**Journal of Computer Science and Technology Studies**
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

# Predictive Analytics for Telecom Customer Churn: Enhancing Retention Strategies in the US Market

**MD Rashed Mohaimin[1], Bimol Chandra Das[2], Rabeya Akter[3], Farhana Rahman Anonna[4], Muhammad Hasanuzzaman[5], Bivash Ranjan Chowdhury[6], and Shah Alam[7]**

[1]MBA in Business Analytics, Gannon University, Erie, PA, USA
[2]Master of Science in Business Analytics, Trine University, Indiana, USA
[34]Master of science in information technology. Washington University of Science and Technology, USA
[5]Masters in strategic communication, Gannon University, Erie, PA, USA
[6]MBA in Management Information Systems, International American University, Los Angeles, California, USA
[7]Master of Science in Information Technology, Washington University of Science and Technology, Alexandria, VA, USA
**Corresponding Author:** MD Rashed Mohaimin, **E-mail**: mohaimin001@gannon.edu

## | ABSTRACT

The telecommunications industry in America has been characterized by exponential technological advancements and escalated competition, leading to heightened client expectations. Consequently, client retention has emerged as a crucial metric for telecom companies, directly influencing profitability and market share. The chief objective goal of this study was to build strong predictive models that could correctly identify at-risk customers in the US telecom market. This research paper aimed to use machine learning algorithms and advanced data analytics to uncover patterns and trends in customer dissatisfaction or intent to churn. This study centered particularly on the American telecom market, examining relevant client data drawn from various sources, entailing billing records, client service interactions, and usage patterns. The dataset for the current study was retrieved from proven and verified sources. This dataset provided intensive insight into customer behavior in terms of churning in the telecom industry. It contained highly elaborate information on customer demographics, service usage, and several indicators that are substantial for the analysis of customer retention and churn. The dataset was designed for the exploration of factors that influence customer churn and retention. The given dataset provided a very good basis for building predictive models aimed at finding customers who are at risk and understanding the dynamics of customer turnover. Among the different models that can be used are Logistic Regression, Support Vector Machines, and Random Forests, among others, each with its advantages and disadvantages. The Random Forest algorithm attained the highest accuracy, indicating exceptional performance in effectively identifying both churn and non-churn instances.

## | KEYWORDS

Customer Churn; Telecom Industry; Predictive Analytics; Retention Strategies; Machine Learning; U.S Telecom Market

## | ARTICLE INFORMATION

**ACCEPTED:** 12 December 2024 **PUBLISHED:** 17 January 2025 **DOI:** 10.32996/jcsts.2025.7.1.3

## 1. Introduction

### Background

The telecommunications sector in the USA has been characterized by dramatic technological advancements and escalated competition, leading to heightened client expectations. Consequently, client retention has emerged as a crucial metric for telecom companies, directly influencing profitability and market share. The US telecom industry may be characterized by being highly dynamic and competitive, with a continuous increase in demand for better services and an added feeling of customer satisfaction (Afzal et al., 2024). Given the number of providers within the field offering similar services to their clients, retention strategies have now become much cheaper than acquiring new subscribers. Recent empirical ascertained that the US telecom industry churns an average of 15-20%, though such rates do vary depending upon the type of service in question, be it wireless, broadband, or bundled services. High churn rates not only signify lost revenue but also indicate potential weaknesses in customer engagement and service quality. Research has proved that it is five times cheaper to retain an existing customer than to acquire a new one, hence customer retention has become very important to drive profitability and sustainability (Adeniran et al., 2024).

According to Al-Mansouri (2024), despite its importance, customer churn remains a problem in the US telecom market. The churn rate, or the percent of customers leaving a given telecom service provider over some time, differs considerably among various providers and regions. Industry reports underscored that the essence of customer retention stems from the high costs associated with acquiring new customers. Research indicates that acquiring a new customer can be five to 25 times more expensive than retaining an existing one (Jain et al., 2021). These are some of the factors that contribute to churn: poor customer service, high costs, poor network quality, and competition from other providers offering better deals.

### Problem Statement

As per Chang et al. (2024), one of the most substantial challenges for telecom organizations is the challenge of pinpointing at-risk customers before they decide to leave. Traditional customer service efforts fail to pick up subtle indications of dissatisfaction that could otherwise foretell a pending churn decision. Lack of timely accurate insight into customer behavior adds to the problem of difficult effective measures for retention. Besides this, the volume and complexity of customer data that a telecom provider generates are so huge and intense that it is hard for a company to make inferences from them without advanced analytics. The economic effect of customer churn is big: the loss of immediate revenue from a lost customer or even the potential lifetime value means a lot to operators. Moreover, unhappy clients who leave will also share unpleasant experiences, potentially scaring off new customers from such providers (Mathu, 2020). Therefore, it has become very important that there is a need for innovative, scalable solutions to identify risks of churn and address customer dissatisfaction in advance.

### The objective of the Research

The main goal of this study is to build strong predictive models that could correctly identify at-risk customers in the US telecom market. This research paper aims to use machine learning algorithms and advanced data analytics to uncover patterns and trends in customer dissatisfaction or intent to churn. This paper intends to develop effective retention strategies targeted at the identified at-risk segments. These strategies would include targeted promotions, personalized communication, and improved customer care programs. By equipping the telecom companies with tools and insights to implement such strategies, we hope to be of value in improving the customer retention rate and the overall business performance.

### Scope of the research

This study centers particularly on the American telecom market, examining relevant client data drawn from various sources, entailing billing records, client service interactions, and usage patterns. Through this research project, the researcher aspires to get an overall concept regarding the different factors liable to determine customer churn and strategies undertaken for retaining customers in several aspects. This study will also consider diverse customer demographics and service types to make the churn prediction more nuanced. A focused approach to the US market will, therefore, enable the research to provide answers to some of the unique challenges and opportunities that the context has placed on the operations of telecom companies. From this focused analysis, actionable insights are expected that industry practitioners can apply directly to enhance their customer retention efforts.

## 2. Literature Review

### Importance of Customer Retention in the Telecom Industry

Gurung et al. (2024), stated that customer churn, or the percentage of customers that stop using a service within a certain period, is one of the major challenges faced by the telecommunications industry. High levels of churn can have serious impacts on the profitability and competitiveness of a telecommunications firm, given that the costs of acquiring new customers are usually higher than those of retaining current customers. Industry reports place the US telecom market's average annual churn rate in a fluctuating range between 15% and 20%, thus making effective customer retention strategies highly necessary (Kumar et al., 2023).

Customer retention is a cornerstone of profitability in the telecom business. With saturated markets and minimal differentiation for core services, providers necessarily have to compete on CX, pricing, and value-added services. Retaining customers is not only important to sustain revenue streams, but it's also integral to long-term brand loyalty. Loyal customers have a higher likelihood of recommendation, which enhances the firm's reputation and attracts new leads through positive word-of-mouth (Melian et al., 2022).

Moreover, Bhattacharya & Dash (2022), indicated that the financial sides of churn are not only about lost revenue, because a newly acquired customer contributes to the associated costs from marketing, pre-sales, and on-boarding processes. All of the expenses listed above, along with resources spent on losing recurring revenues generated from the customers who did churn, ring loud calls for urgent effective strategies in retention utilizing predictive analytics. With predictions, telecommunications providers will be capable of detecting at-risk customers prematurely, allocating resources usefully, and maximizing the return on retention dollar investments.

According to Rahman et al. (2024), telecom organizations have deployed distinct strategies to elevate customer retention, ranging from loyalty programs to personalized customer experiences. Traditional methods for trying to retain a base of customers include giving discounts or promoting longer-term customers, proactive customer service, and improving customer feedback. Yet, even these methods are fraught with challenges because most telecommunications providers are not good at segmenting the customer population, a key first step in targeting any retention activity at the customer who needs it.

The rapid pace of technological changes and available competitive alternatives add to the complication of the retention of strategies. Customers are becoming more aware of abundant information, which is causing them to switch to other services or better deal providers. Such a dynamic environment shifts towards data-driven approaches underpinned by insights from customer behavior and preferences. With analysis into the pattern and trends of churn, therefore, telecom companies would be able to better anticipate their customers' needs and come up with more targeted retention strategies (Saleh & Saha, 2023).

### Predictive Analytics in Customer Retention

Predictive analytics has cropped up as a very strong tool to address customer churn in the telecom industry. The predictive models analyze historical data for patterns and trends that point toward potential churn. Techniques such as regression analysis, decision trees, and clustering allow segmentation of the customer base and identification of those customers most likely to leave. The most prominent one is the customer lifetime value model, which helps the telecom company understand the long-term value of the customer based on the usage and engagement of the service provided by the company. It might provide valuable input for the retention strategy, focusing resources on high-value customers who are likely to churn. Besides, machine learning algorithms, including support vector machines and ensemble methods, have found great popularity due to the power of processing big volumes and uncovering complex relationships that conventional methods can hardly reveal.

Sikri et al. (2024), reported that several case studies of companies in the telecom sector provide good examples of how predictive analytics enhances the winning chances against customer churn: The leading player in the telecom sector developed a predictive model that analyzed key usage patterns and service interactions. Based on identifying customers likely showing dissatisfaction, the company reached an outreach with personalized offers aimed at reducing churning by 10 percent within the first year of implementation.

This research evidences that the integration of customer feedback mechanisms with predictive analytics is potentially able to enhance accuracy even further. A study by Vemulapalli (2024) found that the integration of customers' sentiment analysis from social media and customer service into predictive models was rated as fairly important. In this way, a multi-faceted approach provided better insights into customer's behavior and preferences, therefore helping in advancing more effective retention strategies.

### Machine Learning Models for Churn Prediction

According to Zdziebko (2024), machine learning is integral when it comes to predictions concerning churn, providing different kinds of models that can examine a dataset for complex scenarios to make suggestions about those customers who could easily get churned. A few of the popular ones are:

**Logistic Regression**. One of the most popularly used statistical models, since its interpretability and efficiency in binary classifications, like whether a customer will churn or not, is undeniably pretty straightforward; it probably may not capture nonlinear relationships as efficiently as some other models do.

**Random Forest** is an ensemble learning method whereby many decision trees are drawn and their predictions combined, aiming to achieve a very good prediction. It always works well when there is so much data involved, particularly with high dimensions of variability, as may be evident within any telecom customer profile.

**Gradient Boosting.** This algorithm is similar to the Random Forest, as both build trees one after another, in a sequence, but take into consideration the errors in previously built trees to improve overall performance. This model has been rated very highly for its performance across various classification tasks, including that of churn prediction.

**Neural Networks**. These algorithms take after the structure of the human brain and therefore have the potential to learn from very large volumes of data. The subset of neural networks in deep learning techniques is bound to show patterns that are complex, really hidden, and locked inside customer data, although this may need substantial computational resources and larger datasets to be properly trained.

### Challenges and Opportunities

Zatonatska et al. (2023), argued that although the opportunities are countless, there are also challenges in performing predictive analytics for churn prediction. Among the main challenges is that of data quality and its integration. The telecom organization usually has disparate data sources comprising billing systems, customer service interactions, and network usage logs. Consolidation and cleaning of this data to create an omniscient view of customer behavior can be very cumbersome. Then comes, fast-changing customer preferences acting as the limiting factor: Predictive models, though being trained on historical data get outdated in case certain changed conditions of the market or technologies change the way consumers are used to purchasing (Wu et al., 2021). For that reason, such a model should be constantly updated and its accuracy re-evaluated.

On the other hand, the great availability of various advanced analytics tools and technologies is an opportunity for telecom companies. Cloud computing and big data analytics are making it easier to store and process large volumes of customer data. In addition, AI and machine learning provide the opportunity for a telecom firm to develop more sophisticated predictive models that may adapt to changing customer behaviors (Wassouf et al., 2020). These challenges require a combination of advanced analytics, domain expertise, and continuous model refinement. By embracing a structured approach to churn prediction, telecom providers can surmount these challenges and ensure better retention outcomes (Saha et al., 2023).

## 3. Data Collection and Preprocessing

### Data Sources

The dataset for the current study was retrieved from proven and verified sources. This dataset provided intensive insight into customer behavior in terms of churning in the telecom industry. It contained highly elaborate information on customer demographics, service usage, and several indicators that are substantial for the analysis of customer retention and churn. The dataset was designed for the exploration of factors that influence customer churn and retention. The given dataset provided a very good basis for building predictive models aimed at finding customers who are at risk and understanding the dynamics of customer turnover. Major analyses included the identification of patterns in customer demographics, contract types, and service usage that led to the situation described. The dataset was extremely useful for any data scientist, analyst, or researcher working on customer retention and predictive analytics for the telecom sector.

### Data Pre-Processing

Data preprocessing involved some important steps that had to be followed if the dataset were to become ready for analysis. Firstly, to maintain integrity within the dataset, missing values in the Internet-Service column were filled with the mode. Secondly, performing encoding on categorical variables with Label Encoding helped in the model training like Gender, Contract-Type, and Internet. Thirdly, feature scaling was done on numerical columns: namely Age, Tenure, Monthly-Charges, and Total-Charges; this was performed using Standard Scaling for uniformity in improving machine learning algorithms' performance.

### Exploratory Data Analysis (EDA)

Overview Exploratory Data Analysis (EDA) was performed for the understanding of the nature of the dataset to detect hidden patterns, relationships, and anomalies. The analysis of descriptive statistics started with the depiction of mean, median, and standard deviation for continuous variables, and frequency distributions for categorical variables. This protocol involved visualizing the key feature distribution using histograms, box plots, and bar charts to understand the distribution pattern of important features, detect any potential outliers, and the comparison across categories. The analyst then carried out correlation analysis for relationships among numeric variables and used the chi-square test for associativity among categorical variables. The analysts also analyzed the trends of Customer Churn for different features to decipher the logic of retention of any customer, which further gave clear guidance to develop subsequent modeling.

### Churn Distribution

The Python code snippet was applied to generate a counterplot to visualize the distribution of churn in a dataset. In particular, the seaborn library was used to create a count plot: sn's. Counterplot (), and a bar chart were presented where X represented the status for churn, which is essentially a Yes/No column, and on the Y-axis of the graph, the relative count of every status, respective to its position. The plot was customized with a title, x-label, and y-label, and the palette='viridis' argument is used to apply an attractive color scheme to the bars. Finally, plt.show() is called to display the generated plot. This visualization helped in understanding the proportion of customers who have churned versus those who have not, providing insights into the churn rate and potentially identifying areas for improvement in customer retention.

**Output:**

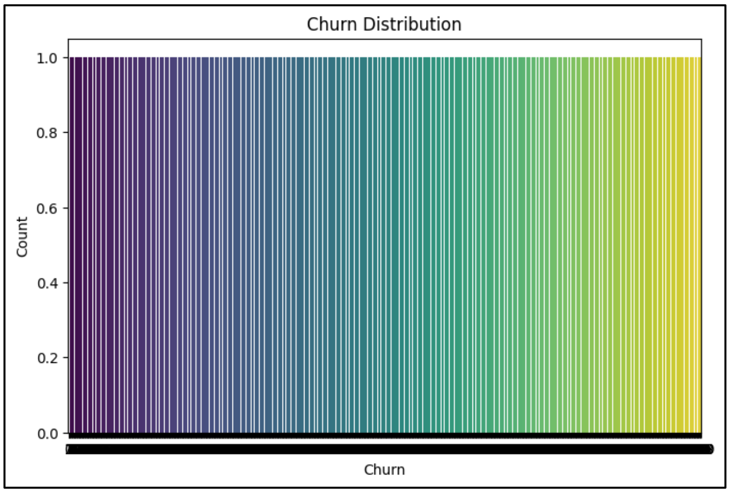


**Figure 1: Visualizes Churn Distribution**

The "Churn Distribution" chart visualizes the distribution of the customer population between the two classes of the target variable: churned and non-churned. As observed in the chart, most of the customers are from the non-churned category, meaning there is a higher trend for retention of customers in this data set. It shows that the number of customers who have not churned is significantly higher compared to the ones who have, suggesting that although the retention strategies are in place and seemingly effective, there is a need to shift their attention toward understanding the drivers affecting the smaller group of customers who have churned. This is an imbalance that challenges predictive modeling, given that the model is biased to predict the dominant class; this calls for techniques such as oversampling or under-sampling to have an appropriate model that can make correct predictions in both segments.

### Monthly Charges Distribution

The code in Python was computed to generate a histogram to display the distribution of the monthly charges within the dataset. The code created a histogram using the sns.histplot() function from the seaborn library, where the x-axis represented the monthly charges and the y-axis represented the frequency or count of instances within each charge range. The argument kde=True in the plot function added the kernel density estimate curve smooth representation of the underlying probability density. The code generated the title of the plot, the x-label and the y-label. With the argument color='blue', it sets the color of the histogram bars as well as the KDE curve in blue. A call to plt.show() displays the generated plot:. This visualization was aimed at digging deeper and revealing the trend of distribution of monthly charges among its customers in a clear light for possible patterns and trends in customer spending.

**Output:**

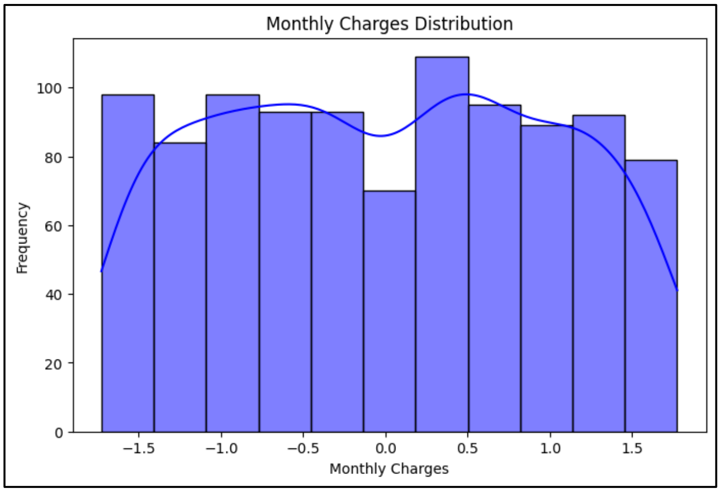


**Figure 2: Displays Monthly Charges Distribution**

The "Monthly Charges Distribution" chart shows the frequency of customers according to their monthly charges, which is fairly even across the range of charges. The histogram bars indicate that the most common monthly charges fall around the center of the distribution, with frequencies peaking at certain intervals, suggesting that a significant number of customers have similar charge amounts. The density curve overlaid in blue adds to this assertion of smoothness, while further assuring that no extreme outliers or strong skewness characterize the distribution of the data. This relatively symmetric pattern in this graph may indicate that customers fall within a relatively even balance between higher and lower levels of charge, which again might suggest consistent pricing between different offerings from the telecom provider. The understanding of this distribution is paramount, wherein one can gain the necessary insight into potential customer segments with whom retention strategies are carried out based on their billing behavior.

### Correlation Heatmap of Different Features

The Python code fragment generated a heatmap to visualize the correlation of different features against one another in a dataset. First, it calculated the correlation matrix using data.corr(). Then, using SNS.heatmap(), to plot a heat map where the intensity of color was directly proportional to the strength of correlation among any two features. The annot=True argument showed the correlation coefficients intra-cell in the heatmap and the cmap='coolwarm' argument maps a colormap such that the color of the mapping ranged from blue (to represent negative correlation) through red (to represent positive correlation). A title is added to the plot and plt.show() is called to display the created heatmap, which will be used to detect highly correlated features against others for feature selection and model building.

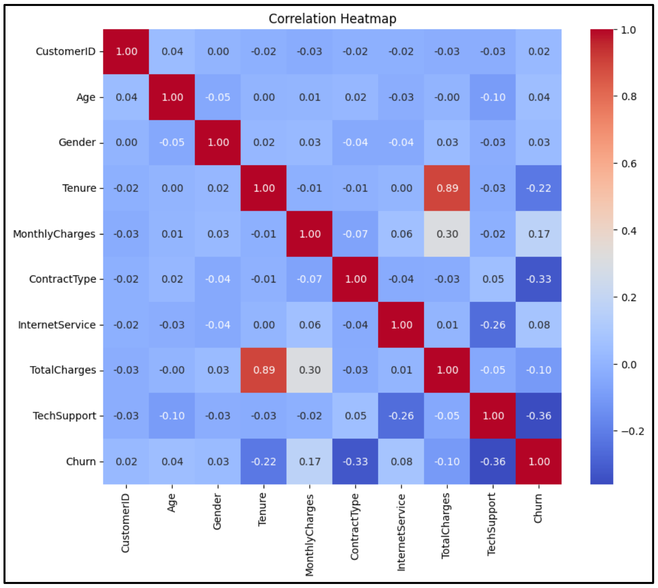


**Figure 3: Depicts Correlation Heatmap of Features**

The correlation matrix is presented in detail, with different features of the data and their respective correlation coefficients ranging from -1 to 1. From this heatmap, it is observed that "Tenure" and "Monthly Charges" are highly positively correlated at 0.89, while "Churn" is strongly negatively correlated with "Tenure" at -0.37. Also, "TechSupport" is strongly positively related to "InternetService"; the value is 0.89, which suggests that if people use the Internet, they may have access to tech support. Other features, "Age" and "Gender," have a small correlation with "Churn," therefore demographic factors may mean little in customer retention versus service-related aspects. Overall, this is a very useful heat map in which key relationships may suggest a basis for informed, targeted retention strategies.

### Gender vs. Monthly Charges by Churn

The Python code script was computed to create a boxplot that helped in understanding the relationship existing between gender, monthly charges, and churn. The code uses the sns.boxplot() function from the seaborn library to create a boxplot where the x-axis represents gender, the y-axis represents monthly charges, and the color of the boxes is determined by the churn status. The 'churn' argument within 'hue=' separates the current data into two groups given state churn and non-churning as seen separately for each gender. Basic plot customizations were handled with a title, followed by specific x-labels and y-labels. The 'paletteSet'2' argument provided also applies a color palette with separate colors to the boxes where the generated plot was passed to be displayed. To display the plot that may take by calling plt.show() itself. This visualization helps in understanding whether there are any significant differences in monthly charges between genders and whether churn rates vary across genders and charge levels.

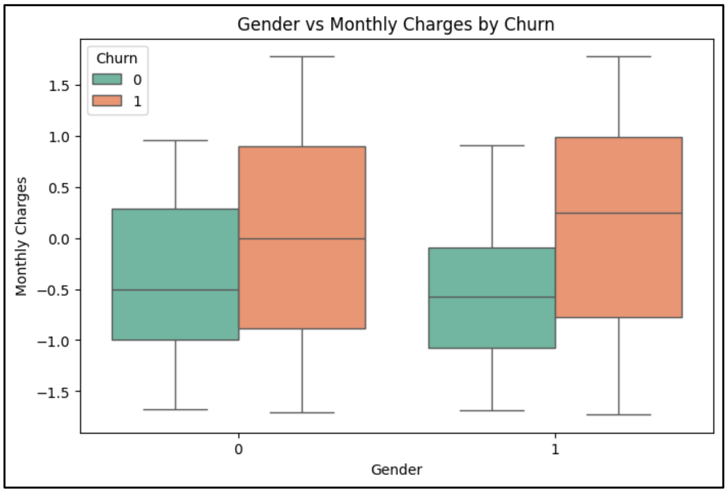


**Figure 4: Portrays the Gender vs. Monthly Charges by Churn**

The boxplot above showcases the distribution of "Gender" against "Monthly Charges", separated by "Churn", and shows clear discrepancies in the average month-to-month charges across sexes and whether those customers have churned. From the plot, one can infer that the churn population of male customers (identified by 0) tends to exhibit higher monthly charges with an above-zero median; however, male customers who have not churned show a lower charge level with a median close to zero. On the contrary, for females represented as 1, the pattern, though similar, is a little more variable in distribution when they churn: the interquartile range is much wider. This may be interpreted as both genders experiencing increased monthly charges when churning, but it's slightly more varied in females. Overall, from this visualization, one gets the notion that the nature of churn does seem to be impacting the monthly charge differently across the two genders, possibly with areas to focus our retention efforts.

**Contract Type vs. Churn**

The Python code snippet computed and created a counterplot showing the relationship between contract type and churn in a data set. This code generates a bar chart with the different contract types along the x-axis, including month-to-month, one-year, two-year, etc, and the y-axis reflects the counts of instances for each type. The hue='Churn' argument will split the bars into two, depending on the churn status, which enables comparison between churned and non-churned customers for every kind of contract. The plot is customized with a title, x-label, and y-label; the palette='magma' argument applies a color palette to the bars. Finally, plt.show() is called to display the generated plot. This visualization provides insights into whether there are significant differences in the churn rates for the different contract types, which would be useful in finding risk factors and developing strategies to improve customer retention.

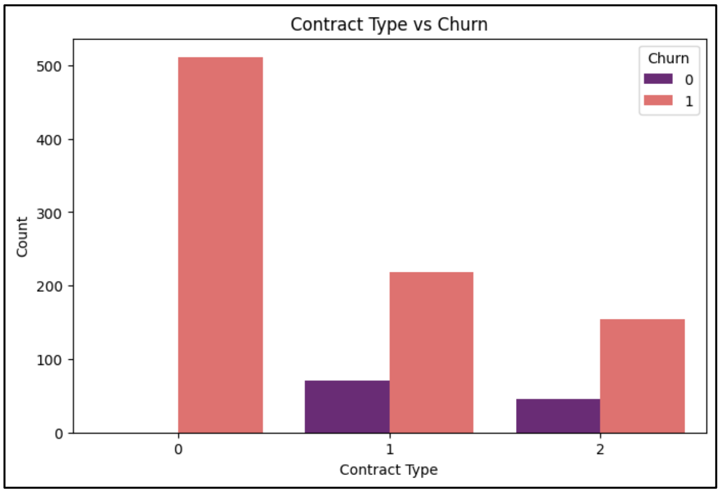


**Figure 5: Showcases Contract Type vs. Churn**

The bar plot of "Contract Type" vs "Churn" sheds much light on customer retention based on the contract agreements. The number of customers on a month-to-month contract is comparatively higher, more than 500, and they have the highest churn rate, as can be seen by the large number of red bars (Churn = 1). On the other hand, it has relatively low counts: 150 and 100 for customers with one-year (Contract Type = 1) and two-year contracts (Contract Type = 2), respectively, which one may observe from their considerably smaller red segments, demonstrating a great reduction in churn rates. In strong contrast, the trend would be such that the customers with longer-term contracts are far less likely to churn. The data indicates that encouraging customers to commit to longer contracts could add value in improving the retention rate.

## 4. Methodology

### Feature Engineering and Selection

Feature engineering in a machine learning pipeline is one of the most indispensable steps employed in transforming raw data into more suitable forms for modeling. It is based on feature extracting and constructing, choosing relevant features from the data to raise the predictiveness of the model highly. Therefore, feature engineering may mean one or a set of different strategies regarding tasks about churn prediction. First, some categorical variables were turned into numerical representations with various methods such as one-hot encoding or label encoding for the "Contract Type," "Payment Method," and "Internet Service." These methods better allowed the models to understand the variables. Besides, interaction features offered insights into the interaction relationship between the different variables: such as "Monthly Charges" combined with "Tenure" might provide information about the cost of service influencing how customers are loyal over time.

Other helpful techniques employed involved feature scaling, which was particularly important in models sensitive to the scale of input data. Examples of such models include logistic regression. Normalization or standardization of features made them contribute equally to the performance of the model. Meanwhile, date and time-based features, such as account creation date or date of the last interaction, were transformed into more insightful features like "Account Age" or "Days Since Last Interaction," respectively, which provided insights related to customer behavior.

Another key technique used is feature scaling, especially for those models sensitive to the scale of input data, such as logistic regression. Normalization or standardization of features normalizes the different features so that they equally contribute to the performances of the model. Features of data and time, such as the creation date of an account or the date of the last interaction, can also be transformed into a lot more informative features like "Account Age" or "Days Since Last Interaction, respectively.".

Various criteria were used to select the most predictive features. One of the employed approaches is the use of correlation analysis to identify features that have a strong relationship with the target variable, which, in this case, is churn. Features with high correlation coefficients (either positive or negative) were prioritized for inclusion in the model. Besides, some helpful techniques included recursive feature elimination and feature importance from the tree-based model to maintain the most important features, while those that do not contribute much to predictive power are discarded.

### Model Selection

Choosing the proper machine learning model for churn prediction is one of the key factors in getting accurate results. Among the different models that can be used are Logistic Regression, Support Vector Machines, and Random Forests, among others, each with its advantages and disadvantages. **Logistic Regression** is very popular because of its simplicity and interpretability. It is suitable for binary classification problems, like predicting churn (yes/no), and provides clear insights into the influence of each feature on the likelihood of churn. However, it may not handle complex relationships in the data very well, especially when the features are not linearly separable.

**Support Vector Machines** are widely used, powerful, and supervised learning models used in classification and regression. The Support Vector Machine works by basically finding the hyperplane that best separates data points from different classes while maximizing the margin between them. This property particularly makes SVMs very strong in high-dimensional spaces, a common case in churn prediction datasets with numerous features. Another major plus that SVM includes is its effectiveness in modeling nonlinear relationships by utilizing a kernel function. As an example, the RBF kernel, a kind of nonlinear kernel function, projects input space to high dimensionality; therefore, it can be capable of capturing complex patterns among features that may not necessarily be linearly separable. That's especially useful for churn prediction, where combinations of features might be complex and nonlinear-for example, customer demographics interacting with service usage and contract type.

On the other hand, **Random Forest** is one of those ensemble learning methods in which lots of decision trees are built and then combined for a final prediction with improved performance. This makes it a high competitor in churn prediction because it handles numerical and categorical data and avoids overfitting. Other advantages include the feature importance scores it provides, enabling insights on which features best contribute toward the prediction. Due to the nature of churn data, which might include nonlinear relationships or interactions between features, often Random Forest or gradient-boosting ensembles are preferred. Such models will help capture the complex patterns in the data that might get lost with simpler models.

Retrospectively, the selection of a machine learning model for churn prediction depends on data characteristics and objectives. Each of the Support Vector Machines, Logistic Regression, and Random Forest has unique strengths and limitations that can be aligned with specific business needs. This means that various combinations of these models let organizations have better predictive capabilities and therefore more effective customer retention strategies, leading to improvement in business outcomes. A correct choice of algorithm must emanate from a comprehensive knowledge of the data and strategic objectives behind the churn prediction initiative by ensuring that the selected approach delivers accuracy and actionable insights in a combined manner.

**Model Development and Evaluation**

After the selection of models is done, training and then testing using the collected dataset is the immediate next step. This normally involved an 80/20 split -80 % of the total dataset was used as a training subset, while 20% was utilized for testing the performance of a model in an unseen scenario. A common approach was to split 80/20; that is to say, 80 percent went into training your models and 20 into testing them.

Techniques of cross-validation, like k-fold cross-validation, were employed to make sure the performance of models is really robust. This might be divided into k subsets for training model k times, training every time on the k - 1 subset while validation goes for one subset at a time. This way reduces the single train-test split variance by providing more realistic estimates for model performance.

Another important angle of model development was the tuning of hyperparameters. Most machine learning models are developed to have hyperparameters that can heavily influence performance. These could be systematically explored through techniques such as Grid Search or Random Search in order to find the best combination of hyperparameters that works for a model. This will help further refine the model to generalize much better on new data.

To assess the model performance, a variety of metrics were used, particularly, accuracy, precision, recall, and F1-Score. The overall impression is usually presented by the accuracy, although it can be misleading because sometimes there is a class imbalance problem. For this reason, other important metrics include precision, recall, and the F1 score. Precision measures the proportion of positive predictions that were true out of all positive predictions, and recall is the proportion of positive predictions out of actual total positives. The F1-score is the harmonic mean of precision and recall, hence it balances the two.

**5. Results and Analysis**

**Model Performance**

**a) Random Forest Classifier**

A suitable Python code snippet modeled the Random Forest classification algorithm. It first instantiated a class of Random-Forest-Classifiers with a random state of 42 for reproducibility. Secondly, it fitted the model on the training data using the fit() method. Then, the model is used to make predictions on the test data using the predict() method. Then, the accuracy of the model is determined by using the accuracy_score() function and is printed out along with a classification report that provides a more detailed breakdown of the model's performance as showcased below:

**Output:**

**Table 1: Exhibits Random Forest Results**

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        23
           1       1.00      1.00      1.00       177

    accuracy                           1.00       200
   macro avg       1.00      1.00      1.00       200
weighted avg       1.00      1.00      1.00       200

Accuracy: 1.0
```

From these results of the Random Forest model, all metrics are excellent in performance, even recording an accuracy of 1.00 on a test set consisting of 200 instances. More importantly, in the class identifying no churn, labeled 0, the precision was 1.00, and both recall and F1-score were also 1.00; this shows all 23 instances were accurately identified with no false positive or false negative instances. For the churn class, labeled as 1, the model also returned a precision, recall, and F1-score of 1.00, correctly predicting all 177 instances. Also, the macro average and weighted average scores were 1.00; this again confirms the consistent performance of the model in both classes, showing that there were no imbalances or misclassifications within the predictions. Overall, these results show the Random Forest model captures the underlying patterns in the data in this instance, with perfect classification.

**b) Support Vector Machines Modelling**

Python code fragment implemented an SVM classifier with a linear kernel. It initiated a class SVC with the kernel parameter set to 'linear' and the random state to 42 for reproducibility. Then it fitted the model on the training data using the fit() method. This model used the test data, X-test, to make its predictions via the predict () method. The accuracy was computed via the accuracy-score() function and printed out with the classification report, which details further the breakdown of the model as displayed below:

**Output:**

**Table 2: Portrays Support Vector Machines**

```
              precision    recall  f1-score   support

           0       0.73      0.48      0.58        23
           1       0.94      0.98      0.96       177

    accuracy                           0.92       200
   macro avg       0.83      0.73      0.77       200
weighted avg       0.91      0.92      0.91       200

Accuracy: 0.92
```

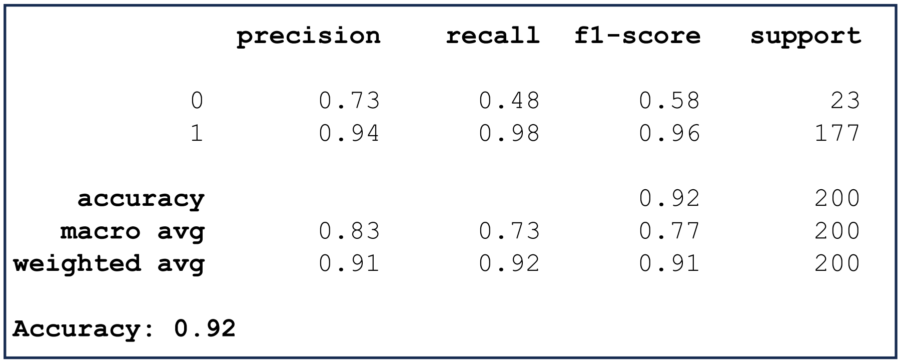
The SVM results had an overall accuracy of 92% on 200 instances, which means that the model classifies most of the cases correctly. The class of non-churn had a precision of 0.73, where even though it correctly predicted 73% of the real instances, it still managed to misclassify some cases as being of a higher class than the non-churn class, leading to the recall being at an all-time low, with a value as low as 0.48 and, hence, did not manage to catch real non-churn cases in their actuality. On the other hand, the churn class did well with 0.94 precision and 0.98 recall; this means the model predicts almost all the actual churning cases out with very few false positives. The macro average scores, 0.83 for precision and 0.73 for recall depict the model's effectiveness across classes, while weighted averages of 0.91 for precision and 0.92 for recall underscore its overall robustness, especially in favor of the churn class. All the same, such disparity in performance between the two classes hints at the areas of improvement for non-churn instances.

**c) Logistic Regression Modelling**

In the code snippet, the implemented model using Logistic Regression was as follows. First, the Logistic-Regression class instantiated with random-state = 42 is for reproducibility, fitting the model with training data using the fit() method. Subsequently, the model predicted on some test data using the predict() method. The accuracy of the model was then evaluated using the accuracy_score() function and printed along with the classification report that provided a more detailed breakdown regarding the performance of the model:

**Output:**

**Table 3: Displays Logistic Regression Results**

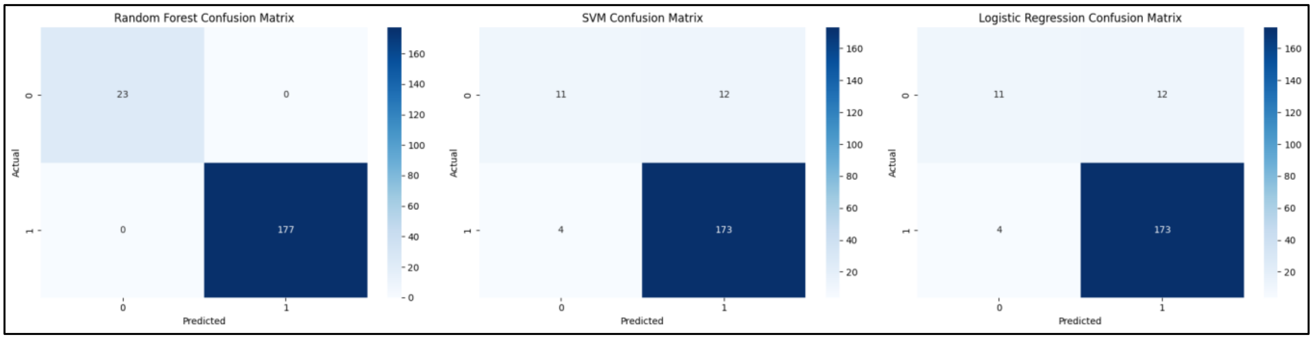


```
              precision    recall  f1-score   support

           0       0.73      0.48      0.58        23
           1       0.94      0.98      0.96       177

    accuracy                           0.92       200
   macro avg       0.83      0.73      0.77       200
weighted avg       0.91      0.92      0.91       200

Accuracy: 0.92
```

The Logistic Regression classification report achieved an overall accuracy of 0.92 from a dataset of 200 instances, indicating that the model performs well in classifying cases. Precision, at class 0, the model only managed to attain a precision of 0.73 hence, it correctly identified only 73% of the predicted cases from instances not churn, while on the other side, recall was just 0.48, which implies that the model could not classify actual non-churn cases correctly. The precision of the churn class was 0.94 and recall 0.98, meaning that the model was successful in capturing almost all the churn cases while keeping the false positive rate low. This further indicates that the performance of both classes is substantiated by a macro average precision and recall of 0.83 and 0.73, respectively, while weighted averages are 0.91 for precision and 0.92 for recall, suggesting overall effectiveness that is dominated by the class of churn. However, the imbalance in performance between the classes was marked and points toward an area for enhancement to better detect instances of non-churn.

**Comparison of All Models**

**Table 4: Exhibits Confusion Matrix of All Models**



The performances varied in terms of identifying an instance as a churn or a non-churn case from the confusion matrix results of models using Random Forest, SVM, and Logistic Regression. In the case of Random Forest, 23 were true negatives (non-churn correctly identified) and 177 true positives, which have correctly been identified as a churn class without any false negatives or false positives; it has perfectly classified the classification concerning the class. The SVM model also presented very high performance, correctly classifying 11 as true negatives and 173 as true positives; however, the model misjudged 12 instances of churn as not being related to it. Logistic Regression had a slightly worse result -11 true negatives, 173 true positives, while in turn facing 4 false negatives-what indeed might attest that it struggled a bit harder than the SVM to find an instance of churn. While all models showed high accuracy, minute misclassifications of the churn cases were observed by both the SVM and Logistic Regression; these are pointers to the areas that need further tuning for improved detection of churning.

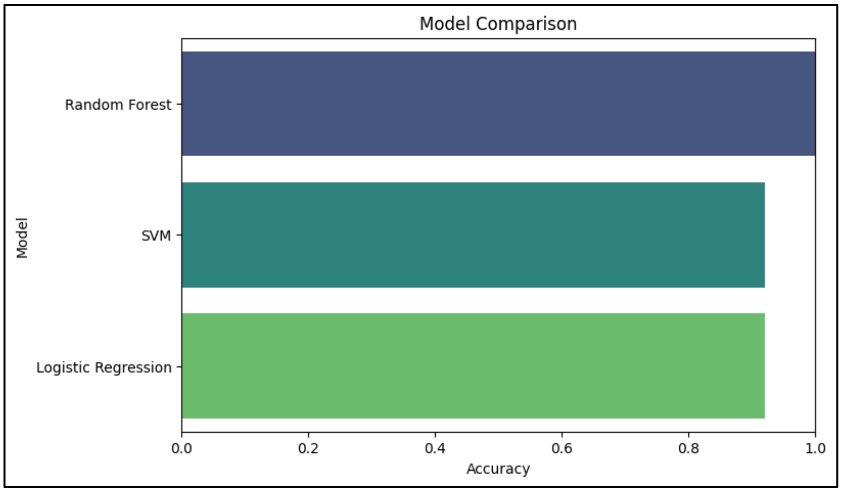


**Figure 6: Depicts Model's Accuracy Comparison**

The model comparison plot shows the performance of three classifications: Random Forest, SVM, and Logistic Regression regarding the classification of churning and non-churning instances. The Random Forest algorithm attained the highest accuracy, nearing 1.0, indicating exceptional performance in effectively identifying both churn and non-churn instances. They perform well in the classification tasks and are slightly below the Random Forest. By contrast, the model with the lowest accuracy among all three models is the Logistic Regression model; this makes it even more puzzling while classifying the instances accurately. From the above visual representation, a strong reflection of better efficiency for the random forest is made, alongside a presentation of the weaknesses in Logistic Regression in need of strengthening to enhance performance. Altogether, all these different models prove varied in effectively handling the problematic issues.

### Feature Importance Analysis

Feature importance analysis is very informative for drivers of customer churn, especially for algorithms like Random Forest and Gradient Boosting, since they intrinsically consider during their training the importance of different features in their splits, hence allowing one to retrieve the most important variables that drive predictions. The idea of Random Forest is that each feature’s importance results from averaging impurity reduction for the specific feature over all trees in the forest. In other words, this means the absolute ranking of the features is allowed regarding their contribution to the predictive power of the model. Gradient Boosting estimates feature importance as the average of contributions of each feature to loss reduction across all trees within an ensemble.

Retrospectively, from the feature importance between the two models, some factors come into key view: The longer customer tenure usually has, the more the business would face fewer churns. Thus, improving customer satisfaction and engagement should be one point of view on which efforts regarding retention will be placed. Moreover, features such as monthly charges and payment methods reveal their impact on the likelihood of churn; for instance, higher fees every month are likely to increase the chances of churning, which indicates that pricing could be a very important factor in customer retention.

### Predictive Insights

Predictive insights from the churn models enable organizations to identify at-risk customers effectively and target them with interventions. By analyzing the predictions of the models, businesses can make out patterns that show a higher likelihood of churning. For instance, customers predicted to have high probabilities of churning often share some characteristics, such as increased complaint frequency, reduced service usage, or even changed payment behavior. These predictive signals help in the focusing of retention efforts and enable organizations to manage customer relationship strategies to utilize scarce resources only on those customers who can most likely leave.

## 6. Retention Strategy Development

### Customer Segmentation

Effective retention strategy development should be underpinned by a detailed segmentation of customers at risk of churning based on churn propensity and other relevant factors. Using predictive analytics, an organization can identify very distinct groups of customers exhibiting different probabilities of churning. For example, based on variables like tenure, usage patterns of service, complaint history, and demographics, customers could be segregated into high-risk, medium-risk, and low-risk.

While the high-risk customers may disengage by showing less use or more complaints, the customers with a moderate risk could have their satisfaction oscillating. This, in turn, allows companies to craft targeted retention strategies that address the peculiar needs and concerns of these segments. Segmentation can be refined further by analyzing the demographics and preferences of the customers. For instance, organizations may find that young customers can engage effectively through digital strategies, while older customers may engage best through a more personal telephone call. This nuance allows for a targeted approach, ensuring that retention efforts resonate with each customer segment's distinctive characteristics and behaviors.

### Targeted Retention Campaigns

Once the customers are segmented, customized retention strategies and campaigns should be developed for each segment to reduce churn. Proactive engagement with high-risk customers is a must. This could be in the form of providing specific incentives, like discounts or loyalty rewards, or improvement in the quality of services concerning certain complaints. For example, if there is dissatisfaction about pricing from a particular segment of customers, campaigns highlighting value-added services tend to work well. Likewise, re-engagement campaigns for customers showing signs of disengagement may involve personalized emails or special promotions that can once again fire their interest.

Communication is the vital key to such retention strategies. The recommendations also include multi-channel engagement for at-risk customers: a mix of digital communications like emails or social media, together with more direct ways of reaching a customer by calling. Messages are personalized, pointing to the history that the customer has built up with the company, and they speak to pain points he or she may have. Reaching out to a customer, for example, who recently downgraded their service with an offer to upgrade at a discounted rate, can prove to be attentive and appreciative.

### Business Impact Analysis

The business benefits to be derived from predictive analytics-driven retention strategies can be considerable. In estimating the potential business impact, a study of CLV and retention rates will come into consideration. For instance, even by shaving a few percentage points of churn, an organization would go a long way toward growing revenue because it is considerably less expensive to retain an existing customer than to win a new one. Furthermore, it means that customers will be much happier, and it leads them to become brand advocates who can create more referrals and business.

Another important aspect is the cost-benefit analysis of the retention efforts. It has to consider the costs associated with running the campaigns for retention, such as marketing expenses, discounts provided, and resources deployed in the improvement of customer service, against the revenue expected from the retained customers. By quantifying these, businesses can realize the return on investment for retention strategies. For instance, if the cost of the campaign is lower than the revenue hike from retained customers, the strategy is considered effective. A fully developed retention strategy with customer segmentation and targeted campaigns will yield better customer loyalty, enhance brand reputation, and provide significant financial benefits to the organization.

## 7. Discussion

### Implications for Telecom Companies

Predictive analytics carved into a telecom company provides enormous insight for improving its customer retention strategy. By applying advanced algorithms that study customer behavior and indicate those who are at risk, these companies can intervene much earlier than actual churn can occur. For example, predictive models can indicate patterns of dis-engagement-such as reduced consumption of service or more complaints which telecom providers can devise a targeted retention strategy.

This targeted approach not only raises customer satisfaction but also efficiently optimizes the marketing spend on high-risk customers who are most likely to respond positively to such interventions. The predictive models will effectively be integrated into business processes, while telecom companies will adopt a data-driven culture of collaboration between data scientists and business units. Real-time analytics systems support timely decision-making by allowing dynamic adjustment in retention campaigns based on current customer behavior. Moreover, the training programs of employees can make them comprehend and utilize predictive insights and make the organization agile to make the business further responsive towards its customers.

### Challenges and Limitations

Apart from the benefits, some setbacks and limitations with predictive analytics do arise in performing customer retention. Particularly, there are some ethical concerns related to data confidentiality of customers; telecom should ensure that GDPR-kind legislations are adhered to strictly, as well as explain themselves transparently while processing information taken from their subscribers. More to that, they need to grant prior information on what this amount of data would be utilized for, while some manner of informed consent availed by the customer should stand uncompromised and their faith also not betrayed.

Furthermore, there are many limitations of model performance due to data quality. Poor-quality data-incomplete or incorrect information may result in misleading predictions that will weaken the retention strategies. Besides, complicated models can create some challenges for interpretation since the stakeholders may not understand the reasoning for certain predictions, which will lower buy-in and implementation. This would also extend to generalizability-lastly, the models could be developed on one set of data but may behave variably well across diversified customer segments or large geographical differences. Defeating these hurdles goes to the betterment of the deployment of predictive analytics while improving customer retention.

## 8. Conclusion

The main goal of this study was to build strong predictive models that could correctly identify at-risk customers in the US telecom market. This research paper aimed to use machine learning algorithms and advanced data analytics to uncover patterns and trends in customer dissatisfaction or intent to churn. This study centered particularly on the American telecom market, examining relevant client data drawn from various sources, entailing billing records, client service interactions, and usage patterns. The dataset for the current study was retrieved from proven and verified sources. This dataset provided intensive insight into customer behavior in terms of churning in the telecom industry. It contained highly elaborate information on customer demographics, service usage, and several indicators that are substantial for the analysis of customer retention and churn. The dataset was designed for the exploration of factors that influence customer churn and retention. The given dataset provided a very good basis for building predictive models aimed at finding customers who are at risk and understanding the dynamics of customer turnover. Among the different models that can be used are Logistic Regression, Support Vector Machines, and Random Forests, among others, each with its advantages and disadvantages. The Random Forest algorithm attained the highest accuracy, indicating exceptional performance in effectively identifying both churn and non-churn instances.

**Funding:** This research received no external funding.
**Conflicts of Interest:** The authors declare no conflict of interest.
**Publisher's Note**: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

## References

[1] Adeniran, I. A., Efunniyi, C. P., Osundare, O. S., Abhulimen, A. O., & OneAdvanced, U. (2024). Implementing machine learning techniques for customer retention and churn prediction in telecommunications. Computer Science & IT Research Journal, 5(8).

[2] Afzal, M., Rahman, S., Singh, D., & Imran, A. (2024). Cross-sector application of machine learning in telecommunications: enhancing customer retention through comparative analysis of ensemble methods. *IEEE Access*.

[3] Al-Mansouri, A. (2024). Anticipating Churn: AI-driven Insights for Sustaining Customer Loyalty in US Business Markets. *Journal of Engineering and Technology*, *6*(1), 1-8.

[4] Bhattacharyya, J., & Dash, M. K. (2022). What do we know about customer churn behaviour in the telecommunication industry? A bibliometric analysis of research trends, 1985–2019. *FIIB Business Review*, *11*(3), 280-302.

[5] Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms*, *17*(6), 231.

[6] Gurung, N., Hasan, M. R., Gazi, M. S., & Chowdhury, F. R. (2024). AI-Based Customer Churn Prediction Model for Business Markets in the USA: Exploring the Use of AI and Machine Learning Technologies in Preventing Customer Churn. *Journal of Computer Science and Technology Studies*, *6*(2), 19-29.

[7] Islam, M. Z., Shil, S. K., & Buiya, M. R. (2023). AI-Driven Fraud Detection in the US Financial Sector: Enhancing Security and Trust. *International Journal of Machine Learning Research in Cybersecurity and Artificial Intelligence*, *14*(1), 775-797.

[8] Kumar, K. P., Kanishkar, P., Raja, V. D., Kumar, T. A., Gopal, S. B., & Gunasekar, M. (2023, December). Telecom Churn Movement Prediction Using Machine Learning. In *International Conference on Intelligent Systems Design and Applications* (pp. 235-243). Cham: Springer Nature Switzerland.

[9] Jain, H., Khunteta, A., & Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: a review. *Telecommunication Systems*, *76*, 613-630.

[10] Mathu, M. (2020). *Reducing Customer Churn In The Telecommunication Industry By Use Of Predictive Analytics* (Doctoral dissertation, University of Nairobi).

[11] Melian, D. M., Dumitrache, A., Stancu, S., & Nastu, A. (2022). Customer churn prediction in telecommunication industry. A data analysis techniques approach. *Postmodern Openings*, *13*(1 Sup1), 78-104.

[12] Mitchell, W. D. (2020). *Proactive Predictive Analytics Within the Customer Lifecycle to Prevent Customer Churn*. Northcentral University.

[13] Rahman, A., Debnath, P., Ahmed, A., Dalim, H. M., Karmakar, M., Sumon, M. F. I., & Khan, M. A. (2024). Machine learning and network analysis for financial crime detection: Mapping and identifying illicit transaction patterns in global black money transactions. *Gulf Journal of Advance Business Research*, *2*(6), 250-272.

[14] Saha, L., Tripathy, H. K., Gaber, T., El-Gohary, H., & El-kenawy, E. S. M. (2023). Deep churn prediction method for telecommunication industry. Sustainability, 15(5), 4543.

[15] Saleh, S., & Saha, S. (2023). Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university. *SN Applied Sciences*, *5*(7), 173.

[16] Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports*, *14*(1), 13097.

[17] Vemulapalli, G. (2024). AI-Driven Predictive Models Strategies to Reduce Customer Churn. *International Numeric Journal of Machine Learning and Robots*, *8*(8), 1-13.

[18] Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *Journal of Big Data*, *7*(1), 29.

[19] Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated churn prediction and customer segmentation framework for telco business. *Ieee Access*, *9*, 62118-62136.

[20] Zatonatska, T., Fareniuk, Y., & Shpyrko, V. (2023). Churn rate modeling for telecommunication operators using data science methods. Marketing i menedžment innovacij, 14(2), 163-173

[21] Zdziebko, T., Sulikowski, P., Sałabun, W., Przybyła-Kasperek, M., & Bąk, I. (2024). Optimizing Customer Retention in the Telecom Industry: A Fuzzy-Based Churn Modeling with Usage Data. *Electronics*, *13*(3), 469.