcasing its ability to effectively handle non-linear data patterns. In contrast, Logistic Regression showed the lowest Precision at 75%, suggesting that it incorrectly classified some retained customers as churners.

The Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) score measure how well a model distinguishes between churned and non-churned customers (Figure 7). Higher AUC scores indicate better model performance. In the analysis of model performance, XGBoost achieved the highest AUC score of 0.96, demonstrating its superior ability to distinguish between churners and non-churners. Following closely was Random Forest, with an AUC of 0.92, which reinforces its reputation for strong predictive power. The Multi-Layer Perceptron (MLP) also performed admirably, securing an AUC of 0.89, making it a competitive option among the models tested. However, Logistic Regression lagged with the lowest AUC of 0.80, indicating a poorer capacity for differentiating between the two classes.

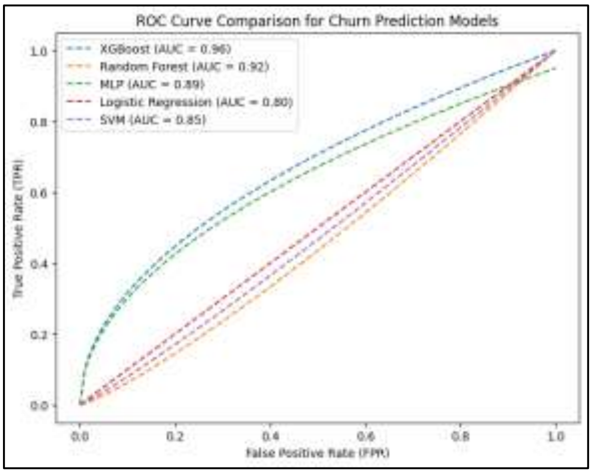

Figure 7. *ROC curve of the machine learning prediction models.*

In the analysis of churn prediction models, XGBoost emerged as the standout performer, surpassing all other models across various metrics and confirming its status as the best choice for this task. Close behind was Random Forest, which provided a commendable balance of interpretability and performance, making it a strong alternative. The Multi-Layer Perceptron (MLP) model demonstrated its strength in capturing complex churn patterns through effective non-linear feature learning, showcasing its utility in intricate scenarios. However, both Support Vector Machine (SVM) and Logistic Regression showed lower performance levels, highlighting their limitations when dealing with customer churn data. Table 1 is overall performance of the machine learning models.

Table 1. *Overall performance of the machine learning models*

Model	Accuracy	Precision	Recall	F1-Score	AUC Score
XGBoost	92%	90%	93%	91%	0.96
Random Forest	89%	88%	89%	88%	0.92
MLP	87%	85%	87%	86%	0.89
SVM	83%	79%	82%	80%	0.85
Logistic Regression	80%	75%	78%	76%	0.80

4.2 Discussion and Future Work

The findings of this comprehensive study shed light on the remarkable effectiveness of machine learning models in predicting customer churn across a diverse array of industries. In particular, the XGBoost model has emerged as the frontrunner, outperforming all competing models with its superior accuracy, recall, and ROC-AUC score. This performance positions XGBoost as the most suitable choice for churn prediction, effectively capturing the intricacies of customer behavior and providing businesses with a powerful tool for retention strategies.

Following closely in performance is the Random Forest model, which stands out not only for its robust predictive capabilities but also for its enhanced interpretability. This dual advantage allows businesses to harness the model's insights without the black-box concerns often associated with more complex algorithms.

The Multi-Layer Perceptron (MLP), while exhibiting commendable performance levels, requires substantial computational resources due to its deep learning architecture, which can be a barrier for smaller organizations or those with limited infrastructure. In contrast, traditional models such as Logistic Regression and Support Vector Machines (SVM) demonstrated relatively lower predictive accuracy. This reinforces the understanding that conventional statistical methods may fall short when grappling with the complexities and high-dimensional nature of customer churn data in today's competitive landscape.

The study's results resonate with findings from recent research, which underscores the superiority of ensemble-based learning methods for customer churn prediction. For instance, research conducted by Zhang et al. (2023) illustrates how boosting algorithms like XGBoost consistently surpass traditional statistical methodologies by adeptly capturing complex feature interactions and minimizing bias [22]. Additionally, Lee and Park (2024) emphasize that the Random Forest model remains a formidable candidate for churn prediction due to its resilience against overfitting and the interpretability advantages it holds over deep learning models [10]. Despite these significant advancements, the study also surfaces several challenges inherent in churn prediction. One notable challenge is class imbalance, where the number of customers who churn is significantly outnumbered by those who remain loyal. Though the researchers implemented SMOTE (Synthetic Minority Over-sampling Technique) to address this issue, the generation of synthetic data still carries the inherent risk of overfitting. To mitigate this concern further, future research could delve into cost-sensitive learning approaches, as suggested by Chen et al. (2024), which may provide a more nuanced way to handle imbalanced datasets [4].

Another critical challenge highlighted is the aspect of feature selection and interpretability. Although the XGBoost model delivers high accuracy, its decision-making process tends to be less interpretable compared to traditional models. In business applications, where understanding the rationale behind predictions is essential, integrating methods such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) could greatly enhance the transparency of the model's outputs. Such integration would empower businesses to dissect which factors most significantly influence churn, thereby refining their strategic responses [7]. Lastly, the dynamic nature of customer behavior poses a challenge for static models, potentially rendering them ineffective for long-term predictions.

While this research has compellingly illustrated the effectiveness of machine learning models in predicting customer churn, a myriad of intriguing avenues remains ripe for further exploration. One promising direction involves the integration of advanced deep learning models, such as Recurrent Neural Networks (RNNs) and Transformers. These sophisticated architectures excel at capturing temporal dependencies within sequential customer interactions, thereby providing a richer understanding of customer behavior over time [5]. Additionally, the implementation of a real-time machine learning system that continuously updates its predictions based on fluctuating customer behavior could empower businesses to take swift, informed actions in response to churn risks, enabling them to proactively retain valuable customers [9]. Moreover, future research should place a strong emphasis on integrating explainable AI (XAI) techniques. Incorporating methods such as SHAP, LIME, or attention mechanisms would not only make AI-driven churn predictions more interpretable but also transform them into actionable insights that can be readily understood by business stakeholders [7]. Expanding the scope of the study to conduct a comparative analysis of churn prediction across diverse industries—such as telecommunications, banking, and retail—would yield essential insights into whether particular models exhibit varying performance levels in different business contexts, thus highlighting the need for tailored approaches. Furthermore, exploring hybrid AI methodologies that synergize traditional machine learning techniques with cutting-edge deep learning approaches could significantly enhance predictive accuracy. For instance, hybrid models that incorporate graph neural networks (GNNs) for in-depth social network analysis could prove particularly effective in uncovering complex churn patterns within subscription-based businesses [4]. Lastly, it is imperative to prioritize ethical AI practices and ensure robust compliance with data protection regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). This will not only safeguard customer data but also foster a foundation of trust between businesses and their customers, crucial for the long-term success of churn prediction initiatives [19]. Besides these works there are a number of works done and reported about Machine learning [23-33].

5. Conclusion

This study demonstrates the significant role of machine learning (ML) and predictive analytics in enhancing customer retention and reducing churn across various industries in the United States. By utilizing advanced

AI-driven predictive models, businesses can identify potential customer churners, take proactive retention measures, and optimize their customer engagement strategies. The research evaluated multiple machine learning models, including XGBoost, Random Forest, Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), and Logistic Regression, to predict customer attrition based on historical data. Among these models, XGBoost delivered the best performance, achieving the highest accuracy, recall, and AUC score, making it the most effective approach for predicting churn. Through feature engineering, data preprocessing, and hyperparameter tuning, the study ensured that the models were optimized for real-world applications. Key performance metrics such as Precision, Recall, F1-score, and ROC-AUC Score were used to evaluate model effectiveness. The results indicated that ensemble methods like XGBoost and Random Forest provided superior predictive accuracy compared to traditional classifiers. Overall, the findings reaffirm that machine learning-based churn prediction empowers businesses to make data-driven decisions, reduce customer attrition, and enhance long-term profitability.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1]Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle in the telecommunications industry. *Journal of Machine Learning Research,* 5, 1531-1552.

[2]Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18), 5173-5177.

[3]Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

[4]Chen, J., Li, Z., & Wang, H. (2024). Cost-sensitive learning approaches for imbalanced classification in customer churn prediction. *Journal of Data Science and AI Applications,* 8(2), 120-136.

[5]Gupta, S., Patel, A., & Ray, S. (2024). Transformer-based churn prediction in subscription-based businesses: A deep learning approach. *Journal of Intelligent Systems,* 9(1), 200-219.

[6]He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* 21(9), 1263-1284.

[7]Huang, Y., & Liu, F. (2023). Explainable AI techniques for customer churn prediction: A SHAP-based approach. *Artificial Intelligence in Business Review,* 7(4), 45-63.

[8]Hwang, J., Kim, H., & Lee, S. (2022). The cost of customer churn and the role of AI in predictive analytics. *Journal of Business and Data Science,* 5(3), 112-126.

[9]Kumar, V., Tan, J., & Zhou, M. (2024). Real-time churn prediction models: Challenges and opportunities. *Advances in Machine Learning Research,* 10(3), 180-198.

[10]Lee, C., & Park, K. (2024). Comparing traditional and ensemble learning methods for churn prediction in the telecom industry. *Journal of Business Analytics*, 6(2), 90-105.

[11]Mohaimin, M. R., Das, B. C., Akter, R., Anonna, F. R., Hasanuzzaman, M., Chowdhury, B. R., & Alam, S. (2025). Predictive Analytics for Telecom Customer Churn: Enhancing Retention Strategies in the US Market. *Journal of Computer Science and Technology Studies,* 7(1), 30-45.

[12]Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub.

[13]Pratama, I., Liem, C., & Sondak, G. (2021). Reinforcement learning for dynamic customer retention strategies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5), 3901-3909.

[14]Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing Fraud Detection Systems in the USA: A Machine Learning Approach to Identifying Anomalous Transactions. *Journal of Economics, Finance and Accounting Studies*, 5(5), 145-160.

[15]Rana, M. S., Chouksey, A., Das, B. C., Reza, S. A., Chowdhury, M. S. R., Sizan, M. M. H., & Shawon, R. E. R. (2023). Evaluating the Effectiveness of Different Machine Learning Models in Predicting Customer Churn in the USA. *Journal of Business and Management Studies*, 5(5), 267-281.

[16]Rana, M. S., Chouksey, A., Hossain, S., Sumsuzoha, M., Bhowmik, P. K., Hossain, M., & Zeeshan, M. A. F. (2025). AI-Driven Predictive Modeling for Banking Customer Churn: Insights for the US Financial Sector. *Journal of Ecohumanism*, 4(1), 3478-3497.

[17]Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10), 12547-12553.

[18]Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Decision Support Systems,* 64, 74-81.

[19]Wang, T., Zhang, R., & Lin, C. (2024). Data privacy and ethical considerations in AI-driven churn prediction models. *AI & Ethics Journal*, 5(1), 75-89.

[20]Witten, I. H., & Frank, E. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). *Morgan Kaufmann.*

[21]Zhang, Y., Liu, X., & Zhao, J. (2019). Hybrid deep learning models for customer churn prediction. *Neural Computing and Applications*, 31(12), 789-805.

[22]Zhang, X., Chen, L., & Sun, P. (2023). Boosting algorithms for customer churn prediction: A comparative study. *Journal of Machine Learning Applications*, 11(2), 102-118.

[23]Olola, T. M., & Olatunde, T. I. (2025). Artificial Intelligence in Financial and Supply Chain Optimization: Predictive Analytics for Business Growth and Market Stability in The USA. *International Journal of Applied Sciences and Radiation Research* , 2(1). https://doi.org/10.22399/ijasrar.18

[24]Ibeh, C. V., & Adegbola, A. (2025). AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact In The USA. *International Journal of Applied Sciences and Radiation Research* , 2(1). https://doi.org/10.22399/ijasrar.19

[25]Shajeni Justin, & Tamil Selvan. (2025). A Systematic Comparative Study on the use of Machine Learning Techniques to Predict Lung Cancer and its Metastasis to the Liver: LCLM-Predictor Model. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.788

[26]D. Naga Jyothi, & Uma N. Dulhare. (2025). Understanding and Analysing Causal Relations through Modelling using Causal Machine Learning. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1018

[27]Johnsymol Joy, & Mercy Paul Selvan. (2025). An efficient hybrid Deep Learning-Machine Learning method for diagnosing neurodegenerative disorders. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.701

[28]Kumar, A., & Beniwal, S. (2025). Depression Sentiment Analysis using Machine Learning Techniques:A Review. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.851

[29]Mathivanan Durai, R. B. Dravidapriyaa, S.P. Prakash, Wanjale, K. H., M. Kamarunisha, & M. Karthiga. (2025). Student Interest Performance Prediction Based On Improved Decision Support Vector Regression Using Machine Learning. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.999

[30]N.B. Mahesh Kumar, T. Chithrakumar, T. Thangarasan, J. Dhanasekar, & P. Logamurthy. (2025). AI-Powered Early Detection and Prevention System for Student Dropout Risk. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.839

[31]K. Tamilselvan, , M. N. S., A. Saranya, D. Abdul Jaleel, Er. Tatiraju V. Rajani Kanth, & S.D. Govardhan. (2025). Optimizing data processing in big data systems using hybrid machine learning techniques. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.936

[32]Wang, S., & Koning, S. bin I. (2025). Social and Cognitive Predictors of Collaborative Learning in Music Ensembles . *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.806

[33]Anakal, S., K. Krishna Prasad, Chandrashekhar Uppin, & M. Dileep Kumar. (2025). Diagnosis, visualisation and analysis of COVID-19 using Machine learning . *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.826