Analyzing the rate of customers leaving Telco services using Regression Statistics, Machine Learning

1st Nguyen An Duc
Faculty of Information Systems
University of Information Technology
Ho Chi Minh City, Vietnam
22520268@gm.uit.edu.vn

2nd Ngo Hoang Phuong Khanh
Faculty of Information Systems
University of Information Technology
Ho Chi Minh City, Vietnam
22520639@gm.uit.edu.vn

3rd Nguyen Thi Xuan Quynh
Faculty of Information Systems
University of Information Technology
Ho Chi Minh City, Vietnam
22521234@gm.uit.edu.vn

Abstract – The application of machine learning for churn prediction is attracting a lot of attention these years. A large amount of research has been conducted in this area and multiple existing results have shown that machine learning methods could be successfully used toward churn predicting using the customer information data. In this paper, we used Telco's customer information dataset to perform some main machine learning models of this project (Logistic Regression, Random Forest and Support Vector Machine).

Keywords – Customer Churn, Predict, Logistic Regression, Random Forest, Support Vector Machine.

INTRODUCTION

Telecommunication is a highly competitive industry. There a several runners in the market, thus managing relationship with the customers has become vital for the service providers. The organizations employ a variety of tactics to increase their revenues such as attracting new customers, selling more services to the existing customers and most importantly retaining the old customers. If a large number of customers churn in a short span of time, the reputation of the provider gets affected, as the businesses nowadays are highly affected by the word of mouth and social media influence. The issues related to the customer support and service satisfaction are the main reasons behind the churn. Working on preventing the existing customers from churning is inexpensive in terms of cost and time both and keeps the performance of the firm stable and strong. Captivating the new customers is said to be around five times costlier that stopping the existing ones to leave **[1]**.

In this project, we employ a variety of machine learning and regression models to detect factors influencing customer churn and to predict the likelihood of customer departure. Our approach incorporates algorithms like Logistic Regression, Random Forest, and Support Vector Machine (SVM).

To evaluate the performance of these predictive models, we assess them based on four key classification metrics: **Accuracy**, **F1-score**, **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)**, and **Precision-Recall curves**. These metrics provide a comprehensive view of each model's effectiveness in distinguishing between churning and non-churning customers, allowing us to determine which model performs best in predicting churn and aiding in decision-making for retention strategies.

RELATED WORKS

In the telecommunications industry, accurately predicting customer churn is crucial, as customer retention can significantly impact profitability and growth. Recent advancements in machine learning have enabled telecom companies to apply data-driven methods for analyzing and predicting churn rates, facilitating more targeted retention strategies. This study utilizes models including Logistic Regression, Random Forest, and Support Vector Machine (SVM) to analyze and predict the rate at which customers leave Telco services.

Logistic Regression is popular for binary classification tasks such as churn prediction, with Md Parvez Ahmed (2024) reporting moderate accuracy and suitability for identifying customers likely to churn **[2]**.

Random Forest, an ensemble technique combining multiple decision trees, has shown resilience against overfitting and adaptability with large datasets, though at the cost of interpretability in ensemble form (Kirk, 2020). In this study, RF was used to predict customer churn because it was quick and could handle missing and imbalanced data **[4]**.

Finally, Support Vector Machine (SVM) has demonstrated high prediction accuracy and training efficiency in high-dimensional data environments, with Xiahou and Harada (2022) affirming its effectiveness in distinguishing churn and non-churn classes, particularly when kernel functions are optimized **[5].**

MATERIALS

A. Dataset

The Telco Customer Churn dataset contains detailed information about a telecommunications company's customer base. It includes data on customer demographics, account information, service usage, and billing details. The primary focus of the dataset is to identify factors influencing customer churn, indicated by a binary column called 'Churn,' which specifies whether a customer has discontinued the service. This dataset is useful for analyzing patterns in customer behavior and developing predictive models using statistical and machine learning techniques to reduce churn and enhance business strategies. The dataset was gathered from Investing website, containing 21 columns and 7024 rows of data.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneSe
0	7590-VHVEG	Female	0	Yes	No	1	No
1	5575-GNVDE	Male	0	No	No	34	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes
3	7795-CFOCW	Male	0	No	No	45	No
4	9237-HQITU	Female	0	No	No	2	Yes

B. Common Items

I. Import the libraries

- Pandas: Calculate mean, median, mode, range (using .max() - .min()), variance (var), and standard deviation (std).
- Matplotlib and Seaborn: Create a histogram, boxplot,

and scatterplot.

- SciPy: Perform T-test and Chi-square test on data such as Monthly Charges and Total Charges.

II. Load and process data

In this dataset, the relevant columns for performing the calculations are:

- MonthlyCharges: Used for calculating mean, median, mode, range, variance, and standard deviation. It can also be visualized through histograms, box plots, and scatter plots to observe spending distributions and detect outliers.
- TotalCharges: Though initially of type "object," it can be converted to a numerical format for similar calculations and visualizations. This column is useful for understanding total customer expenditure and identifying high-spending customers.

III. Data Calculation in Statistics

1. Mean

- Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Mean Symbol is the bar above the letter x (X Bar).
- x_i: ith term in data set.
- n: Number of variables in data set.

2. Median

- Median is the middle value of the given list of data when arranged in an order.
  - Median for odd data:

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{ observation}$$

  - Median for even data:

$$\text{Median} = \frac{\frac{n^{th}}{2} obs. + \left(\frac{n}{2}+1\right)^{th} obs.}{2}$$

- Median is not affected by values at both ends of the distribution.

3. Mode

- Mode is defined as the value that has a higher frequency in a given set of values. It is the value that appears the greatest number of times. If no value repeats, the dataset may have no mode, or it may have multiple modes.
- This is the commonly used tool for measuring Categorical (nominal). Mode is not affected by values at both ends of the distribution.

4. Range

Range is the difference between the lowest and highest values.

$$R = X_{max} - X_{min}$$

5. Variance

Variance is a measure of how data points differ from the mean. In other words, a variance measures how far a set of data (numbers) is spread out from their mean (average) value.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- x_i: Each individual value in the dataset
- $\underline{x}$: The mean of the dataset
- n : The total number of values
- S^2: The symbol of sample variance

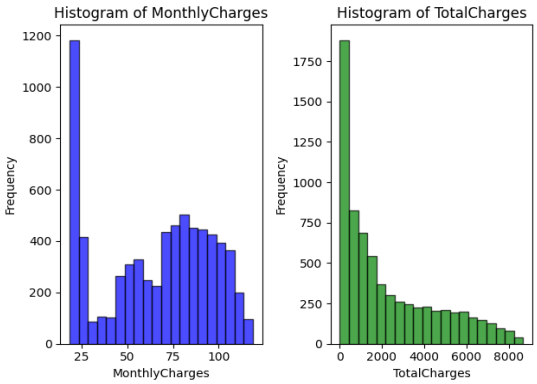
6. Standard deviation

Standard deviation is the positive square root of the variance and a measure of how spread out the data is.

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

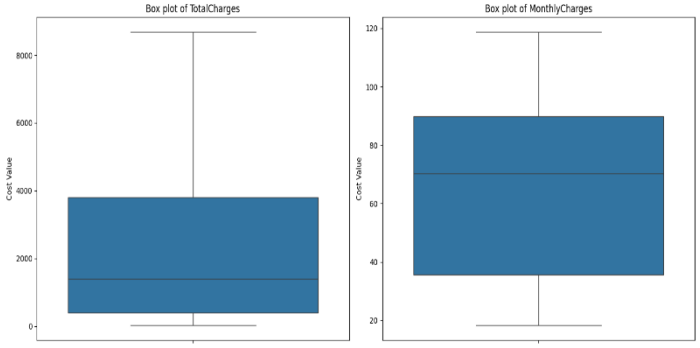
7. Histogram

A histogram represents the distribution of numerical data by grouping values into bins or intervals. It shows the frequency of values in each interval, giving insights into the spread and shape of the data.



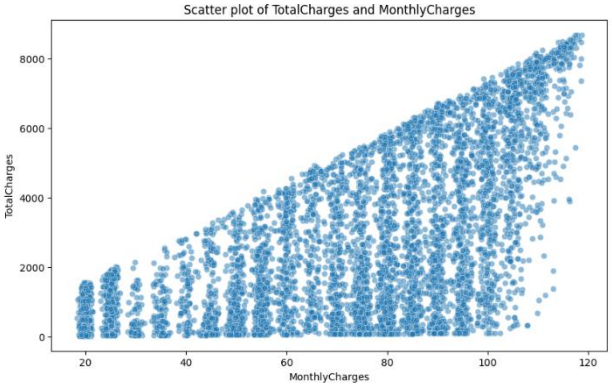
8. Box Plot

A box plot visually summarizes the distribution of a dataset based on five statistics: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. It also highlights potential outliers.



9. Scatter Plot

A scatter plot displays individual data points based on two numerical variables, often revealing relationships or trends between them.



10. Confidence Interval

A confidence interval gives the probability within which the

true value of the parameter will lie. The confidence level (in percentage) is selected by the investigator. The higher the confidence level is the wider is the confidence interval (less precise). Before learning the confidence interval, one must understand the basic statistics formulas and z-score formula. The formula for the confidence interval is given below:

- If n ≥ 30, Confidence Interval = $\bar{X} + Z_C\left(\frac{\sigma}{\sqrt{n}}\right)$
- If n<30, Confidence Interval = $\bar{X} + t_C\left(\frac{S}{\sqrt{n}}\right)$

- n = Number of terms
- $\bar{x}$ = Sample Mean
- σ = Standard Deviation
- z_c = Value corresponding to confidence interval in z table
- t_c = Value corresponding to confidence interval in t table

11. T-Test (Two-Sample T-Test)

A T-test compares the means of two groups to see if they are significantly different from each other.

$$t = \frac{\underline{x_1} - \underline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- $\underline{x_1}, \underline{x_2}$: Means of the two groups.
- s_1^2, s_2^2: Variances of the two groups.
- n_1, n_2: Sample sizes of the two groups.

12. Chi-Square Test

A chi-square test evaluates if there is a significant association between categorical variables.

$$x^2 = \sum \frac{(O - E)^2}{E}$$

- O: Observed frequency.
- E: Expected frequency based on the hypothesis.
- x^2: The chi-square statistic, indicating the difference between observed and expected frequencies.

C. The Result

Calculate the values of Mean, Median, Mode, Range, Variance, Standard Deviation, T-Square and Chi-Square for the mothly charges and total charges of the dataset

```
Basic Statistics:
                    MonthlyCharges   TotalCharges
Mean                     64.798208    2283.300441
Median                   70.350000    1397.475000
Mode                     20.050000      20.200000
Range                   100.500000    8666.000000
Variance                905.165825 5138252.407054
Standard Deviation       30.085974    2266.771362

T-Test Results:
T-Statistic: -82.06412329277843
P-Value: 0.0
There is a statistically significant difference between MonthlyCharges and TotalCharges.

Confidence Intervals:
MonthlyCharges: (64.09489704074434, 65.50151934150821)
TotalCharges: (2230.310779269899, 2336.290102413833)

Chi-square Test Results:
Chi-square: 4596.39409002918
P-Value: 0.0
There is a significant relationship between binned MonthlyCharges and TotalCharges.
```

METHODS

A. Logistic Regression

Logistic regression is the one machine-learning algorithm that is not a black box model. Normally black box models are complex but the logistic regression tells what it does actually. Logistic regression can be binary, multinomial or ordinal. The logistic regression takes the real-valued inputs and makes the prediction like input class belonging to the class 0. If the prediction is > 0.5 then it takes the output as class 0 otherwise it takes output as class 1(here class 0 refers non-churners and 1 refers to churners). Logistic regression is achieved by taking the log odds of $P_{\frac{i}{1}} - P_i$ where P is the probability of being churn or not churn. P always will come in rage 0 to 1.

$$Z_i = ln\ ln\ (P_{\frac{i}{1}} - P_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Here β is the coefficients to be learned and $\boldsymbol{X}_1 \ldots \boldsymbol{X}_n$ is the independent variables. Here churn is the dependent variable and rest features are independent variables.

Taking the exponent of both side:

$$P_i = E\ (\langle\, y = 1 | X_i \,\rangle) = e_{\frac{z}{1}} + e_z = e_{\alpha+\beta ixi}\ /1 + e_{\alpha+\beta ixi}$$

In the machine learning algorithm, the value of the coefficient is estimated by using stochastic gradient descent. It just calculates the value of prediction for each instance in the training set and calculates the error for each prediction. In addition, this process continues until the model is accurate enough. In addition, the coefficient keeps updated in the process. For updating the coefficient, the following equation is used:

$$b = b - a * (y - p) * p * (1 - p) * X$$

Here b is the coefficient that is updating and $\alpha\alpha$ is the value that is updated at the start of the model. This is the learning rate of the coefficient, how much it will change. Normally the right value of $\boldsymbol{\alpha}$ is 0.1 to 0.3. and P is the prediction that is making in the model output **[6]**.

B. Random Forest

The random forest method was presented by Breiman as a new development method for decision trees. The general principles of ensemble training techniques are based on the assumption that their accuracy is higher than that of other singular training algorithms. Because it is a combination of several prediction models. It is more accurate than a single model and reduces existing weaknesses. Several decision trees are used in this algorithm. A subset of data is given to each tree. These trees can make decisions and build their classification model with this subset of data. The random forest algorithm is currently one of the best learning algorithms, and due to its good performance in solving the problem of customer churn, it was chosen for classification in this research **[7]**.

Random Forests utilize the bagging technique for their training algorithm. In greater detail, the Random Forests operate as follows: for a training set TSn ={$(x_1.y_1).\cdots.(x_n.y_n)$}, bagging is repeated B times, and each iteration selects a random sample with a replacement from TSnTSn and fits trees to the samples:

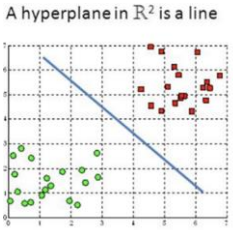
1. Sample n training examples, $X_b.Y_b$.
2. Train a classification tree (in the case of churn problems) f_b on the samples $X_b.Y_b$.

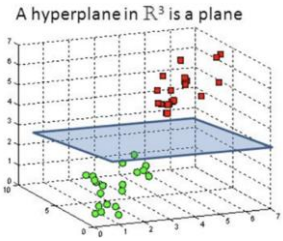
After the training phase, Random Forests can predict unseen samples x′ by taking the majority vote from all the individual classification trees x′.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b\,(x') \quad \textbf{[8]}$$

C. Support Vector Machine (SVM)

Support Vector Machines (SVM) algorithm is extremely useful for classification of binary by identifying the correct hyperplanes (Fig), which is a boundary dividing space into two layers





The hyperplanes:

$$D(x) = w^T x + b \; for - 1 < D(x) < 1$$

- w: an m-dimensional vector of training set
- b: scalar and represents the bias
- T: dot product operation **[9]**

METHODS EVALUATION

Confusion Matrix

The confusion matrix is a tool that easily and effectively shows the performance of the classifier and has the advantage of being easy to interpret the results. A confusion matrix can be used to evaluate the performance of any models or algorithms **[10]**.

	Observed 0	Observed 1
Estimated 0	*TN*	*FN*
Estimated 1	*FP*	*TP*

The four boxes in the classification table are all assigned a name: *FN*, *FP*, *TP*, and *TN*.

- *TN* stands for true negative. Here, the customers are observed as not being churners, and the model has also classified the customers as non-churners.
- *FP* stands for false positive. Here, customers are observed as being non-churners, but the model has classified the customers as churners.
- *FN* stands for false negative. Customers are observed as being churners, but the model has classified the customers as non-churners.
- *TP* stands for true positive. Here, customers are both observed and classified as churners **[11]**.

Precision, Recall, and Accuracy can be calculated using the following formulas:

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Accuracy} = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{TP+TN}{TP+FP+TN+FN}$$ **[12]**

F1-score is one of the critical measures to evaluate imbalanced classification models. It works based on false positive and false negative rates.

$$\text{F1-score} = \frac{2 \times Precision \times Recall}{Precision+Recall}$$ **[13]**

The ROC curve

In the field of churn prediction, the Receiver Operating Characteristic (ROC) Curve is widely recognized as a prominent ranking metric for evaluating the performance of classifiers. This metric enables the assessment of a classifier's ability to differentiate between classes by providing a visual representation of the True Positive rate and False Positive rate of predicted values, as calculated under various threshold values.

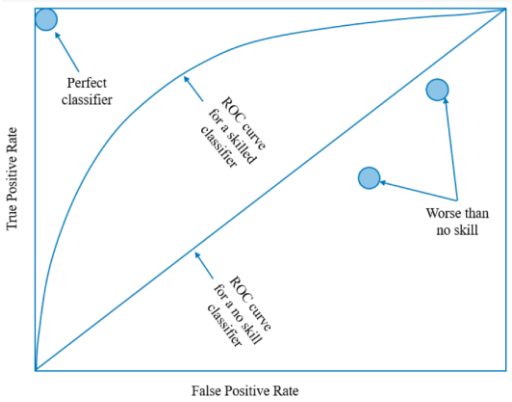
The True Positive rate (recall or sensitivity) is calculated as follows:

$$\text{TruePositiveRate} = \frac{TP}{TP+FN}$$

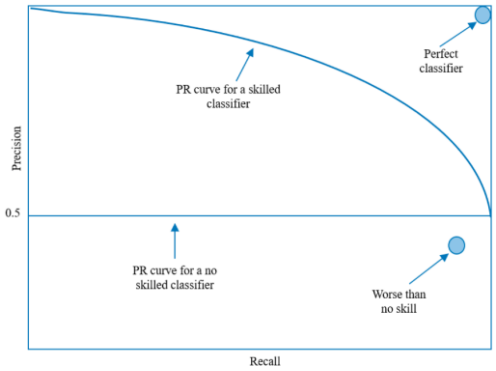
And the False Positive rate is calculated as follows:

$$\text{FalsePositiveRate} = \frac{FP}{FP+TN}$$

Each point on the plot represents a prediction made by the model, with the curve being formed by connecting all points. A line running diagonally from the bottom left to the top right on the plot represents a model with no skill, and any point located below this line represents a model that performs worse than one with no skill. Conversely, a point in the top left corner of the plot symbolizes a perfect model.



The Area Under the ROC curve can be calculated and utilized as a single score to evaluate the performance of models. However, it should be noted that the ROC curve can be effective for classification problems with a low imbalanced ratio and can be optimistic for classification problems with a high imbalanced ratio. In such cases, the Precision–Recall curve is a more appropriate metric because it focuses on the performance of the classifier on the minority class.



The ROC curve is a widely used method for evaluating the performance of machine learning models. The ROC curve plots the True Positive rate against the False Positive rate at various threshold settings, with each point on the curve representing a predicted value by the model.

A horizontal line on the plot signifies a model with no skill, while points below the diagonal line indicate a model that performs worse than random chance. Conversely, a point located in the top left quadrant of the plot represents a model with perfect performance.

In addition to the ROC curve, the Area Under the ROC curve (AUC) is also a commonly used metric for evaluating the performance of machine learning models. The AUC provides a single score for comparing the performance of different models. In cases where the dataset has a high imbalanced ratio, the Precision–Recall AUC (PR AUC) may be more informative as it specifically focuses on the performance of the minority class. However, if the imbalanced ratio of the dataset is not excessively high, such as the dataset utilized in this study, the use of PR AUC may not be necessary for the

evaluation **[12]**.

A. Logistic Regression

	precision	recall	f1-score	support
0	0.85	0.91	0.88	1033
1	0.68	0.55	0.61	374
accuracy			0.81	1407
macro avg	0.76	0.73	0.74	1407
weighted avg	0.80	0.81	0.80	1407

Class 0 (Non-churn):

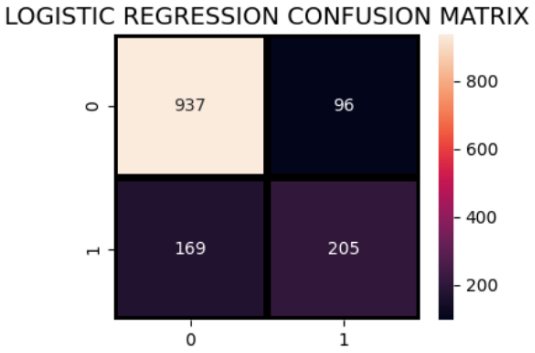
- Precision: 0.85, meaning when the model predicts a customer won't churn, it's correct 85% of the time.
- Recall: 0.91, indicating the model correctly identifies 91% of actual non-churning customers.
- F1-score: The F1 score balances precision and recall. Here, 0 (Non-churn) achieves 0.88.
- Support: 1033 customers didn't churn.

Class 1 (Churn):

- Precision: 0.68, meaning when the model predicts churn, it's correct 63% of the time.
- Recall: 0.55, meaning the model identifies 53% of actual churning customers, suggesting it misses nearly half.
- F1-score: achieves 0.61, reflecting a performance imbalance with the model being more accurate at predicting non-churners.
- Support: 374 customers churned.

Overall Metrics:

- Accuracy: 0.81, indicating the model correctly predicts churn status for 81% of customers.
- Macro Avg: Averages metrics without considering class imbalance (0.76 for precision, 0.73 for recall, 0.74 for F1-score).
- Weighted Avg: Averages metrics considering class distribution, which shows the overall effectiveness of the model (0.80 for precision, 0.81 for recall, 0.80 for F1).

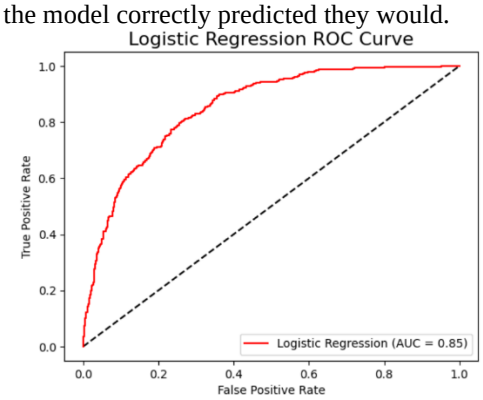


In this confusion matrix:

- Label 0 likely represents "No Churn" (customers who did not cancel the service).
- Label 1 likely represents "Churn" (customers who canceled the service)

Analysis of Values in the Matrix:

- True Negatives (TN): 937 customers did not churn, and the model correctly predicted they would not churn.
- False Positives (FP): 96 customers did not churn, but the model incorrectly predicted they would.
- False Negatives (FN): 169 customers churned, but the model incorrectly predicted they would not.
- True Positives (TP): 205 customers churned, and the model correctly predicted they would.



The y-axis represents the True Positive Rate (TPR), also known as recall or sensitivity, which measures the proportion of actual positive cases (e.g., churners) that are correctly identified by the model. It ranges from 0 to 1, where 1 indicates perfect sensitivity (i.e., all actual positives are identified correctly).

The x-axis represents the False Positive Rate (FPR), which is the proportion of actual negative cases (e.g., non-churners) that are incorrectly predicted as positives (churners). A lower FPR indicates better model performance. The FPR also ranges from 0 to 1.

A curve closer to the top-left corner indicates a model with a better classification performance, as it achieves a high TPR and a low FPR.

The AUC value of 0.85 (85%) indicates that the Logistic Regression model performs well in classification. This means:

- The model has a high predictive ability: it correctly distinguishes between churn and non-churn customers 85% of the time.
- A value closer to 1 indicates high accuracy, while a value near 0.5 indicates poor model performance.

B. Random Forest

	precision	recall	f1-score	support
0	0.82	0.92	0.87	1033
1	0.69	0.45	0.55	374
accuracy			0.80	1407
macro avg	0.75	0.69	0.71	1407
weighted avg	0.79	0.80	0.79	1407

Class 0 (Non-churn):

- Precision: the precision is **0.82**, meaning that 82% of the time when the model predicts a customer as "non-churn," it is correct.
- Recall: for class 0, recall is **0.92**, indicating that the model identifies 92% of actual "non-churn" customers correctly.
- F1-score: the F1-score is **0.87**, showing that the model performs well for "non-churn" predictions.
- Support: 1033 didn't churn

Class 1 (Churn):

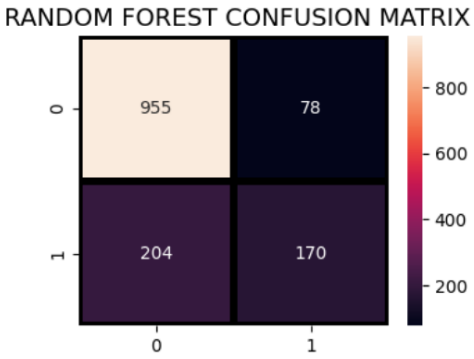
- Precision: the precision is 0.69, which is lower, indicating the model has a higher chance of mistakenly labeling customers as "churn" when they aren't.
- Recall: recall is 0.45, which means the model only captures 45% of actual "churn" cases, missing a

significant portion.

- F1-score: the F1-score is 0.55, which reflects the poorer performance for identifying churn cases.
- Support: 374 customers churn

Overall Metrics:

- Accuracy: 0.80 – The model's overall accuracy is 80%, meaning that it correctly classifies 80% of the total instances.
- Macro Avg:
  - Precision: 0.75 – The average precision across both classes.
  - Recall: 0.69 – The average recall across both classes.
  - F1-score: 0.71 – The average F1-score, suggesting moderate performance.
- Weighted Avg: Takes the number of instances (support) of each class into account, providing an overall metric that is more reflective of class distribution.

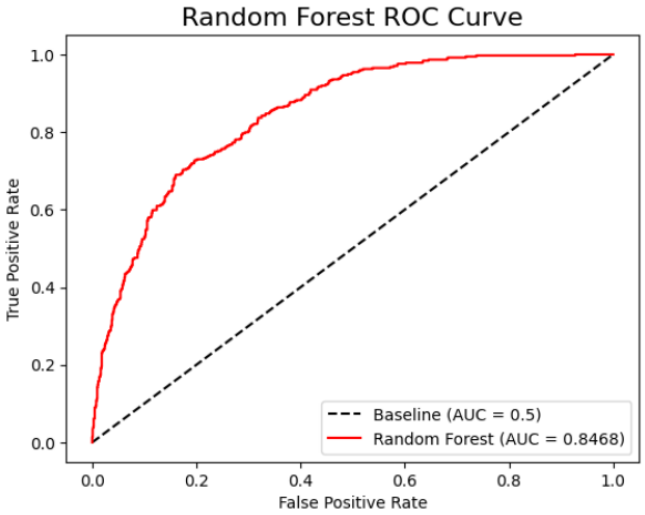


In this confusion matrix:

- Label 0 likely represents "No Churn" (customers who did not cancel the service).
- Label 1 likely represents "Churn" (customers who canceled the service)

Analysis of Values in the Matrix:

- True Negatives (TN): There are 955 instances where the model correctly predicted "No Churn" when the customer did not churn.
- False Positives (FP): There are 78 instances where the model incorrectly predicted "Churn" when the customer did not churn.
- False Negatives (FN): There are 204 instances where the model incorrectly predicted "No Churn" when the customer actually churned.
- True Positives (TP): There are 170 instances where the model correctly predicted "Churn" when the customer actually churned.



The x-axis represents the False Positive Rate (FPR), which indicates the proportion of actual negatives (non-churners) that are incorrectly classified as positive (churners).

The y-axis represents the True Positive Rate (TPR), also known as Recall or Sensitivity, which indicates the proportion of actual positives (churners) that are correctly classified.

Any point above this line indicates a model with better-than-random performance, and the farther the curve from the line, the better the model.

The ROC curve for the Random Forest model (shown in red) is significantly above the diagonal, indicating that the model performs better than random guessing.

The curve reaches high TPR values (close to 1.0) while maintaining relatively low FPR values, especially in the initial part of the curve. This suggests that the model can accurately identify most churn cases without too many false positives.

The AUC value indicates that the Random Forest model performs well in distinguishing churn and non-churn customers.

AUC values closer to 1 represent better model performance, while values near 0.5 indicate poor performance.

C. Support Vector Machine (SVM)

	precision	recall	f1-score	support
0	0.82	0.93	0.87	1033
1	0.69	0.45	0.54	374
accuracy			0.80	1407
macro avg	0.76	0.69	0.71	1407
weighted avg	0.79	0.80	0.78	1407

Class 0 (Non-Churn):

- Precision: 0.82 – The model is 83% accurate when predicting non-churn customers. This means that 83% of the customers predicted as non-churn are correctly labeled.
- Recall: 0.93 – Out of all actual non-churn customers, the model successfully identifies 86% of them.
- F1-score: 0.87 – A high F1-score indicates a good balance between precision and recall for non-churn customers.
- Support: 1033 – There are 1,549 actual non-churn customers in the dataset.

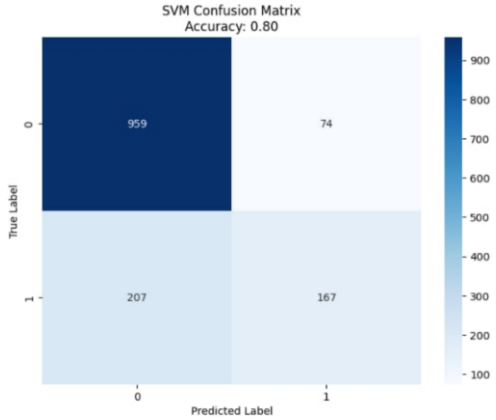
Class 1 (Churn):

- Precision: 0.69 – When the model predicts a customer will churn, it is correct 58% of the time. This indicates the presence of false positives.

- Recall: 0.45 – The model identifies only 52% of the actual churn cases, suggesting that many customers who churn are not being correctly predicted by the model (false negatives).
- F1-score: 0.55 – The relatively low F1-score shows that the model's performance on churn prediction is moderate and could be improved.
- Support: 374 – There are 561 actual churn customers in the dataset.

Overall Model Performance:

- Accuracy: 0.80 – The model's overall accuracy is 80%, indicating that it correctly predicts the class (churn or non-churn) in 80% of cases.
- Macro Avg:
  - Precision: 0.76 – The average precision
  - Recall: 0.69 – The average recall, showing the model's ability to identify actual positive and negative cases.
  - F1-score: 0.71 – The macro-average F1-score reflects that the model's performance on both classes is balanced but leans towards moderate.
- Weighted Avg:
  - Precision weighted: 0.79.
  - Recall weighted: 0.80.
  - F1-score weighted: 0.78.



True Negative (TN):

- 959. This is the number of samples that truly belong to class 0 (not churn) and are correctly predicted as class 0.

False Positive (FP):

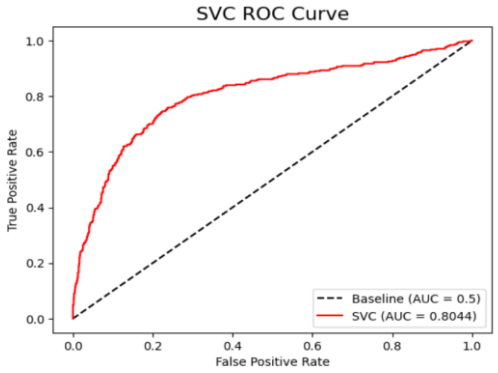
- 74. This is the number of samples that truly belong to class 0 but are incorrectly predicted as class 1 (churn). This leads to a false alarm.

False Negative (FN):

- 207. This is the number of samples that truly belong to class 1 (churn) but are incorrectly predicted as class 0. These cases are where the model fails to detect churn even though they truly belong to the churn group.

True Positive (TP):

- 167. This is the number of samples that truly belong to class 1 (churn) and are correctly predicted as class 1.



Significance of the ROC Curve:

- The ROC Curve shows the relationship between the False Positive Rate (FPR) and the True Positive Rate (TPR) as the classification threshold changes.
- The X-axis represents the FPR: The rate at which customers are incorrectly predicted to churn while they actually remain.
- The Y-axis represents the TPR: The rate at which customers are correctly predicted to churn.

AUC Value (Area Under the Curve):

- The AUC of the SVC is 0.8044, indicating that the model performs well in distinguishing between customers who will churn and those who will stay.
  - AUC = 1: The model is perfect.
  - AUC = 0.5: The model has no classification ability (equivalent to random guessing).

Baseline Line (AUC = 0.5):

- The dashed black line (Baseline) represents a model with random predictions, where the TPR equals the FPR.

CHALLENGES ENCOUNTERED

During the implementation of the research project "Analyzing the Rate of Customers Leaving Telco Services Using Regression Statistics and Machine Learning," we encountered several challenges, including:

- Difficulty in Data Processing: Customer churn data often includes missing values, categorical variables, and imbalanced classes, all of which required systematic preprocessing to ensure the feasibility and accuracy of our prediction models.
- Challenges in Model Selection: Building effective churn prediction models requires a deep understanding of various machine learning algorithms. We needed to make critical decisions on model selection, balancing interpretability and accuracy, while addressing complex patterns in customer behavior.
- Complexity in Model Evaluation: Evaluating churn prediction models involved using multiple statistical metrics. We tested our models for accuracy, precision, recall, and F1-score, but found that achieving a balance between precision and recall remained challenging, particularly given the impact of false positives in a churn context.

In the future, we aim to address these challenges by:

- Enhancing Data Processing Techniques: We plan to explore more advanced data processing methods, such as feature engineering and

synthetic data generation, to improve the robustness and accuracy of our models.

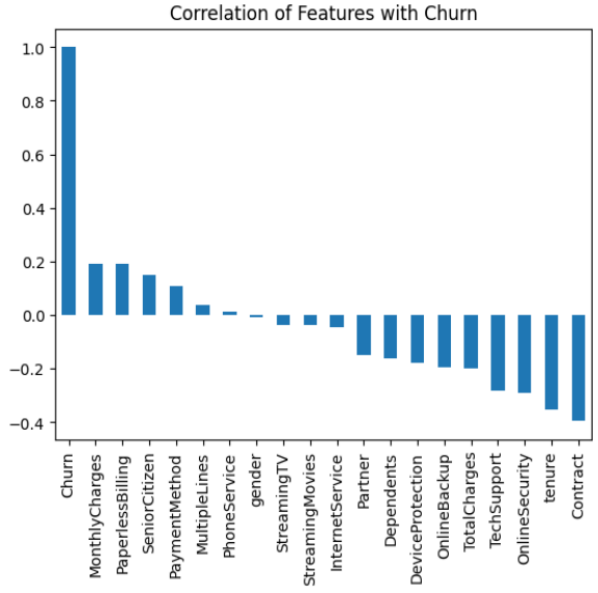
- Utilizing Advanced Modeling Techniques: We will continue researching advanced models, including ensemble methods and deep learning, to better capture the complexities of customer churn behavior.
- Strengthening Evaluation Metrics: We intend to use metrics that better capture customer churn dynamics, such as Area Under the Precision-Recall Curve (AUPRC) and F2-score, to refine model evaluation and improve the interpretability of results.
- Collaborating with Industry Experts: To improve model performance, we plan to connect with industry professionals specializing in churn analysis, allowing us to exchange insights and adopt best practices.

With these strategies, we are confident in our ability to enhance the accuracy and effectiveness of our customer churn prediction models in the future.

CONCLUSION

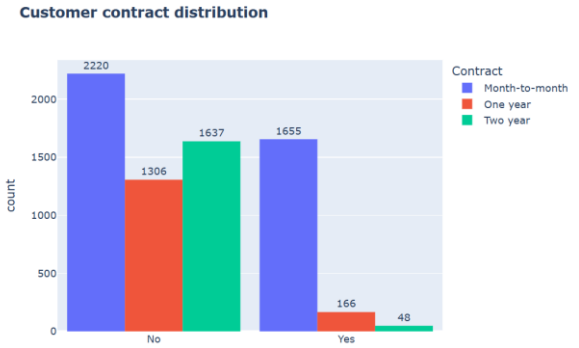
A. Dataset

To understand more about the dataset, we made a chart about the correlation of features with churn. The correlation chart below highlights the relationship between various features in the dataset and customer churn. By examining these correlations, we can identify which factors most strongly influence whether a customer decides to leave or stay with the service.



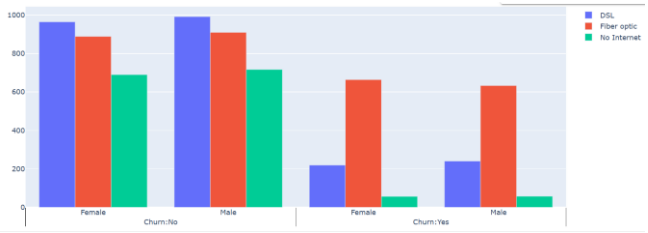
The following section outlines how these features affect churn, the underlying reasons for their impact, and recommended solutions to improve customer retention.

1. Contract Type



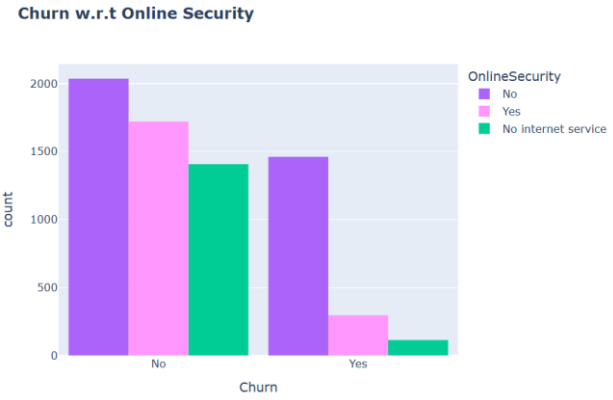
Customers on month-to-month contracts exhibit significantly higher churn rates compared to those on one-year or two-year contracts. This is likely due to the lack of long-term commitment, making it easier for these customers to switch providers. To address this, it is recommended to encourage customers to opt for longer-term contracts by offering attractive discounts, bundled services, or loyalty benefits.

2. Internet Service Type



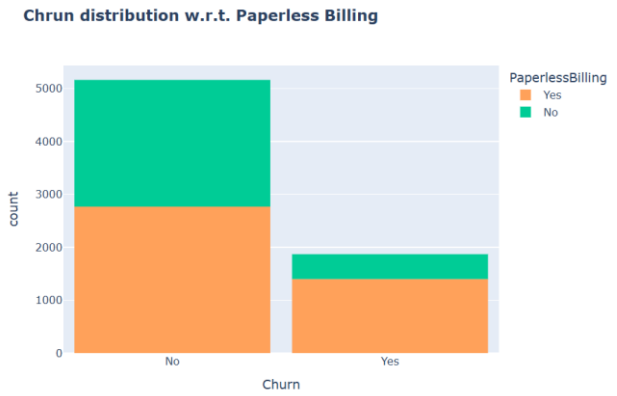
The analysis reveals that Fiber Optic customers are more likely to churn compared to those using DSL or no internet services. This could be attributed to dissatisfaction with pricing, reliability, or service quality. To mitigate this issue, efforts should be focused on improving the reliability and quality of Fiber Optic services.

3. Online Security Services



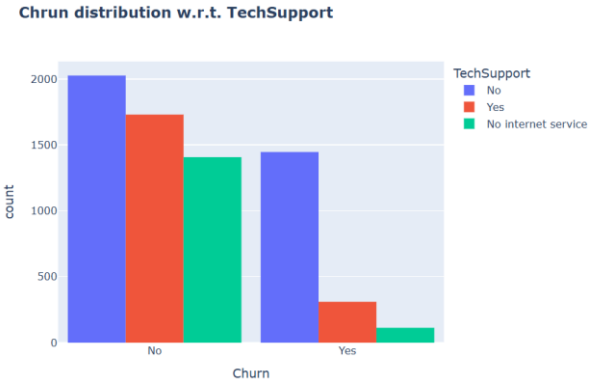
Customers who do not subscribe to Online Security services are more prone to churn. This suggests that these customers may not perceive sufficient value in the service or may feel less secure. To address this, the company should consider offering free trials, discounts, or bundling Online Security with other services. Promoting the benefits of these services can help improve adoption rates and customer retention.

4. Paperless Billing



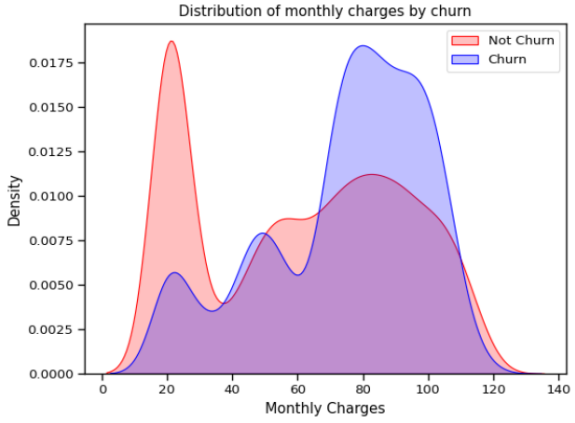
Paperless Billing is associated with higher churn rates compared to traditional billing methods. This could stem from customer dissatisfaction with the digital billing process or a lack of familiarity with the system. To resolve this, it is recommended to improve the usability and transparency of the digital billing platform. Additionally, customers uncomfortable with paperless billing should be offered the option to switch to traditional billing methods.

5. Technical Support



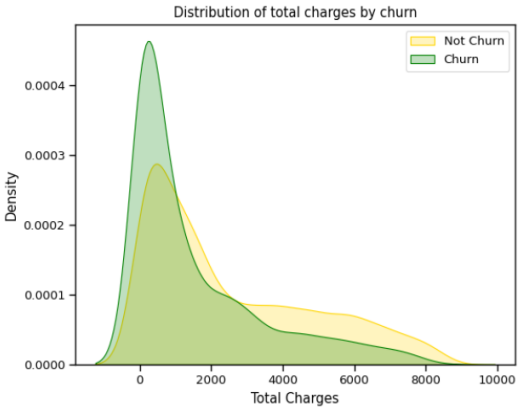
The absence of Tech Support services correlates with higher churn rates, indicating that customers value support during technical issues. To improve retention, the company should enhance the accessibility of Tech Support through multiple channels, such as live chat, phone, and self-help portals. Providing free or discounted trials for Tech Support services and actively marketing its benefits can further encourage adoption.

6. Monthly Charges



Higher Monthly Charges are strongly linked to increased churn rates, as customers may perceive the cost as not aligning with the value of services received. Introducing tiered pricing plans, loyalty discounts, or rewards for consistent payments could make pricing more attractive and reduce the likelihood of churn among high-paying customers.

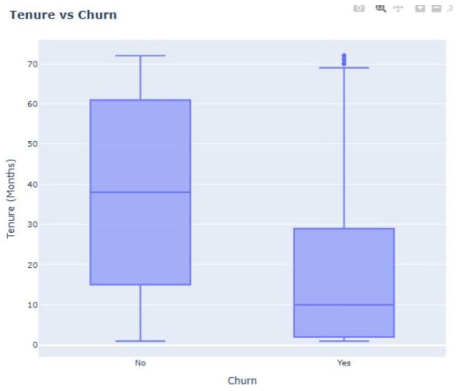
7. Total Charges



Customers with lower Total Charges, typically indicating shorter tenures, are more likely to churn. This suggests that new customers may be leaving due to unmet expectations or initial dissatisfaction. To counter this, the company should implement robust onboarding programs, offer welcome promotions, and conduct satisfaction surveys during the early stages of the customer lifecycle to identify and address any concerns.

In conclusion, the analysis of the dataset provides a comprehensive understanding of the factors influencing customer churn. By identifying key, we have gained valuable insights into customer behavior and the root causes of churn. This understanding allows us to propose targeted strategies for retention, ensuring a data-driven approach to improving customer loyalty and satisfaction.

8. Tenure and Churn



Observations from the chart:
"No" group (did not churn):
Wider interquartile range, with a median around 50 months, indicating that most customers who remain loyal tend to stay for a long time.
Tenure ranges approximately from 0 to over 60 months.
"Yes" group (churned):
Narrower interquartile range and a lower median (around 15 months).
Customers who churned generally have shorter tenures, mostly below 25 months.
A few customers had longer tenures but still churned (these are the outliers).

B. Models

After comparing models on the Telco Customer Churn dataset, we identified the model that provided the best predictive results as follows:

- Logistic Regression: Achieved an accuracy of 81% with a precision of 0.85 and an F1-score of 0.68, providing a clear and interpretable baseline for churn prediction.
- Random Forest: Delivered the best accuracy at 80%, with a precision of 0.69 and an AUC-ROC score of 0.82, making it highly effective for

predicting churn.

- Support Vector Machine (SVM): Had an accuracy of 77%, with precision at 0.83 for non-churn cases, though it struggled with recall on churn cases.

However, to advance customer churn prediction technology, we need to focus on the following issues:

- Improving model accuracy: Although Random Forest and other models have shown promising results in predicting churn, further improvements are needed to enhance accuracy and minimize misclassifications, ensuring reliable predictions.
- Algorithm optimization: We need to explore and apply optimization techniques, such as hyperparameter tuning and feature engineering, to boost the performance and accuracy of our churn prediction models.
- Developing real-world applications: Churn prediction models are valuable for customer retention strategies, helping identify at-risk customers in telecommunications. Expanding practical applications will enhance the value of this technology.
- Researching new prediction models: As new machine learning algorithms and models emerge, studying advanced approaches like deep learning may improve the effectiveness of churn prediction.

In summary, to advance churn prediction technology, we must focus on improving model accuracy, optimizing algorithms, developing real-world applications, and researching new prediction models. These efforts will strengthen our ability to forecast customer churn and support strategic retention efforts in the telecom industry.

ACKNOWLEDGMENT

First of all, we sincerely express our gratitude to Dr. Tran Van Hai Trieu and TA. Nguyen Minh Nhut for their invaluable expertise, guidance, and dedicated support throughout this project. Their insightful advice and enthusiastic assistance have been crucial, and we believe the successful completion of this report would not have been possible without their mentorship and encouragement.

This project has provided each team member with an opportunity to collaborate closely, enhancing teamwork skills, learning from one another, and, most importantly, applying knowledge in a real-world setting. Through this experience, we gained valuable insights, both from applying what we have been taught and from discovering new approaches to improve our work.

Despite our best efforts, limited time, experience, and knowledge mean that some imperfections are inevitable. We respectfully welcome constructive feedback to help us expand our understanding and improve our skills, equipping us better for future projects and practical applications.

Finally, we wish Dr. Tran Van Hai Trieu and TA. Nguyen Minh Nhut abundant health and continued success in their meaningful work of guiding and inspiring future generations.

REFERENCES

[1] Adeniran, Ibrahim Adedeji, et al. "Implementing machine learning techniques for customer retention and churn prediction in telecommunications." *Computer Science & IT Research Journal* 5.8 (2024).

[2] Ahmed, Md Parvez, et al. "A Comparative Study of Machine Learning Models for Predicting Customer Churn in Retail Banking: Insights from Logistic Regression, Random Forest, GBM, and SVM." *Journal of Computer Science and Technology Studies* 6.4 (2024): 92-101.

[3] Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context. International Journal of Electrical and Computer Engineering, 8(4), 2367.

[4] Al-Sultan, Sultan Yahya, and Ibrahim Ahmed Al-Baltah. "An Improved Random Forest Algorithm (ERFA) Utilizing an Unbalanced and Balanced Dataset to Predict Customer Churn in the Banking Sector." *IEEE Access* (2024).

[5] Xiahou, Xiancheng, and Yoshio Harada. "B2C E-commerce customer churn prediction based on K-means and SVM." *Journal of Theoretical and Applied Electronic Commerce Research* 17.2 (2022): 458-475.

[6] Jain, Hemlata, Ajay Khunteta, and Sumit Srivastava. "Churn prediction in telecommunication using logistic regression and logit boost." Procedia Computer Science 167 (2023): 101-112.

[7] Taherkhani, Leila, et al. "Analysis of the Customer Churn Prediction Project in the Hotel Industry Based on Text Mining and the Random Forest Algorithm." Advances in Civil Engineering 2023.1 (2023): 6029121.

[8] Imani, Mehdi, and Hamid Reza Arabnia. "Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis." *Technologies* 11.6 (2023): 167. 16

[9] Ly, Tran Van, and Dao Vu Truong Son. "Churn prediction in telecommunication industry using kernel Support Vector Machines." Plos one 17.5 (2022): e0267935.

[10] Yun, Hongwon. "Prediction model of algal blooms using logistic regression and confusion matrix." International Journal of Electrical and Computer Engineering 11.3 (2021): 2407-2413.

[11] Chang, Victor, et al. "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models." Algorithms 17.6 (2024): 231.

[12] Imani, Mehdi, and Hamid Reza Arabnia. "Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis." Technologies 11.6 (2023): 167.

[13] Sina Mirabdolbaghi, Seyed Mohammad, and Babak Amiri. "Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions." Discrete Dynamics in Nature and Society 2022.1 (2022): 5134356.