Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Evaluating the Effectiveness of Different Machine Learning Models in Predicting Customer Churn in the USA

MD. Sohel Rana[1], **Anchala Chouksey**[2] **Bimol Chandra Das**[3], **Syed Ali Reza**[4] **Muhammad Shoyaibur Rahman Chowdhury**[5], **Mir Mohtasam Hossain Sizan**[6], **and Reza E Rabbi Shawon**[7]

[1]*Doctorate in Business Administration and Management, University of Cumberlands*

[2]*Masters of Science in Financial Mathematics, University of North Texas, Denton, Texas*

[3]*Master of Science in Business Analytics, Trine University, Indiana, USA*

[4]*B.Sc. In Electrical & Electronics Engineering, Ahsanullah University of Science & Technology*

[5]*Information Technology, Gannon University, Erie, PA*

[6]*Master of Science in Business Analytics, University of North Texas*

[7]*MBA Business Analytics, Gannon University, Erie, PA*

Corresponding Author: MD. Sohel Rana, **E-mail**: mrana32911@ucumberlands.edu

| ABSTRACT

Customer churn is deemed as the process by which customers stop using the product or service of a company now the burning issue facing organizations in the USA across all sectors: telecommunication, retail, banking, and subscription-based services. In the current competitive marketplace, companies in America face substantial challenges in retaining customers. The utmost objective of this study was to compare the performance of various machine learning algorithms in terms of predicting customer churn, thereby identifying the most effective techniques for accurately forecasting churn within US businesses. The scope of this study focused on contrasting machine learning algorithms for customer churn forecasting using extensive datasets derived from US businesses across various industries. The dataset of customer churn applied in this study is a rich set of data points developed to capture several dimensions of customer behavior and interaction with the business. For predicting customer churn, distinctive machine learning algorithms were considered, notably, Logistic Regression, Random Forest, and Gradient Boosting. The performance evaluation metrics of the models encompassed accuracy, precision, recall, F1-score, and ROC-AUC. While all three models perform similarly, SVM appeared to have the highest accuracy among the three algorithms. The adoption of machine learning for churn prediction extends considerable benefits for businesses in the United States alone, especially in highly competitive fields like telecommunications, retail, and subscription services. By leveraging predictive analytics, firms can identify high-risk customers and proactively engage with them to reduce churn rates and thereby improve customer loyalty.

| KEYWORDS

Customer churn, customer retention, churn forecasting, machine learning, prediction models, economic impact, USA businesses

| ARTICLE INFORMATION

ACCEPTED: 02 October 2023 **PUBLISHED:** 29 October 2023 **DOI:** 10.32996/jbms.2023.5.5.23

I. Introduction

Context and Importance

Al-Najjar et al. (2022), reported that in the current competitive marketplace, companies in America face substantial challenges in retaining customers. Customer churn is deemed as the process by which customers stop using the product or service of a company

now the burning issue facing an organization across all sectors: telecommunication, retail, banking, and subscription-based services. As per Ahmad et al. (2022), High churn rates do not just shrink revenues; they also hike up customer acquisition costs since the general trend for customer acquisition tends to be costly compared with retaining previous ones. Indeed, the economic consequences of churn are profound.

According to Agarwal et al. (2022), increasing customer retention rates by as little as 5% can lead to profit increases in the range of 25% to 95%. Despite its significance, churn also is intransigent. For instance, annual churn rates in the US telecommunications sector range up to 20%. A decline in churn has emerged as the most important profitability and growth driver in subscription businesses, such as online streaming or SaaS providers. The need to understand and predict churn, therefore, becomes a very relevant angle of strategic business planning in the USA (de Lima et AL., 2022).

Motivation

The motivation of the research is the dire need for accurate churn prediction models that may inform customer retention strategies. Businesses operate in an age characterized by rapid technological advancement and, subsequently, an exponential proliferation of data. However, converting this into actionable insight remains a challenge. Traditional churn prediction methods, usually based on simple statistical techniques, cannot capture the full complexity of customer behavior or the many factors that influence churn. By contrast, machine learning models offer a sophisticated approach to data analysis, enabling businesses to uncover patterns and relationships within the data that may not be immediately apparent.

The potential of machine learning algorithms to enhance prediction accuracy is specifically relevant in the setting of customer churn. By reaping the output of algorithms working on past data and subsequent new information upon it, businesses make predictive models aimed at identifying people who are possibly going to discontinue their purchasing from them and know the reasons one would do it. This knowledge will help businesses with targeted retention strategies by engaging customers at risk effectively and personalizing their offerings to meet the changing needs of the customers. In this respect, the exploration of different machine learning models for churn prediction is a very important area of research that promises enhancement in customer loyalty and ensures sustainable business growth.

Research Objective

The utmost objective of this study is to compare the performance of various machine learning algorithms in terms of predicting customer churn, thereby identifying the most effective techniques for accurately forecasting churn within US businesses. Logistic regression, XG-Boost, random forests, etc., are examples of a set of machine learning algorithms that are going to be within the scope of the study. The present study is designed to undertake an in-depth review of their strengths and weaknesses concerning churn prediction through the various evaluation metrics available to measure the predictive accuracy, precision, recall, and other related performance metrics of such models. This research will also look into the characteristic features and data attributes that result in effective churn prediction. The knowledge of which factors are most predictive of churn could enhance the interpretability of the models and inform strategic decision-making for businesses. This study ultimately aspires to offer actionable insights that can guide organizations in implementing data-driven retention strategies, thereby reducing churn rates and fostering long-term customer relationships.

Scope of the Research

The scope of this study is focused on contrasting machine learning algorithms for customer churn forecasting using extensive datasets derived from US businesses across various industries. It will contain all types of customer attributes, transactional data, engagement metrics, and demographic information that are rich enough to form the foundation for analysis. The study, therefore, aims to find common patterns and trends in churn behavior across a wide range of industries, underpinning variability that may be specific to certain industry contexts. Most importantly, it will also consider the aspect of model interpretability and explainability, considering that the reasoning behind churn prediction is important to enhance the trust of the stakeholders involved and to derive appropriate retention strategies. The findings of this study are expected to add a lot to the literature and practical value for businesses in their quest to understand and effectively manage customer churn amid an increasingly competitive marketplace.

II. Literature Review

Customer Churn

Morozov et AL. (2020), demystified that customer churn is usually defined as the proportion of customers who cease doing business with a company in a fixed period. It is a key metric to measure customer loyalty, generally speaking, and the well-being of the business. But from a business perspective, churn is not only important as a direct driver of revenues and profitability; it is also indicative of the satisfaction level of the customers about a company's service delivery and engagement policy. High rates of churn

can therefore serve as warning signs that there may be deep-seated problems regarding the level of service being offered or product provided or failures to meet customer expectations. Consequently, understanding the factors contributing to churn is essential for businesses that seek to maintain a competitive edge in increasingly saturated markets (Momin et AL., 2020).

He et AL. (2020), asserted that customer churn is much more than the financial costs resulting from such situations. Due to a higher churn rate, a company usually experiences various long-term negative factors that directly and indirectly include weakened brand loyalty, damaged brand reputation, dissuading prospects, and shrinking market share. Besides, Lalwani et AL. (2020), contended that the economic implications are crystal clear: it costs from five to twenty-five times more to acquire a new customer than it does to retain an existing customer. This very reality makes effective churn management strategies critical, entailing proactive outreach, targeted marketing, and mechanisms for thorough feedback from customers that can help identify issues before actual churn occurs.

However, there are many challenges in managing and predicting churn. First, there is complexity in customer behavior driven by possibly any one of various factors, including economic circumstances, competitive forces, and the preference of a customer. The other limitation of the traditional approaches to churn prediction is their grounding in historical data and simple statistical analyses that do not consider the dynamic nature of relationships customers enter with a company (Jamjoom, 2021). Companies may hardly identify customers who are at risk in time, and that would make them adopt reactive rather than proactive churn management strategies. Therefore, there is an urgent need for more sophisticated approaches to handle the complexity of customer churn and present actionable insights (Matuszelański, & Kopczewska, 2022).

Traditional vs. Machine Learning Approaches

Traditional churn prediction methods have focused on statistical techniques, including logistic regression models, cohort analysis, and descriptive statistics. Most of these methods are based on the use of historical customer data in the identification of patterns and trends that enable a business to estimate churn rates and target its retention efforts (Guliyev, 2021). Traditional methods can provide a general view of churn, but in most cases, they cannot represent such a complex multivariate problem as customer behavior (Celia & Osmanoglu, 2022). For example, logistic regression might oversimplify the relationships among the variables and miss critical interactions and nonlinearities that could affect the predictions of churn.

One of the noteworthy limitations of conventional methods is their dependence on predefined hypotheses, which can lead to biases and missed opportunities for deeper insights. Many of these methods also require heavy feature engineering and assumptions regarding the underlying data distribution, which may not be representative of the true complexity of real-world customer interactions. This makes a business unprepared to respond quickly enough in the face of rapid changes in market conditions or consumer preferences (Fuji et AL., 2022).

On the contrary, Faritha Banu et AL. (2022), argue that machine learning approaches avail much more dynamic and flexible frameworks towards churn prediction by using algorithms capable of learning from data and thus adapting to newer information, to carve out those hidden patterns and relationships that the other traditional methods have failed or overlooked. Machine learning algorithms, for example, can handle large volumes of data and identify complex relationships among its variables, making businesses enable predictive analytics of churn. Additional aspects of machine learning models are their capabilities to operate in high-dimensional spaces with no strict assumptions regarding data distribution, further broadening their area of applicability in predicting the probabilities of churn.

More importantly, machine learning models can provide constructive insight into the very drivers of this churning, hence enabling businesses to craft better retention strategies. The algorithms of machine learning can also target customers who have a high probability of churning out, considering their customer attributes, transaction behaviors, and engagement measures, and can offer intervention for the same. This proactive move not only works in the interest of an organization to improve the likelihood of retaining good business but also increases the value system within an organizational culture toward basing decisions on data-driven grounds (Beeharry & Tsokizep Fokone, 2022).

Machine Learning in Churn Prediction

As per Ahmad et AL. (2019), Machine learning for churn prediction has attracted lots of attention lately, resulting in various myriad models being applied to solve this problem. Logistic regression is one of the most popular, though among the oldest, having nowadays been extended to handle more complex features and interactions. Logistic regression can provide a baseline model that offers interpretable results, which a business can use to identify key predictors of churn. Another very popular machine learning method for churn prediction is decision trees. Decision trees create an interpretable and easy-to-visualize model by recursively splitting the data into subsets based on feature values. It may show in a very straightforward way what the most relevant factors of churn are. However, decision trees can be prone to overfitting, especially in the case of complex datasets. To overcome this weakness, ensemble methods such as random forests have emerged as one of the most powerful alternatives. Random forests

combine multiple decision trees to improve predictive performance and robustness, effectively reducing overfitting while maintaining interpretability (Al-Najjar et AL, 2019).

Agarwal et AL. (2022), posited that another ensemble technique that recently gained more traction in churn prediction is gradient boosting. This outcome is accomplished additively; each new model tries to correct the mistakes made by the previously constructed model. In time, this yields an amazingly accurate prediction model that would be able to capture most of the complex relationships among the variables. Gradient boosting algorithms, with their families including XG-Boost and Light-GBM, have gained wide popularity due to speed and efficiency, hence their applicability to big data that one usually deals with in a churn prediction task. Neural networks, and especially deep learning models, are relatively recent developments in machine learning applications for churn prediction. While they do require huge amounts of data and computational resources, these models have promised unprecedented capabilities concerning the extraction of complex patterns from large datasets. First, neural networks can learn relevant features from raw data without requiring extensive feature engineering. This is useful in situations where customer interactions may be multifeatured, as might be seen in telecommunications or e-commerce.

Various studies have been done on the performance of various machine learning models while predicting customer churn. For instance, in one of the most recent studies, logistic regression, decision trees, and random forests are some of the algorithms used to predict the possibility of a customer churning in a telecommunication company. The results indicated that among those, the best-performing algorithm is random forests; hence, this highlights the model's ability to capture such complex relationships that may exist between variables in the data. Other works have involved the use of techniques like gradient boosting. They were effective in forecasting churn across many industries, such as finance and retail. The works have also pointed out feature selection as one of the major ways of enhancing model performance. Various works have proven that relevant features such as customer demographics, transaction history, and engagement metrics greatly enhance the precision of churn predictions. This finding supports the idea that businesses should strive to be holistic in their approach to collecting and analyzing data, ensuring they capture every important aspect of customer behavior (DE Lima et AL., 2022).

III. Data Collection and Preprocessing

Dataset Description

The dataset of customer churn applied in this study is a rich set of data points developed to capture several dimensions of customer behavior and interaction with the business. Transaction history, such as transaction details, the frequency of transactions, and monetary values associated with each customer to show spending habits, is contained within the dataset. It also involves demographics: the age, gender, geographical location, and account age of the customers, in considering segment analyses for determining variability in demographic components on churn behavior. Interaction logs are included to document interactions on customer service support requests and submitted feedback. Records of interactions provide insight into general customer satisfaction or pain points that may give rise to churning. This is a rich and diverse dataset from multiple internal systems: a company's customer relationship management system, sales databases, and customer support systems that facilitate robust churn analysis and predictive modeling.

Data Preprocessing

Data preprocessing is an important step in preparing the Customer Churn dataset for analysis because it ensures that the machine learning models employed are valid and effective. First, we treated missing values mainly in the Total-Charges column by imputation-actually replacing the missing entries with either the mean or median of the available values, thus conserving the distribution of the dataset without too much bias. Secondly, we used one-hot encoding for the categorical variables: gender and the type of contract. This turned the features into a numerical format that will easily integrate with machine learning algorithms while capturing the categorical nature of data without assuming ordinality. Thirdly, we performed feature scaling to normalize numerical features such as tenure and monthly charges; for these, we used the z-score normalization technique so that such features have an average of zero and a standard deviation of one, which is important for some algorithms sensitive to feature scales, such as those involving gradient descent-based methods. The comprehensive preprocessing step provided a sound basis for the modeling effort that followed; it improved the convergence of many machine-learning methods used in this study.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a vital statistical technique used to analyze and summarize the main characteristics of a dataset, often employing visual methods to uncover patterns, trends, and anomalies that may not be immediately apparent through traditional statistical analysis. For example, in EDA, such research on customer churn is the most critical task: it allows an analyst to infer the underlying structure in data, and recognize the relations of different features like customer demographic data, transaction history, or engagement measures. This can enable us to take a glance at the distribution of the important variable, check the

correlation, and even find some outliers that might be useful for the choice critical to the proper machine learning model and strategy of feature engineering. On the contrary, deep insights are provided by EDA into data supporting the finding not only of the best predictive model but also help in formulating hypotheses on what drives customer churn to make informed decisions and effective targeted retention.

Distribution of Churn

The Python code snippet was computed aimed at visualizing customer churn distribution from an available dataset. Through the Seaborn library, it developed a count plot that plots customers who have churned and those who have not on this plot. Adding a title and label on a bar plot and improving readability by the color palette on the plot made it a valuable insight into comprehending the churn rate. Furthermore, with the understanding of the churn rate, identifying areas for improvement may also be found in improving customer retention.

Output:

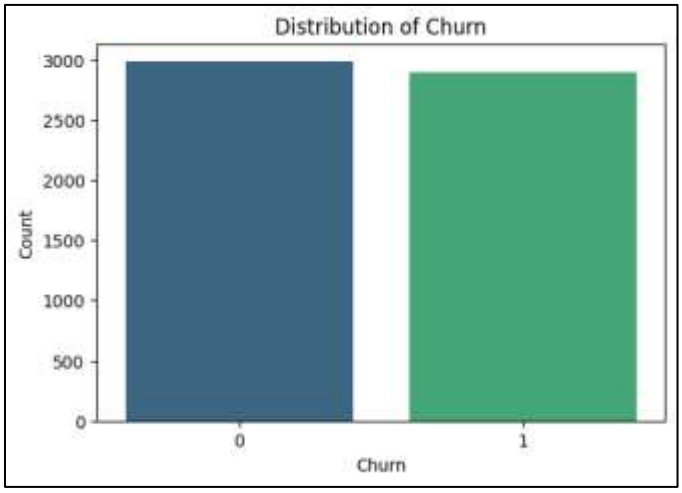

Figure 1: Displays the Distribution of Churn

Above is the histogram of the distribution of churn: containing two bars representing counts of customers who did not churn (0) and those who did churn (1). By looking at the heights, one can see that there is a relatively fair balance between these two, with a slight majority of customers having not churned, represented by the taller bar for the non-churned. Precisely, the number of non-churning customers is just over 2,500, while the number of churning customers is also close to 2,500, indicating that half of the customers in the dataset remained loyal and the other half did not continue their relationship with the business. This balanced distribution is important to the analysis because it shows the dataset is appropriate for training machine learning models without any issue of huge class imbalance that could otherwise skew predictive outcomes toward biased results. The understanding of this distribution is quite basic in making up effective retention strategies, which have to address both customer segments.

Monthly Charges vs. Total Charges

The executed Python code created a scatter plot to plot monthly charges versus the total charges of a set of customers. It colored the customers depending on whether they were in the churn group or not. The seaborn library first created a scatter plot and then set figure size, plot title, x, and y labeled, as well as the legend for showing which customers have churned or not. This visualization might help to find out if there are patterns or a trend in the association between monthly charges, total charges, and customer churn to help get at the root causes behind this customer attrition.

Output:

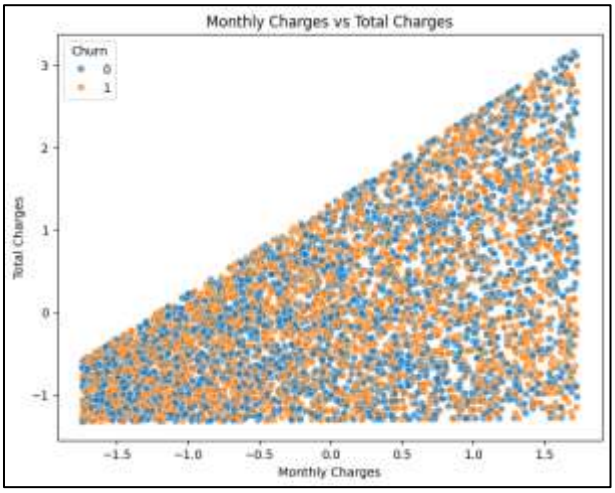

Figure 2: Monthly Charges vs. Total Charges

The scatter plot of Monthly Charges vs. Total Charges shows a positive trend and interesting churn insights about customers. Indeed, each point is a customer, where colors represent churn or not. The strongest feature is a positive relation between Monthly Charges and Total Charges indicated by the upward-sloping pattern of the data points. The customers with higher Monthly Charges develop greater Total Charges over time, meaning that bigger spending each month corresponds to potentially longer customers. However, the orange points reflecting churn are throughout the Total Charges and centered within the range of mid to high Monthly Charges, which conveys that despite how much the customers have spent, they may opt to leave anyway. As explained, this implies that not only the billed amount is in question but all factors attributing toward the satisfaction or retention of subscribers, as more-than-regional charges just by themselves ensure no loyalty in customers. So, the scatter plot denotes the churn dynamics' nature and the need for handling a multi-dimensionality in analyzing a particular problem that constitutes customer behavior in general.

Distribution of Tenure

The provided Python code fragment is designed to visualize the distribution of customer tenure in a dataset. Using the seaborn library, it creates a histogram to display the frequency of customers with different tenure lengths. The plot is customized with a title, labels, and a blue color for better readability. Additionally, a kernel density estimation (KDE) curve is overlaid on the histogram to provide a smooth representation of the underlying probability density function. This visualization would help understand the typical tenure of customers, identify potential churn points, and inform strategies for improving customer retention.

Output:

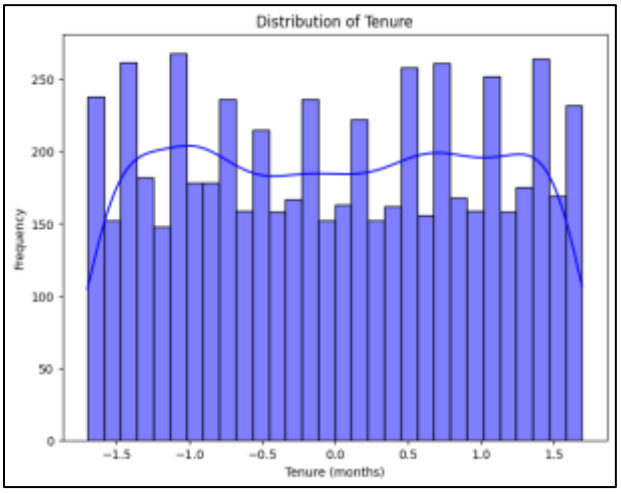

Figure 3: Depicts Distribution of Tenure

Above is the scatter plot showing the distribution of tenure. A histogram is overlaid with a smoothed line that is meant to help show customer retention over time. Bars illustrate frequency across different ranges of tenure, which seems reasonably well-distributed, indicating that customers are spread across various lengths of time with the service. The peaks in the histogram indicate that there are significant clusters of customers who have been with the service for certain lengths of time, while the smoothed line indicates the overall trend in customer retention. The consistency in frequency across these tenure ranges indicates a lack of significant bias towards either early leaving or long-time staying; however, higher frequencies at some tenure points might indicate potential periods of dissatisfaction or satisfaction. This distribution thus highlights the importance of understanding how tenure correlates with customer experiences and may necessitate targeted retention strategies at critical junctures in a customer's lifecycle.

Correlation Heat Map

The implemented code in Python was computed to plot the correlation between some numeric features in the provided dataset with a variable, "Churn". First of all, this is a list defining numeric features - "tenure", "Monthly Charges", and "Total Charges". Using Seaborn, plot a heatmap from an array of cross-correlation coefficients of these three features against "Churn". The plot was customized by adding a title, annotations for each cell to represent the value of the correlation, a colormap to visualize better, and a format for the values. Such a visualization helped in highlighting which of the numerical features had the maximum strength concerning customer churn and may bring important factors that could potentially cause customer attrition.

Output:

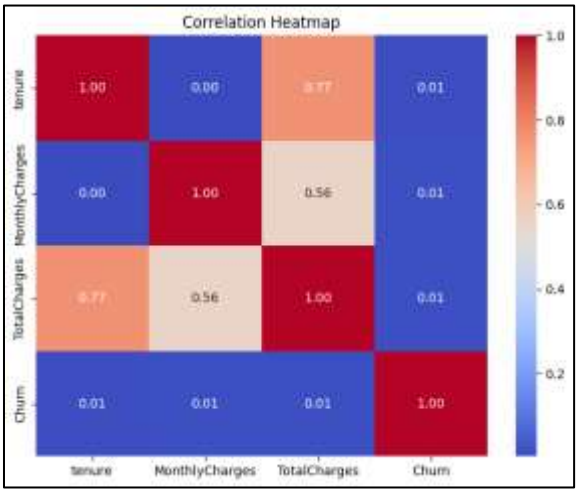

Figure 4: Portrays Correlation Heatmap

Above is the correlation heat map that relates tenure to Monthly Charges Total Charges and the rate of churning. In essence, it shows the relationship of these variables in numerical form. For example, there is a strong positive correlation between tenure and Total Charges of 0.77, which means the longer a customer has been in service, the more the total charge. Monthly Charges are also moderately positively correlated with Total Charges, with a value of 0.56, which means that the more expensive the monthly fee, the higher the contribution to the total charges. On the other hand, the correlation of churn with the rest of the variables is very low, close to zero, which means that neither tenure, neither Monthly Charges nor Total Charges are strong predictors of churn behavior. This finding implies that while financial metrics may reflect customer engagement, they do not directly influence the likelihood of a customer leaving the service, highlighting the need for additional factors to be considered in churn analysis.

Churn by Gender

The executed code snippet was computed to counterplot the relationship of gender to churn in the dataset. This code generates a counterplot using the seaborn library, defining figure size and customizing the plot by adding a title, labels, and color palette. This plot will display the number of male and female customers who have churned versus those who have not. This visualization can help determine if there is a significant difference in churn rates between genders, potentially revealing insights into factors that contribute to customer attrition.

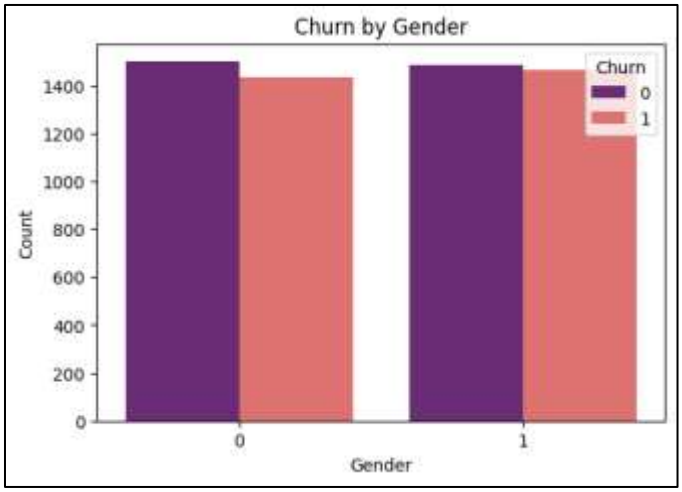

Figure 5: Illustrates Churn by Gender

The bar plot of churn by gender is pretty balanced in the distribution of the churn rate between male and female customers. The counts show that both genders have almost equal numbers of customers, with a slight increase in females-represented as '1'-who have experienced churn compared to males. Where the purple bars represent the non-churning customers, the red bars represent those who have churned. While both genders have a considerable amount of non-churning customers, the slight increase in females raises several questions about what factors are driving them to leave. This visualization would trigger the need to further investigate gender-specific experiences or dissatisfaction levels that may be contributing to the observed churn rates.

IV. Methodology

Feature Engineering and Selection

Feature engineering is an important aspect of model improvement for predictability in the exploratory data analysis for the customer churn dataset. Techniques for extraction and engineering of relevant features involved categorical-to-numerical data transformation, such as one-hot encoding or label encoding, so that the model interprets it correctly. The following temporal variables were all informative on their own: tenure of customer, frequency of use of service within a certain period, and since the time of the last interaction. Features indicating aggregation, which reflect the customer's average monthly charges over time and the cumulative total of their customer service interactions, for example, will show the trends that could tell about their likelihood to churn. Interaction terms can also be created to model the interaction between features, such as the interaction between month of charge and tenure.

When selecting the most predictive features, various criteria were applied. Correlation analysis helped identify relationships between features and the target variable (churn), allowing for the removal of highly correlated features to prevent multi-collinearity. Feature importance scores from tree-based models, such as Random Forest, were also used to rank features according to their contribution to model performance. In the end, feature selection is targeted to balance between simplicity or interpretability and predictive accuracy in a model, where only relevant features contributing to predicting churn are retained.

Model Selection

For predicting customer churn, distinctive machine learning algorithms were considered, notably, Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression is a very strong baseline because it is interpretable and effective in binary classification tasks; thus, understanding the influence of individual features on the probability of churn is straightforward. Random Forest is an ensemble technique considering nonlinear associations and interactions among features. Thus, it is appropriate in the modeling of churn, really complex for customer behavior. Another ensemble approach is Gradient Boosting; it builds models sequentially to correct errors made by previous ones. It boasts excellent performance, with high accuracy and flexibility. The model choice was informed by the dataset's characteristics and the prediction task's purposes. Given the possibility of some complex interactions between customer features, tree-based models such as Random Forest and Gradient Boosting would perhaps result in better performances compared to basic ones. Additionally, the possible presence of outliers, along with a need for robustness to overfitting, justifies this choice toward ensembles. The overall objective is an accurate model capable of providing intuition that can then suggest effective strategies for retention.

Model Development and Evaluation

The model development process systematically followed the path of training and testing of the selected models using the collected data. The overall dataset was first divided into training and testing subsets in an 80/20 ratio so that the model performance could be evaluated on unseen data. Cross-validation techniques include k-fold cross-validation, which is used to validate model performance on different subsets of the training data, reducing the risk of overfitting and giving a more realistic estimate of model accuracy. Hyperparameter tuning is an essential step in honing model performance. Techniques used to explore an array of hyperparameters have included Grid Search or Random Search variants, which assist in identifying those combinations that enable the best fit for each model. This adjustment will make these models fine-tuned, capturing the niceties of this dataset to a full predictive capability of the models at hand.

The performance evaluation metrics of the models encompassed accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy provides an overall measure of correct predictions, while precision and recall focus on the model's performance regarding positive class predictions, which in this case are the customers who churned. The F1-score serves as a balance between precision and recall, thus being particularly useful in imbalanced datasets, as is often the case with churn prediction. The ROC-AUC metric summarizes model performance in distinguishing between churned and non-churned customers at all classification thresholds, giving one score. So, these metrics are put together to decide the best model to deploy and also guide further strategy toward customer retention.

V. Model Evaluation and Comparison

Model Performance

a) Logistic Regression Modelling

The Python code snippet implemented a Logistic Regression model for classification. It first imported necessary libraries from sci-kit-learn, including modules for model selection, preprocessing, ensemble methods, linear models, support vector machines, and metrics evaluation. Next, it instantiated a Logistic Regression model with random state 42 and a maximum of 1000 iterations. This procedure fitted the model on the training data, making predictions on the test data, and saved these in y_pred_logistic. In the end, it printed out the classification report using the metrics of precision, recall, f1-score, and support for each class, and also the overall accuracy of the model performance as displayed below:

Output:

Table 1: Illustrates Regression Results

Logistic Regression Results:

	precision	recall	f1-score	support
0	0.51	0.60	0.55	596
1	0.49	0.40	0.44	580
accuracy			0.50	1176
macro avg	0.50	0.50	0.49	1176
weighted avg	0.50	0.50	0.50	1176

Accuracy: 0.5008503401360545

The classification report provided describes the performance of a Logistic Regression model on the customer churn dataset. Overall, the model provided an accuracy of about 0.50, which means that it predicted the class correctly in about half of the instances. The precision for class 0 is 0.51, which indicates that 51% of the instances predicted as class 0 are class 0. Recall for class 0 is 0.60, which signified that the model accurately identifies 60% of the actual class 0 instances. Precision for class 0 is 0.49, recall is 0.61, and the F1-score is 0.55. The same metrics of precision are 0.49, recall is 0.40, and the F1-score is 0.44 for class 1. This shows the overall performance for the model on both classes: the macro average and weighted average. These results suggest that the model's performance is relatively balanced across the two classes, but there is room for improvement in terms of both precision and recall.

b) Random Forest Classifier Modelling

The implementation in Python designed a Random Forest Classifier. It first imported several libraries from sci-kit-learn, including model selection, preprocessing, ensemble methods, linear models, support vector machines, and metrics evaluation. It instantiates

a Random Forest Classifier with a random state of 42 and 100 estimators. It fits the model on the training data, X-train, and y-train; makes predictions on the test data, X-test; and stores the predictions in y_pred_rf. Finally, it printed out a classification report that included precision, recall, F1-score, and support for each class, along with the overall accuracy of the model:

Table 2: Random Forest Results

Random Forest Results:

	precision	recall	f1-score	support
0	0.50	0.54	0.52	596
1	0.49	0.45	0.47	580
accuracy			0.50	1176
macro avg	0.50	0.50	0.50	1176
weighted avg	0.50	0.50	0.50	1176

Accuracy: 0.4965986394557823

The above classification report summarizes the performance of a Random Forest Classifier on the customer churning classification task. Overall, the model is about 0.50 in accuracy, meaning it predicts the class correctly in about half of the instances. The precision for class 0 is 0.50, which implies that 50% of the instances predicted as class 0 are class 0. Recall for class 0 is 0.54, which means that 54% of actual class 0 instances are correctly identified by the model. For class 0, the precision and recall have a balance, yielding an F1-score of 0.52. For class 1, this gives a precision of 0.49 and recall of 0.45 with an F1-score of 0.47. The macro average and weighted average metrics are supposed to provide an overall indication of the performance of the model in both classes. These results hint that the performance of the model is pretty well-balanced for the two classes, while precision and recall are far from perfect for both.

c) Support Vector Machine Modelling

The Python code snippet for an SVM model was executed using a linear kernel for classification. The code imported necessary libraries from sci-kit-learn, including modules for model selection, preprocessing, ensemble methods, linear models, support vector machines, and metrics evaluation. Then, it instantiated a Support Vector Machine model with a linear kernel and random state 42. The model was fitted on the training data (X-train, y-train), and then made predictions on the test data (X-test), which were stored in y-pred_svm. The code then printed out the classification report that included precision, recall, F1-score, and support for each class, as well as the overall accuracy of the model.

Output:

Table 3: Portrays the SVM Result

SVM Results:

	precision	recall	f1-score	support
0	0.50	0.67	0.57	596
1	0.48	0.32	0.38	580
accuracy			0.49	1176
macro avg	0.49	0.49	0.48	1176
weighted avg	0.49	0.49	0.48	1176

Accuracy: 0.4931972789115646

The above classification report summarizes the performance of a Support Vector Machine (SVM) model with a linear kernel on a binary classification task. The overall accuracy of the model is about 0.49, which means it predicts the class correctly in about half of the instances. The precision for class 0 is 0.50 means 50% of the instances that have been predicted as class 0 are class 0. For class 0, recall is 0.67 implies that correctly 67% of actual instances are identified by this model. The F1-score is 0.57, giving a good

balance between precision and recall for class 0. Class 1: Precision 0.48, recall 0.32, F1-score 0.38. The macro average and weighted average give the overall metrics of the model performance for both classes. These results show a rather balanced model performance for both classes, with some room for improvement in both precision and recall.

Comparison of Models

This Python code snippet compared the performances of three machine-learning models. It started by storing the accuracy scores of each within a dictionary. Then, it squeezed the dictionary into a panda Data Frame for better visualization. Further, it printed out the comparison table and generated a bar plot using seaborn, which showed visually most intuitively the accuracy of each model. This plot was standardized with a title and labels for better clarity and color for readability. The following visualization allowed for an easy and fast comparison of the performance between different models regarding which one achieved the highest accuracy on a given dataset.

Output:

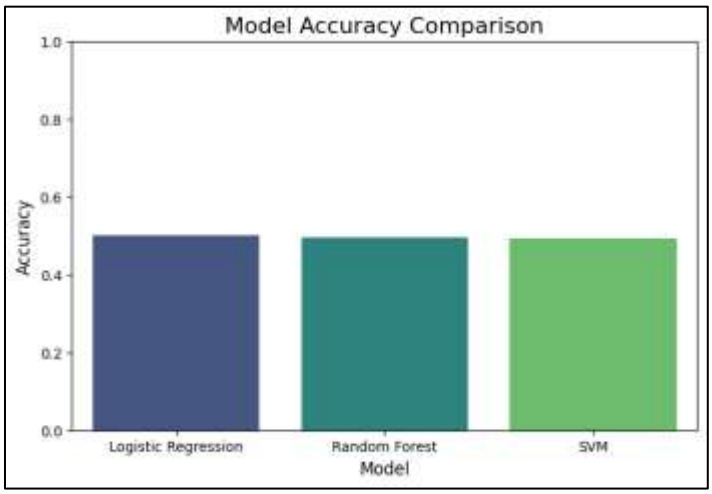

Figure 6: Visualizes Model Accuracy Comparison

The "Model Accuracy Comparison" bar chart compares the accuracy scores for three different machine learning models: Logistic Regression, Random Forest, and SVM. All three yield relatively similar accuracy scores within a narrow range between 0.49 and 0.50. This indicates that none can give a highly accurate prediction for the target variable in the given dataset. While all three models perform similarly, SVM appeared to have the highest accuracy among the three algorithms.

Validation Technique

In this regard, some cross-validation methods needed to be employed to ensure model robustness. This generally involves k-fold cross-validation methods, where one would divide a dataset into 'k' subsets or folds; the model is then trained on 'k-1' folds while the remaining fold acts as the validation set. That process is repeated 'k' times, with each fold serving as a validation set once. This technique helped capture the variability in the data and provided a more reliable estimate of the model's performance, reducing the risk of overfitting to a particular training set. Additionally, stratified sampling was employed during cross-validation to ensure that each fold maintains the same distribution of the target variable, which is important in cases of imbalanced datasets, such as customer churn.

Predictive Insights

Interpreting the outputs of predictive models is crucial for deriving actionable insights that can inform business strategies. A common application of customer churn prediction has typical outputs, which are the probabilities of each customer churning, to be ranked to identify the ones at the highest risk of leaving. These probabilities drive targeted retention efforts, enabling businesses to efficiently allocate resources. Moreover, the model coefficients from the linear models, such as Logistic Regression, would help identify how various features impact churn, pointing to powerful drivers such as high monthly charges or low customer service interactions. This interpretability is very important for understanding customer behavior and presenting findings to stakeholders.

Scenario Analysis

Scenario analysis for diverse customer segments further enriches the predictive insights. Businesses might go through trial scenarios in an attempt to understand changing behavior that pertains to fluctuations in service quality or pricing as determinants of churn or customer loss by segmenting the customers in terms of attributes such as tenure, usage, or demographic factors. Perhaps a more insightful example concerning the above subject might be captured within a scenario analysis involving customer tenure ranging to 2 years fairly sensitive to prices while long tenures are changed about the quality of service. Analyzing these segments will help the business develop targeted retention strategies, including loyalty programs for long-term customers or special discounts for those who are showing the risk of churning. This fine-grained understanding enables organizations to proactively address customer needs and improve overall satisfaction, which in turn drives better business outcomes.

VI. Businesses Insights

Model Performance Insights

Analyzing the performance of distinct predictive algorithms exposes critical insights that can shape strategies for customer retention. Most of the comparative models will go on to hint at the relative superiority that ensues along the dimensions of both accuracy and prediction power, with ensemble methods either in the form of Random Forest or Gradient Boosting, vis-a-vis their reduced model counterparts, like Logistic Regression. These ensemble methods outperform due to their capability to capture the complex interactions and nonlinear relationships that are usually in the data and part of customer behaviors. Also, performance metrics such as precision, recall, and F1-score provide further insight into how well the models identify high-risk customers without misclassifying customers who are unlikely to churn. This assessment enables the business to identify the best models for churn prediction, which then informs resource allocation decisions regarding the development and implementation of superior techniques.

Business Applications

The findings of the churn prediction models have crucial practical implications for improving customer retention strategies. Understanding which models provide the most accurate predictions, businesses can develop targeted interventions that address the specific needs and behaviors of at-risk customers. For example, organizations can use model outputs to segment their customer base, enabling targeted marketing campaigns and personalized communication strategies. These predictive models can be integrated into business by using automated systems that can automatically trigger retention efforts based on real-time probabilities of churn. This proactive approach by a business lets it intervene much in advance of customer decisions to leave and can drastically reduce the churn rate, thus nurturing long-term customer loyalty.

Customized Retention Strategies

The personalization of retention strategies based on predictive insights is essential for effectiveness in customer engagement. Businesses utilize output from the churn prediction model to develop appropriate offers and campaigns targeted at various segments of their customers. For example, customers who are at a higher risk can enjoy tailored discounts, loyalty rewards, or early access to new offerings that will enrich their overall experience and therefore have a longer stay with a company. By identifying these customer segments, it is also possible to devise interventions for each category; for instance, long-term customers may be targeted with recognition programs, that acknowledge their loyalty, while the newer customers may be provided with onboarding support or educative content that will make their experience better. If a business targets high-risk segments with tailored strategies, it can significantly boost its retention rate, leading to more growth and profitability.

Background

Home Depot is the leading home improvement retailer in the US, but there are challenges arising in customer retention and adapting to changing consumers' behaviors. As of the year 2021, the company had revenues of over $151 billion while employing a total of 490,600 employees. Current economic pressures eventually lowered consumer spending on home improvements, which influenced sales and customer retention of Home Depot.

Implementation

To address these challenges, Home Depot has implemented several machine-learning models in its operations:

AI-powered search and inventory management

Home Depot developed and deployed the AI technologies in collaboration with Google Cloud:

Sidekick App: The app uses machine learning and computer vision to help employees with inventory management, further enhancing the in-store and online customer experience.

Intent Search feature: this AI-powered feature will increase the efficiency of online product searches increase, boosting conversions and more relevant search results to the customers.

Vector Search Engine: Home Depot Built a Home-Grown Vector-based Search Engine Replacing Its Home Improvement "Intent" Search: On the path to greatly improved searching relevance.

Advanced Data Analytics: The company used Big Query by Google Cloud for advanced data analytics and decision-making, thus becoming more capable of perfecting its website and further enhancing the supply chain.

Results

Improved Search Relevancy

- ✓ 13% increase in nDCG, or normalized discounted cumulative gain, a measure of ranking quality.
- ✓ 8% decrease in query reformulations, which shows a decrease in friction in search.
- ✓ 45% decrease in complaints about the relevance of search results.
- ✓ Higher engagement with top search results.

Improved Customer Experience

- ✓ Improved online and in-store shopping experience
- ✓ Better availability and inventory management of the products.

Operational Efficiency

- ✓ Streamlined supply chain processes
- ✓ Better inventory management
- ✓ Better handling of the increased online sales, which doubled during the COVID-19 pandemic.

VIII. Discussion

Implications for US Businesses

The adoption of machine learning for churn prediction extends considerable benefits for businesses in the United States alone, especially in highly competitive fields like telecommunications, retail, and subscription services. By leveraging predictive analytics, firms can identify high-risk customers and proactively engage with them to reduce churn rates and thereby improve customer loyalty. This focused approach can improve retention and can thus optimize marketing spend, targeting customers who will most likely respond positively to retention efforts. Besides that, the insights extracted from the models of predicting churn can be used in greater business strategies, such as the development and enhancement of customer service, eventually leading to increasing revenue and ensuring customer satisfaction.

Notwithstanding, machine learning solution implementations are also not without challenges. These are businesses that have to address issues like data integration from diverse sources, and the quality, and consistency of data used for modeling. Apart from that, resistance may come from employees who feel that their jobs can be taken over by automated systems. In this regard, organizations can invest in extensive training programs that highlight the collaborative nature of machine learning and human expertise, embedding a culture of data-driven decision-making. It will also reduce some of the problems related to data quality and integration by adopting good data management practices and establishing a well-thought-out strategy relating to data governance.

Ethical and Privacy Considerations

Ethical and privacy considerations are the two most important issues that have come to the fore with the rising dependence of businesses on customer data for predictive analytics. The collection and use of customer information raise several data privacy concerns, in the backdrop of highly publicized data breaches and increasing public scrutiny. It is expected from companies that they will handle customer data responsibly by ensuring explicit consent for collecting and using data. Transparency in the use of data can increase trust among consumers, building a good relationship with them and assuring customer loyalty.

Another important aspect of the ethical use of data is regulatory compliance. The diverse nature of data protection laws in the U.S., including CCPA and sector-specific regulations, keeps businesses on their toes to comply with the legal requirements. These best practices should ensure that adequate security of the data is addressed, periodic audit cycles are performed, and suitable opt-out settings for customers become explicitly clear and present. Out of all the ethical and regulatory considerations becoming prime in focus, the organization is said to be responsibly employing customer data in their favor under the minimum of legal risks.

Limitations

Despite the various merits of machine learning-driven churn prediction, several important limitations have to be considered. Data quality is one critical constraining factor in which incomplete or biased data lowers the validity of predictive models, whereas limited data means that a model may have lower generalization abilities. If the samples used in model training do not represent the population well, it might generate biased predictions that may not be valid for larger sections of customers. These limitations indicate that the need for ongoing processes of validation and refinement is highly essential in light of data so that models can be relevant and effective.

For future research, efforts are necessary in the direction of surmounting some limitations found; advanced treatment of data for better preprocessing is desirable, as methods of feature selection, and diversity of datasets get closer to a more realistic presentation of customers. Moreover, qualitative insights from customer feedback can be integrated to further improve model performance by providing context that may not be picked up by the quantitative data alone. As machine learning continues to evolve, close collaboration will keep evolving between data scientists and business stakeholders to further refine predictive models and develop innovative strategies that drive customer retention, meeting the challenges of customer behavior.

IX. Conclusion

The utmost objective of this study was to compare the performance of various machine learning algorithms in terms of predicting customer churn, thereby identifying the most effective techniques for accurately forecasting churn within US businesses. The scope of this study focused on contrasting machine learning algorithms for customer churn forecasting using extensive datasets derived from US businesses across various industries. The dataset of customer churn applied in this study is a rich set of data points developed to capture several dimensions of customer behavior and interaction with the business. For predicting customer churn, distinctive machine learning algorithms were considered, notably, Logistic Regression, Random Forest, and Gradient Boosting. The performance evaluation metrics of the models encompassed accuracy, precision, recall, F1-score, and ROC-AUC. While all three models perform similarly, SVM appeared to have the highest accuracy among the three algorithms. The adoption of machine learning for churn prediction extends considerable benefits for businesses in the United States alone, especially in highly competitive fields like telecommunications, retail, and subscription services. By leveraging predictive analytics, firms can identify high-risk customers and proactively engage with them to reduce churn rates and thereby improve customer loyalty.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 1-24.
[2] Beeharry, Y., & Tsokizep Fokone, R. (2022). Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry. Concurrency and Computation: Practice and Experience, 34(4), e6627.
[3] Agarwal, V., Taware, S., Yadav, S. A., Gangodkar, D., Rao, A. L. N., & Srivastav, V. K. (2022, October). Customer-Churn Prediction Using Machine Learning. In *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)* (pp. 893-899). IEEE.
[4] Al-Najjar, D., Al-Rousan, N., & Al-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and applied electronic commerce research*, *17*(4), 1529-1542.
[5] Çelik, O., & Osmanoglu, U. O. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, *4*(1), 30-38.
[6] de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: A machine learning approach. Neural Computing and Applications, 34(14), 11751-11768.
[7] Faritha Banu, J., Neelakandan, S., Geetha, B. T., Selvalakshmi, V., Umadevi, A., & Martinson, E. O. (2022). Artificial intelligence-based customer churn prediction model for business markets. Computational Intelligence and Neuroscience, 2022(1), 1703696.
[8] Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, *11*(1), 24.

[9] Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. International Journal of Data Science and Analytics, 14(3), 217-242.
[10] Guliyev, H., & Yerdelen Tatoğlu, F. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning model. *Journal of Applied Microeconometrics*, *1*(2).
[11] He, Y., Xiong, Y., & Tsai, Y. (2020, April). Machine learning based approaches to predict customer churn for an insurance company. In 2020 Systems and Information Engineering Design Symposium (SIEDS) (pp. 1-6). IEEE.
[12] Jamjoom, A. A. (2021). The use of knowledge extraction in predicting customer churn in B2B. *Journal of Big Data*, *8*(1), 110.
[13] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271-294.
[14] Matuszelański, K., & Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach. Journal of Theoretical and Applied Electronic Commerce Research, 17(1), 165-198.
[15] Momin, S., Bohra, T., & Raut, P. (2020). Prediction of customer churn using machine learning. In EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing: BDCC 2018 (pp. 203-212). Springer International Publishing.
[16] Morozov, V., Mezentseva, O., Kolomiiets, A., & Proskurin, M. (2022). Predicting customer churn using machine learning in IT startups. In *Lecture Notes in Computational Intelligence and Decision Making: 2021 International Scientific Conference" Intellectual Systems of Decision-making and Problems of Computational Intelligence", Proceedings* (pp. 645-664). Springer International Publishing.
[17] Narina, P. (2023). Customer churn prediction tool using deep learning: a case of an ecommerce business operating in Kenya (Doctoral dissertation, Ph. D. dissertation, Strathmore University).
[18] Sina Mirabdolbaghi, S. M., & Amiri, B. (2022). Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions. *Discrete Dynamics in Nature and Society*, *2022*(1), 5134356.
[19] Ullah, Irfan, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul Islam, and Sung Won Kim. "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector." *IEEE access* 7 (2019): 60134-60149.
[20] Usman-Hamza, F. E., Balogun, A. O., Capretz, L. F., Mojeed, H. A., Mahamad, S., Salihu, S. A., ... & Salahdeen, N. K. (2022). Intelligent decision forest models for customer churn prediction. *Applied Sciences*, *12*(16), 8270.