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

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

Nisha Gurung[1], **Md Rokibul Hasan**[2] ✉ **Md Sumon Gazi**[3] **and Faiaz Rahat Chowdhury**[4]
[1234]MBA Business Analytics, Gannon University
Corresponding Author: MD Rokibul Hasan, **E-mail**: prorokibulhasanbi@gmail.com

| ABSTRACT

Understanding consumer churn is pivotal for companies in the USA to develop efficient strategies for consumer retention and reduce its negative effects on revenue and profitability. To start with, understanding client churn entails pinpointing the factors that contribute to it. This research paper delved into the application of machine learning algorithms such as Random Forests and Decision Trees for designing churn prediction models and exploring key factors that churn probabilities. The dataset used in this study was sourced from the prominent UCI repository of machine learning databases, preserved at the University of California, Irvine. This dataset provided extensive information on a total of 3333 clients, facilitating in-depth analysis and insights. Models performance evaluation comprised examining the model's efficiency using a confusion matrix. Random Forest seemed to be a relatively better performing model than Decision Tree for this specific classification task. In particular, Random Forest attained higher accuracy (96.25%), precision (91.49), Recall (83.49%), F-measure (0.87), and Phi coefficient (0.85). By deploying Random Forest and Decision Tree models, government companies can get an in-depth comprehension of the factors that lead to consumer churn. As a result, this information may enable them to tailor targeted retention strategies and interventions. By effectively retaining consumers, government organizations can maintain a stable customer base, leading to sustained revenue and economic growth.

| KEYWORDS

Customer Churn; AI; Machine Learning; Radom Forest; Decision Tree.

| ARTICLE INFORMATION

ACCEPTED: 01 April 2024 **PUBLISHED:** 11 April 2024 **DOI:** 10.32996/jcsts.2024.6.2.3x

1. Introduction

As per Banu et al., (2022), customer attrition, or customer churn, denotes the loss of subscribers or customers. It has substantial negative financial effects on organizations across various industries in the USA like insurance, telecommunications, banking, etc. Detecting and preventing client churn using timely intervention and retention strategies has therefore become of utmost priority. With the invention of technologies such as machine learning and artificial intelligence (AI), organizations are now capable of tailoring highly predictive consumer churn prediction models that can assist in pinpointing at-risk consumers and taking preemptive measures. This research paper examines the application of machine learning algorithms such as Random Forests and Decision Trees. for designing churn prediction models and exploring key factors that churn probabilities.

According to Krishna, & Reddy (2023), in the contemporary highly competitive business surrounding, obtaining new clients is much more expensive than retaining the current ones. While the price of retaining current clients normally ranges between 15-23% of the original acquisition cost, consumer re-acquisition costs have been ascertained to be 5-10 times higher. Besides, with consumers becoming increasingly transient and aware, client loyalty is reducing over time. As a consequence, companies across industries are aggressively concentrating on retention to sustain profitable growth and leverage lifetime value from customers.

A pivotal component of any retention strategy is comprehending consumer churn behavior and pinpointing at-risk clients on time. The traditional strategies of monitoring churn comprised tracking basic metrics like service cancellations, payment defaults, etc., and taking reactive interventions post-churn. Nonetheless, with the rapid inventions of technologies and the presence of large customer datasets, organizations can now take a more analytics-driven, proactive, approach to minimize churn via predictive modeling (Optimove, 2023). Progression in artificial intelligence and machine learning have facilitated the advancement of robust consumer churn prediction models that can assist foresee churn risk with a high level of accuracy and facilitate preemptive interventions.

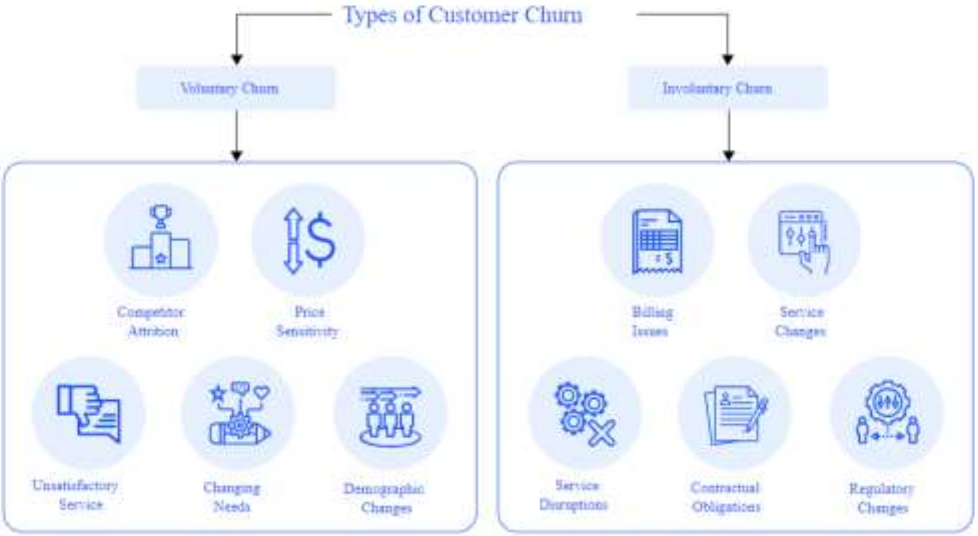
This study explores the application of Artificial Intelligence and machine learning technologies for developing churn prediction models that can efficiently pinpoint consumers likely to defect soon. This research paper delves into the key machine learning models employed for predictive modeling along with elements influencing churn probability estimations. This study equally discusses the business benefits of executing AI-based predictive analytics and presents recommendations on optimizing retention efforts via targeted churn reduction strategies informed by such models.

2. Literature Review

2.1 Understanding Customer Churn

As per Willoughby (2023), understanding consumer churn is pivotal for companies in the USA to develop efficient strategies for consumer retention and reduce its negative effects on revenue and profitability. To start with, understanding client churn entails pinpointing the factors that contribute to it. Client churn can happen because of various reasons, such as dissatisfaction with services or products, pricing issues, competitive offers, or changes in consumer preferences. By pinpointing these factors, organizations can get insights into the underlying causes of churn and take proactive measures to address them.

2.2 Types of Customer Churn



2.2.1 Voluntary Churn

According to Shalini & Kavitha (2023), voluntary churn revolves around the intentional decision made by consumers to terminate their relationship with a company. This may be attributed to factors, such as dissatisfaction with the service or product, attractive offers and discounts from competing organizations, evolving consumer needs, or financial limitations.

2.2.2 Involuntary Churn

On the other hand, involuntary churn happens when clients discontinue their relationship with an organization because of circumstances beyond their control. Illustrations of such incidents are the death of a client, relocation to a location where the service is not accessible, or facing technical challenges that remain unresolved (Shalini & Kavitha, 2023).

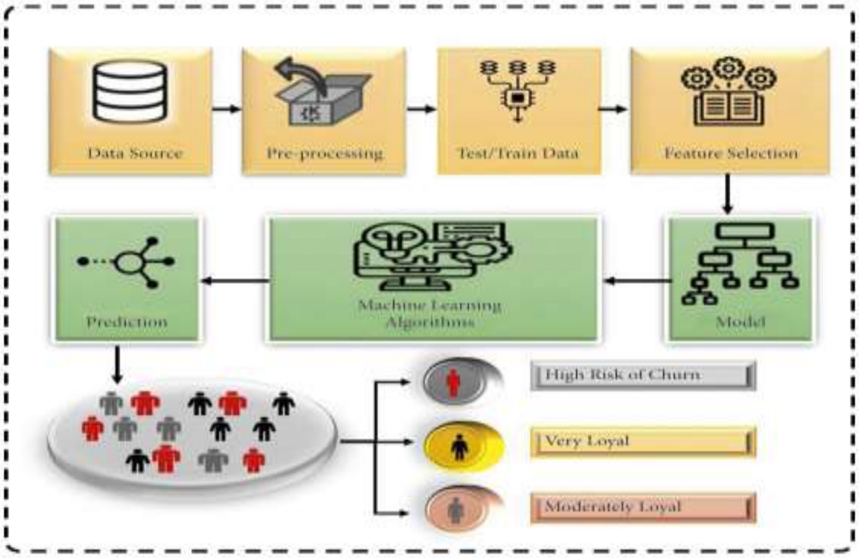
2.3 Financial Implications of Customer Churn

Takyar (2023), indicated that the financial implications of client churn on the organization in the USA cannot be underestimated. When clients churn, companies experience a direct loss of revenue. The loss of loyal clients disrupts the revenue stream and diminishes overall profitability. besides, acquiring new consumers to replace the churned clients can be a costly endeavor. Companies frequently incur higher advertising and marketing expenses to attract and onboard new consumers, leading to increased customer acquisition costs. Furthermore, consumer churn has indirect financial effects that transcend beyond instant revenue loss. For instance, churned consumers may share their negative experiences with others, consequently, leading to a tarnished brand reputation. Negative word-of-mouth can hinder prospective clients from engaging with the organization, resulting in a decrease in market share and a loss of future revenue opportunities.

2.4 The Role of AI and Machine Learning in Churn Prevention

According to Banu et al. (2022), the employment of Artificial Intelligence (AI) and Machine Learning (ML) in churn prediction provides several benefits. Firstly, these inventions can manage numerous volumes of data and complicated patterns, facilitating more accurate churn predictions compared to conventional mainstream methods. Therefore, companies can deploy advanced Machine Learning algorithms, such as random forests, logistic regression, decision trees, and neural networks, to unveil hidden insights and pinpoint factors that substantially influence churn probabilities. Secondly, AI and ML facilitate real-time churn detection. By progressively assessing data in real-time, companies can detect early warning signs of churn and intervene swiftly with suitable retention measures. This proactive approach enables companies to retain customers before they decide to switch to a competitor.

3. Methodology



3.1 Dataset

The dataset used in this study was sourced from the prominent UCI repository of machine learning databases, preserved at the University of California, Irvine. This dataset provided extensive information on a total of 3333 clients, facilitating in-depth analysis and insights. Each client's profile is portrayed by 20 variables, including various attributes and attributes. Moreover, the dataset comprises a crucial target variable entitled 'churn,' which acts as an indicator of whether the client has chosen to terminate their relationship with the organization or has remained an active consumer (proAIrokibul, 2022). This 'churn' variable plays a paramount role in comprehending consumer behavior and is instrumental in developing accurate churn prediction models.

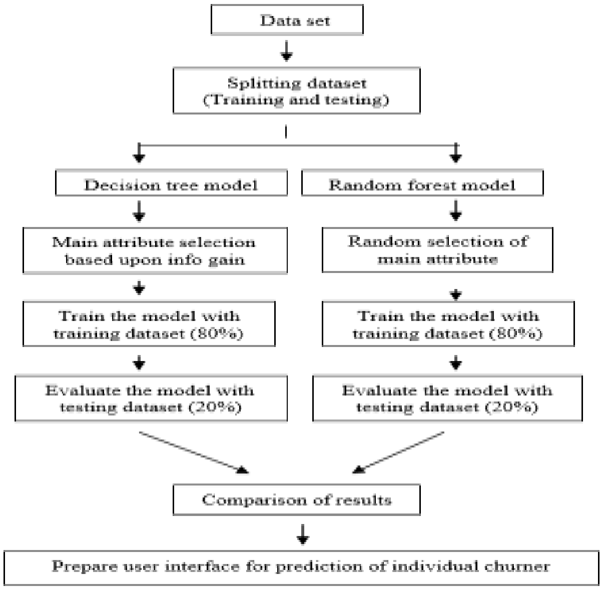
3.2 Data Pre-Processing

The analyst started the data pre-processing by cautiously assessing the dataset to pinpoint any irrelevant or missing data points. In scenarios where there were missing values, the researcher adopted suitable imputation methods to approximate the missing values grounded on the available data (proAIrokibul, 2022). In scenarios where particular data points were considered irrelevant, the analyst made a strategic decision to exclude them from the dataset. After eliminating the irrelevant or missing data, the analyst progressed to transform categorical data into numerical values utilizing methods such as label encoding, one-hot encoding, or. This phase was instrumental in facilitating statistical and mathematical analysis of the data. Subsequently, the transformation of categorical data was performed, where the researcher normalized and standardized the data to guarantee that all features had equal significance and were on a comparable scale. These methodologies assisted in mitigating the effects of outliers enhanced the accuracy of machine learning algorithms, and reinforced the overall quality of the data analysis.

3.4 Feature Engineering and Selection

The primary dataset was first subdivided into two subsets, most notably, the training data, which comprised 80% of the dataset, and the testing data, which constituted the remaining 20%. Within the training dataset, a detailed evaluation exposed that out of the overall 2670 incidents, a substantial majority of 2286 incidences (approximately 85.75%) were labeled as non-churners, represented by a churn status of false. By contrast, a smaller subset of 380 incidents (approximately 14.25%) were pinpointed as churners, demonstrated by a churn status of true. This distribution of non-churners and churners was pivotal for training the model since it enabled the models to learn patterns and attributes related to both churn and non-churn cases.

3.4.1 Proposed Model


Data set
Splitting dataset (Training and testing)
Decision tree model
Random forest model
Main attribute selection based upon info gain
Random selection of main attribute
Train the model with training dataset (80%)
Train the model with training dataset (80%)
Evaluate the model with testing dataset (20%)
Evaluate the model with testing dataset (20%)
Comparison of results
Prepare user interface for prediction of individual churner

3.4.2 Decision Trees

The decision tree is a non-parametric technique that is extensively adopted for solving both regression and classification problems. This algorithm performs by partitioning the data into sub-categories premised on the most substantial features, progressing this procedure until a stopping procedure is satisfied. The resulting architecture bears a similarity to a tree, with every internal node representing a feature, each branch portraying a decision rule, and every leaf node symbolizing a numeric value and class label. In the setting of this research, a decision tree algorithm can be optimized to identify the most impactful components that influence the churn rate of telecommunication consumers. By examining the decision rules of the tree, the analyst can infer insights into the factors that are most strongly associated with customer churn and consequently tailor interventions to counteract the threat of consumer attrition. The entropy of the overall samples can be computed using the expression below:

$$Entropy(Ent) = -\sum_{i=1}^{n}[p_i^+ \log_2(p_i^+) + p_i^- \log_2(p_i^-)]$$

Where P_i^- and P_i^+ stand for the negative and positive target variables.

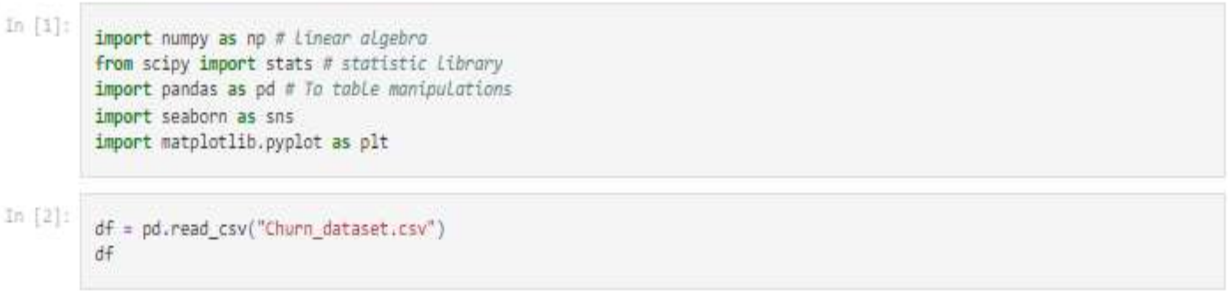
3.4.3 Random Forest

Random Forest, is a renowned ensemble technique comprising decision trees, it has been widely adopted for both regression and classification tasks. In this study, Random Forest was adopted as the secondary classifier to assess the model's performance. The Random Forest technique is greatly accredited in the data science field because of its accuracy and flexibility in terms of handling diverse datasets. The Random Forest model initiates the procedure of developing multiple decision trees by considering distinct characteristics at each split node. In the setting of predicting churn among telecommunication subscribers, a Random Forest algorithm can be trained. This algorithm uses various elements such as service usage, client demographics, and billing data to forecast if a customer is likely to churn or not. By utilizing the joint predictive power of multiple decision trees within the Random Forest ensemble, the algorithm can efficiently analyze and interpret the complex association between these elements and customer churn behavior.

3.5 Experimental Results

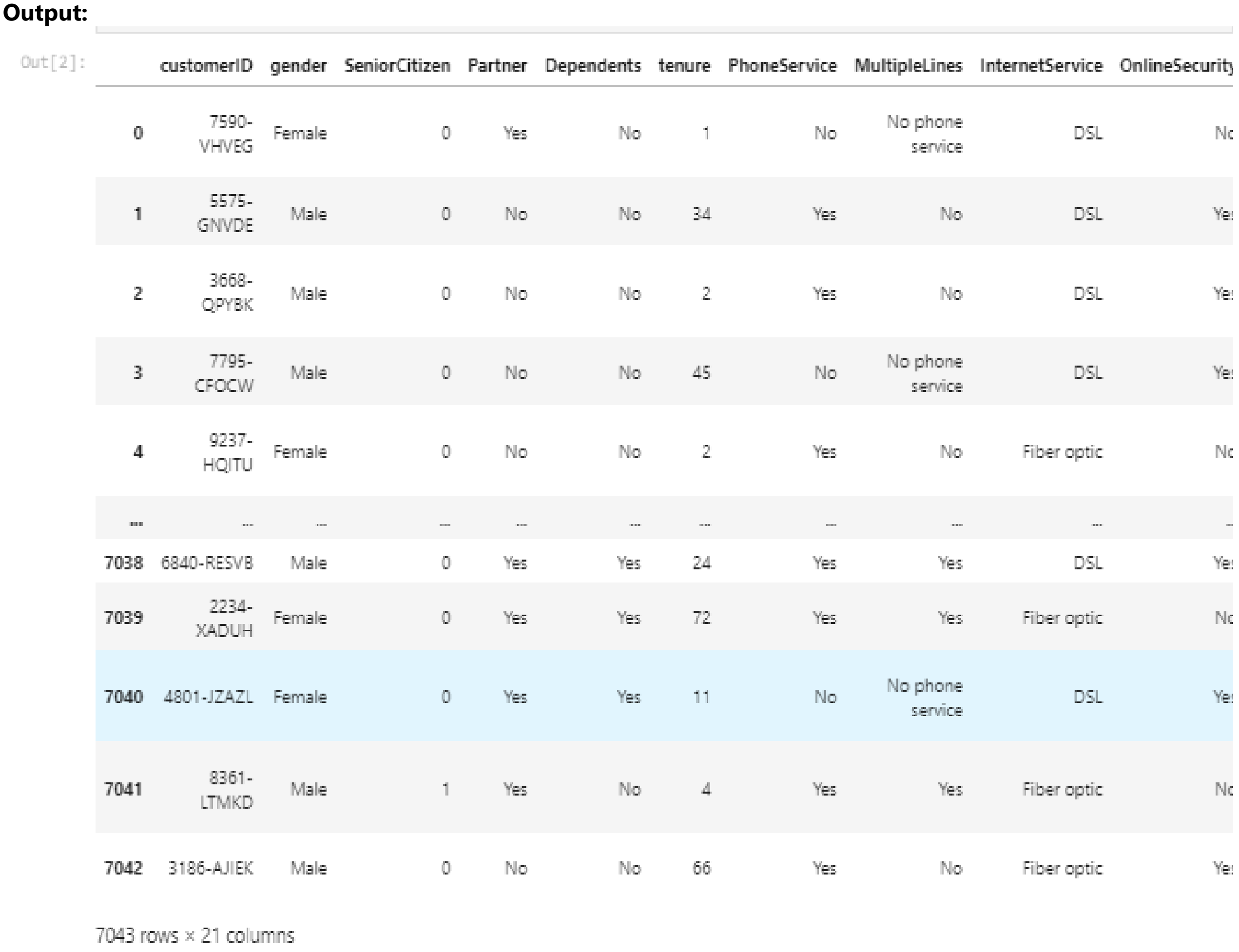
In this research Python program was used where the panda's library was employed for efficient data manipulation and analysis. Besides, matplotlib was employed to generate a wide spectrum of data visualizations, comprising charts and graphs. Furthermore, NumPy equally played a paramount role in computing scientific calculations. Moreover, the sci-kit-learn library played a paramount role in the designing of computational models for this research.

3.5.1 Importing libraries



```
In [1]:  import numpy as np # linear algebra
          from scipy import stats # statistic library
          import pandas as pd # To table manipulations
          import seaborn as sns
          import matplotlib.pyplot as plt

In [2]:  df = pd.read_csv("Churn_dataset.csv")
          df
```

Output:

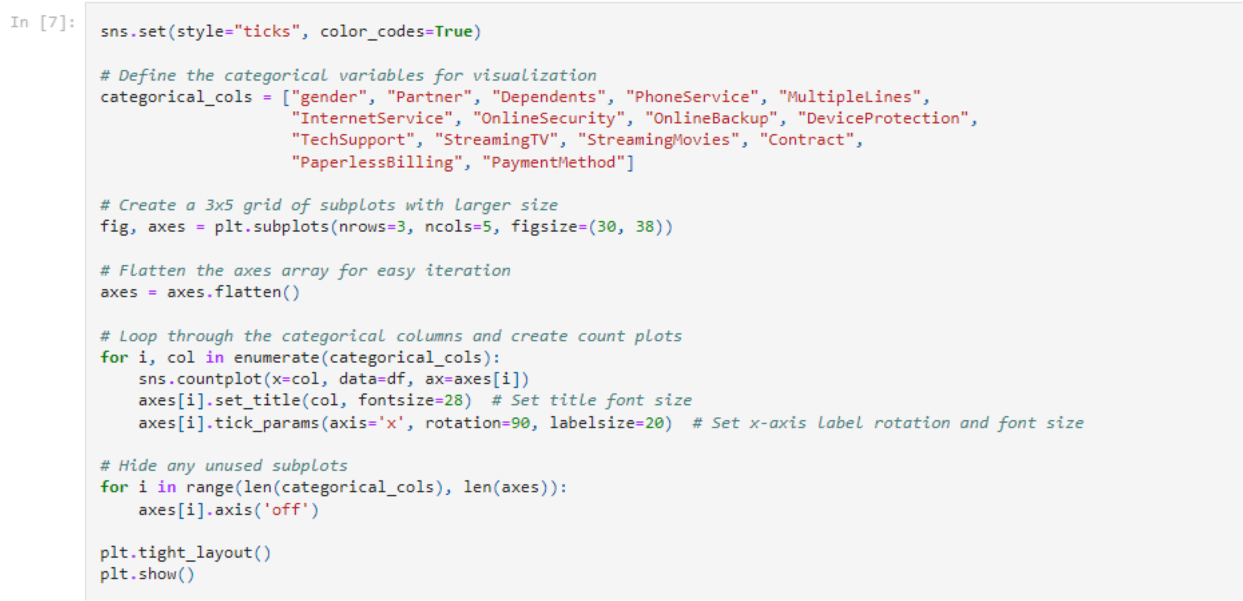


Out[2]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurit
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	Nc
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Ye
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Ye
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Ye
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	Nc
...	...	...	...	...	...	...	...	...	...	..
7038	6840-RESVB	Male	0	Yes	Yes	24	Yes	Yes	DSL	Ye
7039	2234-XADUH	Female	0	Yes	Yes	72	Yes	Yes	Fiber optic	Nc
7040	4801-JZAZL	Female	0	Yes	Yes	11	No	No phone service	DSL	Ye
7041	8361-LTMKD	Male	1	Yes	No	4	Yes	Yes	Fiber optic	Nc
7042	3186-AJIEK	Male	0	No	No	66	Yes	No	Fiber optic	Ye

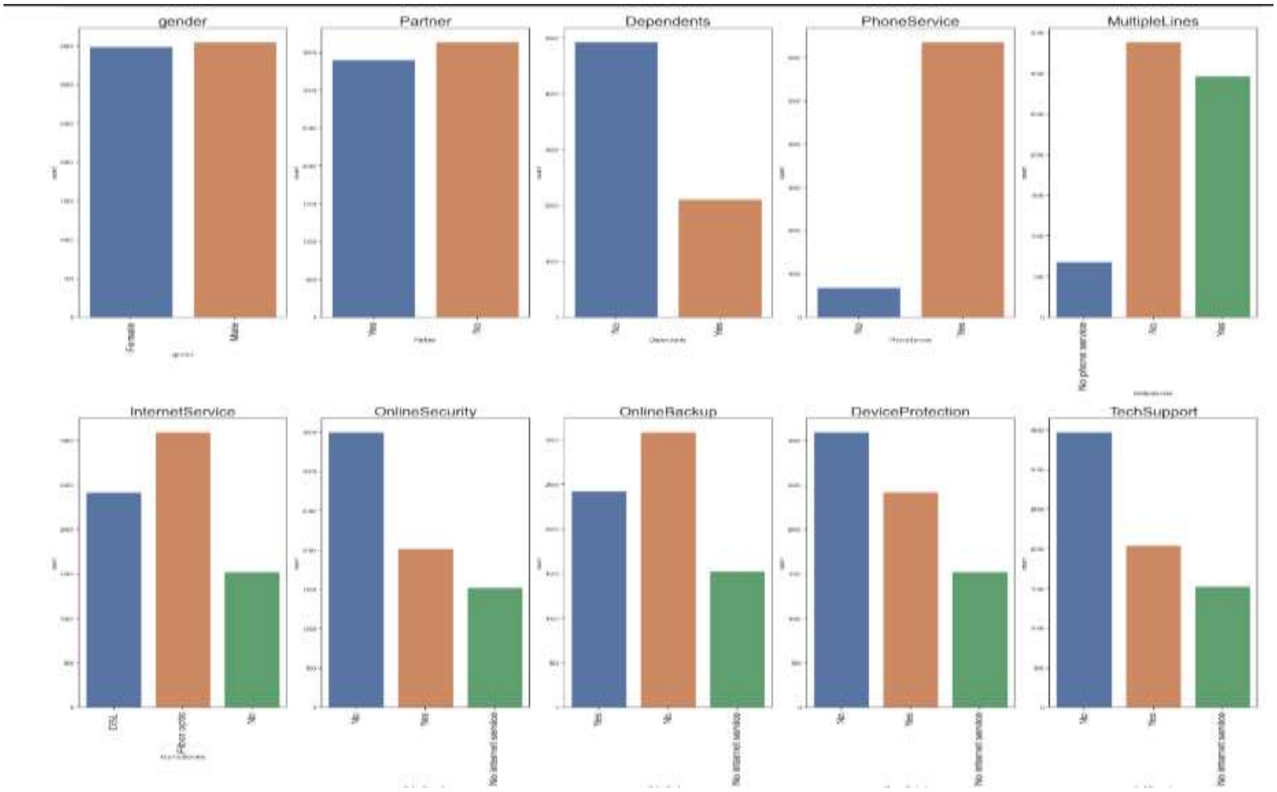
7043 rows × 21 columns

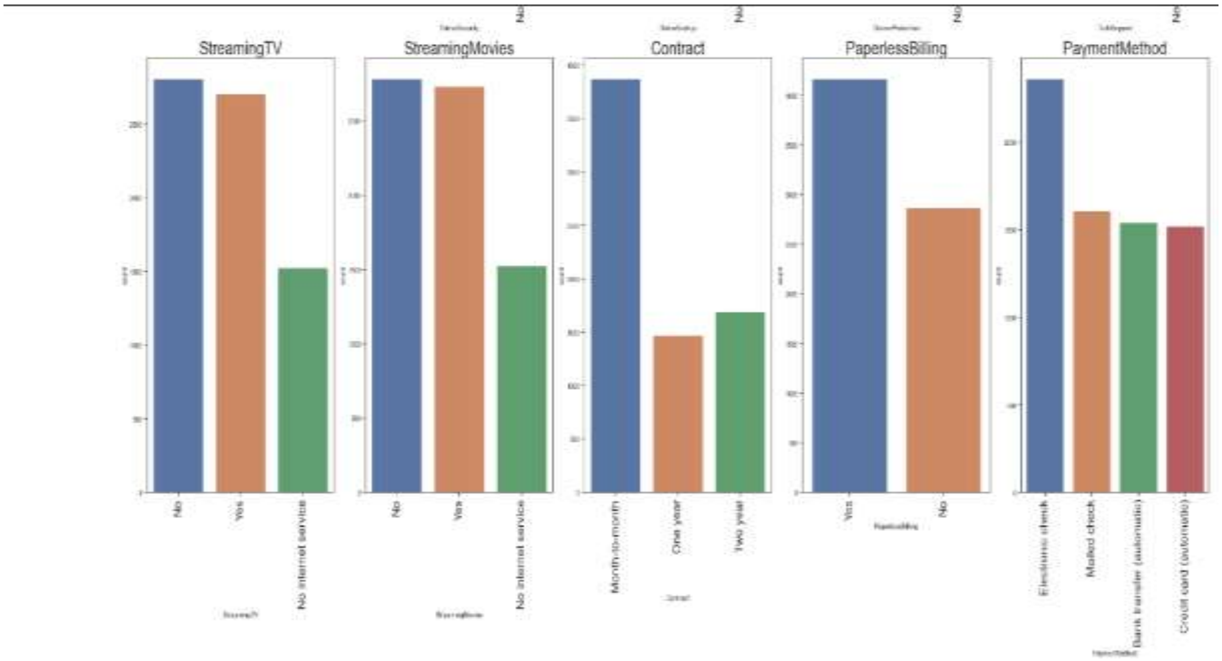
Afterward, the researcher then conducted exploratory data evaluation and analysis to have a comprehensive and in-depth look at the descriptive statistics of the data.



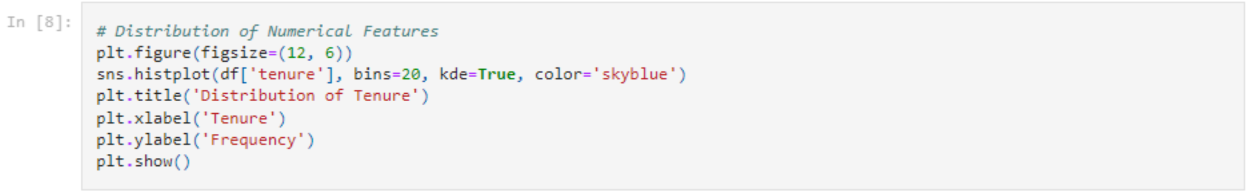
```
In [7]:  sns.set(style="ticks", color_codes=True)

         # Define the categorical variables for visualization
         categorical_cols = ["gender", "Partner", "Dependents", "PhoneService", "MultipleLines",
                             "InternetService", "OnlineSecurity", "OnlineBackup", "DeviceProtection",
                             "TechSupport", "StreamingTV", "StreamingMovies", "Contract",
                             "PaperlessBilling", "PaymentMethod"]

         # Create a 3x5 grid of subplots with larger size
         fig, axes = plt.subplots(nrows=3, ncols=5, figsize=(30, 38))

         # Flatten the axes array for easy iteration
         axes = axes.flatten()

         # Loop through the categorical columns and create count plots
         for i, col in enumerate(categorical_cols):
             sns.countplot(x=col, data=df, ax=axes[i])
             axes[i].set_title(col, fontsize=28)  # Set title font size
             axes[i].tick_params(axis='x', rotation=90, labelsize=20)  # Set x-axis label rotation and font size

         # Hide any unused subplots
         for i in range(len(categorical_cols), len(axes)):
             axes[i].axis('off')

         plt.tight_layout()
         plt.show()
```

Output:



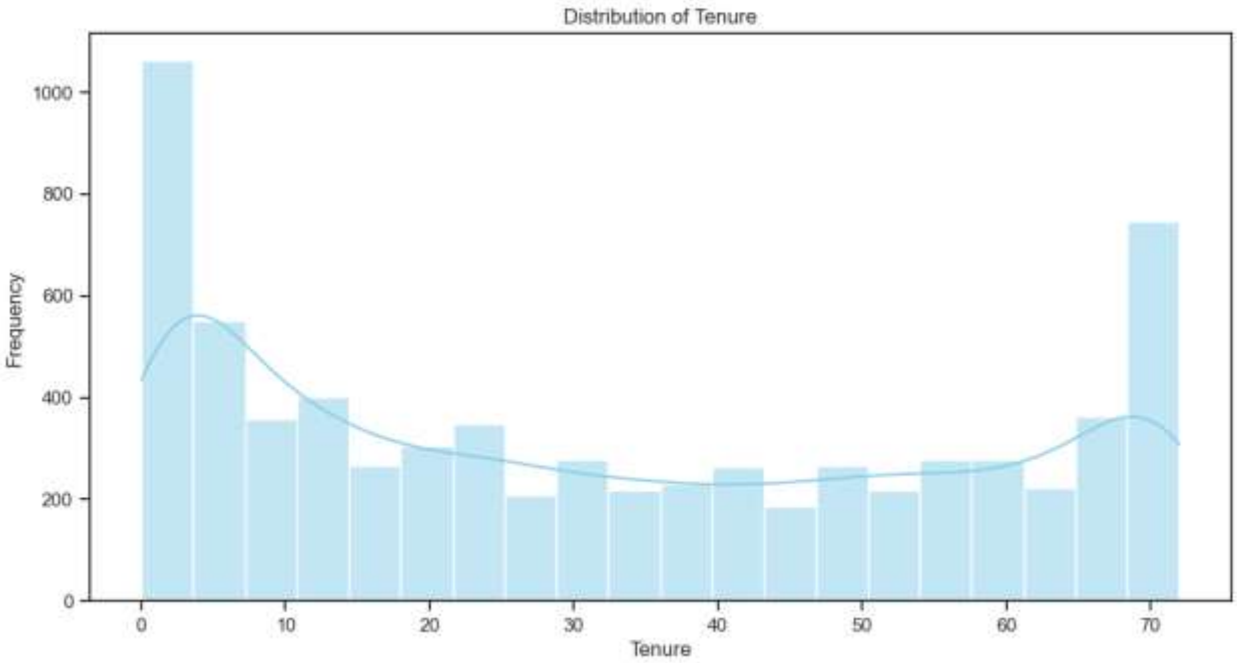


Subsequently, the researcher proceeded to generate histograms to determine the tenure distribution:



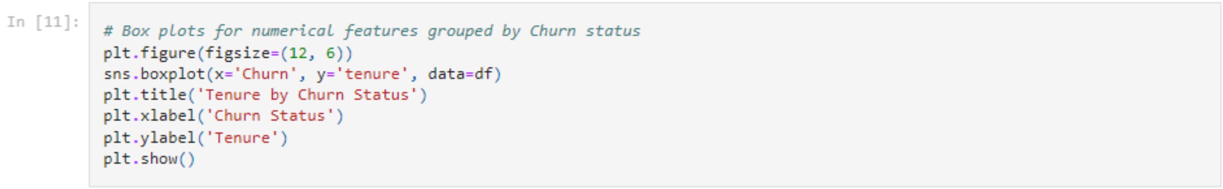
```
In [8]:  # Distribution of Numerical Features
         plt.figure(figsize=(12, 6))
         sns.histplot(df['tenure'], bins=20, kde=True, color='skyblue')
         plt.title('Distribution of Tenure')
         plt.xlabel('Tenure')
         plt.ylabel('Frequency')
         plt.show()
```

Output:



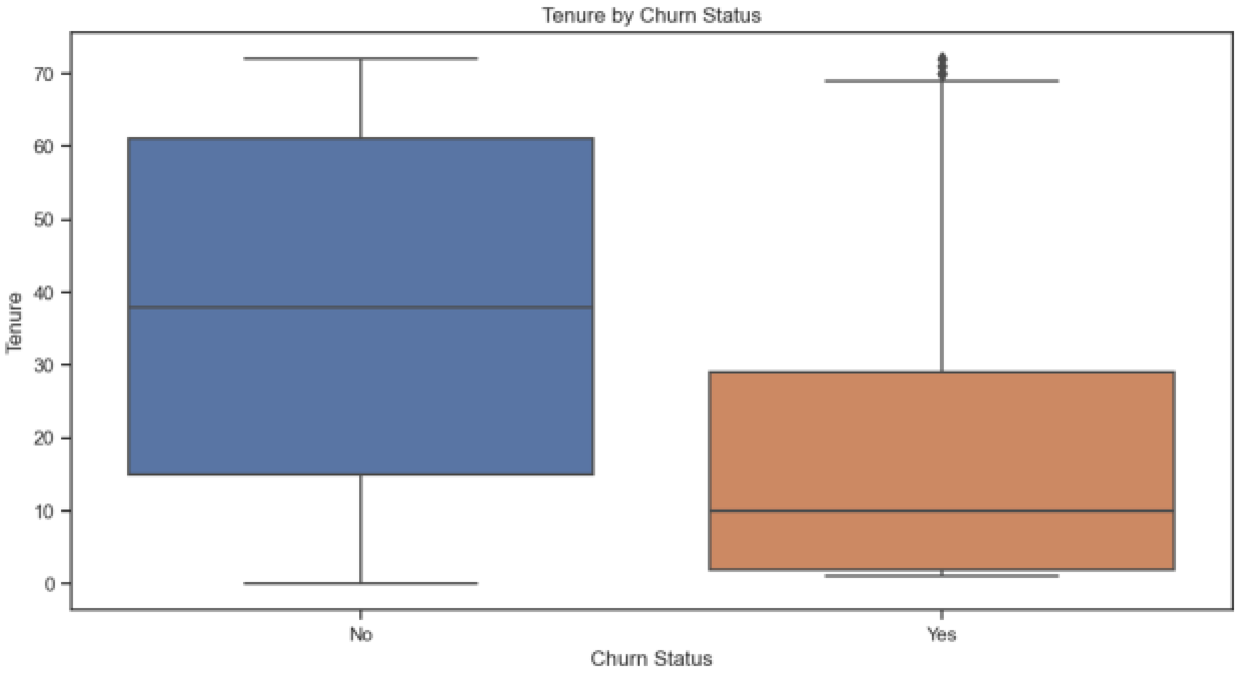
From the histogram above, it was evident that the majority of customers had less than 10 years of tenure, with a peak of around 5 years. There was also a substantial number of customers with tenure between 10 and 30 years and a smaller number of customers with more than 00 years of tenure. The researcher equally established that customer churning was highest in the first few years of the tenure, this was attributed to the fact that some customers leave for new opportunities or because they do not find good services at the company.

Apart from that, the researcher equally targeted to explore the relationship between churn and tenure status:



```
In [11]:  # Box plots for numerical features grouped by Churn status
           plt.figure(figsize=(12, 6))
           sns.boxplot(x='Churn', y='tenure', data=df)
           plt.title('Tenure by Churn Status')
           plt.xlabel('Churn Status')
           plt.ylabel('Tenure')
           plt.show()
```

Output:

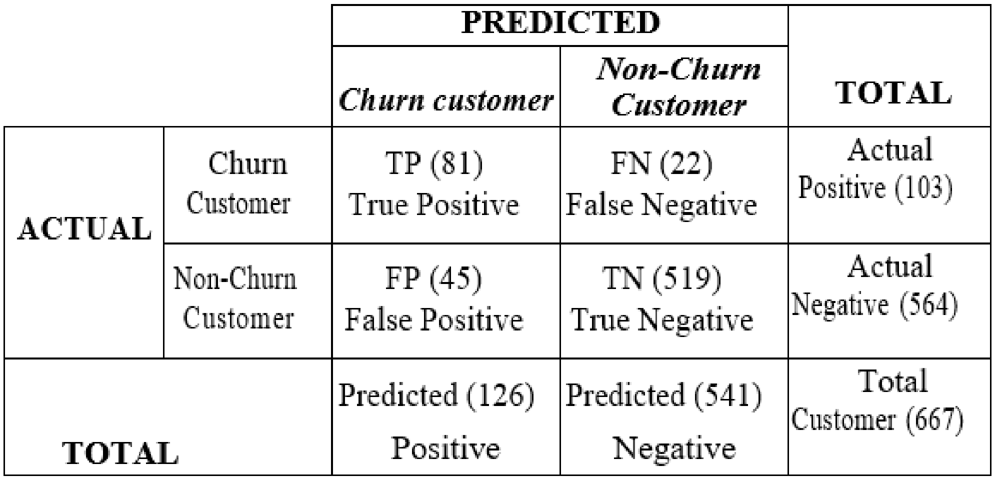


By referring to the above box plot, it was evident that there were few outliers in the "No" churn category, but none in the "Yes" churn group. This indicated that there was a small number of long-term consumers who did not churn and there were no long-term consumers who churned. Overall, the box plot indicated that consumer tenure had a strong predictor of churn for these organizations. Consumers who churn tend to do so relatively early in their customer relationship.

3.6 Models performance Evaluation

Models performance evaluation comprised examining the model's efficiency using a confusion matrix. This matrix is a renowned tool for assessing predictive performance and evaluates the model against a specific testing dataset (20% of the overall data) developed during data pre-processing. It entails four key metrics: false positive (FP), true positive (TP), false negative (FN), and true negative (TN). These metrics are instrumental for consumer retention strategies, affirming accurate identification of clients to retain and evading unnecessary promotions for false consumers.

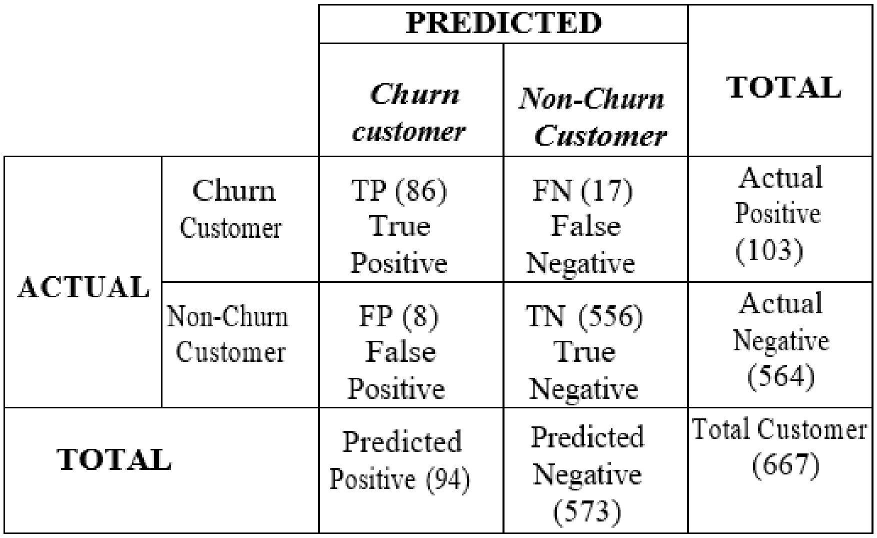
By referring to Table 1, the testing dataset comprised 667 subscribers, with 103 pinpointed as churners (having changed service providers) and 564 identified as non-churners (staying with their present provider). Post-assessment, the prediction demonstrated that out of the 667 subscribers, 126 were predicted to switch providers while 541 are anticipated to stay. The confusion matrix exposed a discrepancy of 23 subscribers between the predicted and actual values.



		PREDICTED		**TOTAL**
		Churn customer	***Non-Churn Customer***	
ACTUAL	Churn Customer	TP (81) True Positive	FN (22) False Negative	Actual Positive (103)
	Non-Churn Customer	FP (45) False Positive	TN (519) True Negative	Actual Negative (564)
TOTAL		Predicted (126) Positive	Predicted (541) Negative	Total Customer (667)

Table 1: Displays Confusion Matrix result for the Decision Tree Algorithm

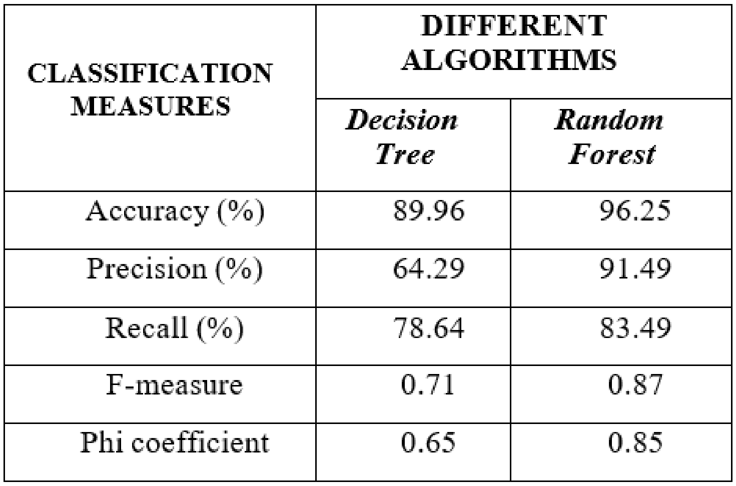
Table 2 below displays the testing dataset of 667 subscribers, applying the random forest classifier algorithm. The dataset encompassed 103 churners (customers who have already changed providers) and 564 non-churners (customers who stayed with the current provider). After the assessment, the algorithm predicted that 94 customers would switch providers, while 573 would proceed with the current provider. The confusion matrix highlighted a discrepancy of 9 subscribers between the predicted and actual values.



		PREDICTED		**TOTAL**
		Churn customer	***Non-Churn Customer***	
ACTUAL	Churn Customer	TP (86) True Positive	FN (17) False Negative	Actual Positive (103)
	Non-Churn Customer	FP (8) False Positive	TN (556) True Negative	Actual Negative (564)
TOTAL		Predicted Positive (94)	Predicted Negative (573)	Total Customer (667)

Table 2: Exhibits the Confusion Matrix results for the Random Forest

3.7 Classification Results



CLASSIFICATION MEASURES	DIFFERENT ALGORITHMS	
	Decision Tree	*Random Forest*
Accuracy (%)	89.96	96.25
Precision (%)	64.29	91.49
Recall (%)	78.64	83.49
F-measure	0.71	0.87
Phi coefficient	0.65	0.85

By referring to the above table and the metrics used for analysis, Random Forest seemed to be a relatively better performing model than Decision Tree for this specific classification task. In particular, it attained higher accuracy (96.25%), precision (91.49), Recall (83.49%), F-measure (0.87), and Phi coefficient (0.85).

3.8 Business Impact

The majority number of businesses in the USA have been using traditional strategies of monitoring churn such as basic metrics like customer surveys, service cancellations, payment defaults, etc., and taking reactive interventions post-churn. However, with the rapid inventions of technologies and the presence of large customer datasets, companies are finding it difficult to navigate the large volume of data which makes algorithms such as Decision Trees and Random Forests Ideal for Churn detection and prevention.

3.8.1 How to Use This Model

Step 1: Data Acquisition

- The process begins with a dataset. This data could be sourced from various sources in the organization and may need preprocessing before being used in the model.

Step 3: Data Splitting

- The dataset can then be subdivided into two parts, most notably, training data and testing data.
- ➢ **Training data** (approximately 80% of the dataset) should be utilized to train the algorithm. Subsequently, the algorithm learns to pinpoint relationships and patterns within the data.
- ➢ **Testing data** (approximately 20% of the dataset) should be used to evaluate and assess the model's performance on unseen data.

Step 4: Decision Tree vs Random Forest Selection

- In this phase, should perform two experiments separately with both decision tree and random forest.

Step 5: Models Evaluation

- After training the models, the algorithms' performance can then be evaluated using the testing data.

Step 6: Comparison of Results

- Subsequently, the results for both the decision tree and random forest should be compared and contrasted and the best algorithm should be chosen.

Step 7: User Interface for Prediction

- After assessment, a user interface should be prepared to allow the data analyst to make churn predictions on customer incidences using the chosen model.

3.8.2 Benefits of Using the Model on the USA Economy

1. **Healthy and Stable Economy:** By assisting companies in America to retain customers, the proposed models can contribute to a relatively healthy and stable US economy. In particular, reduced churn translates to elevated consumer lifetime value for businesses in the USA, leading to potential job growth and investment.
2. **Increased Customer Retention:** By deploying Random Forest and Decision Tree models, government companies can get an in-depth comprehension of the factors that lead to consumer churn. As a result, this information may enable them to tailor targeted retention strategies and interventions. By effectively retaining consumers, government organizations can maintain a stable customer base, leading to sustained revenue and growth.
3. **Cost Savings:** Capturing new clients can be expensive for businesses in the USA. By minimizing churn by deploying the proposed algorithms, government organizations can save costs related to customer acquisition. Retaining current clients is considered to be more cost-effective than continuously seeking new ones. Subsequently, government agencies can allocate resources more effectively, resulting in improved profitability and a stronger economy.

4. Conclusion

This research paper examined the application of machine learning algorithms such as Random Forests and Decision Trees for designing churn prediction models and exploring key factors that churn probabilities. The dataset used in this study was sourced from the prominent UCI repository of machine learning databases, preserved at the University of California, Irvine. This dataset provided extensive information on 3333 clients, facilitating in-depth analysis and insights. Models performance evaluation comprised examining the model's efficiency using a confusion matrix. Random Forest seemed to be a relatively better performing model than Decision Tree for this classification task. By deploying Random Forest and Decision Tree models, government companies can get an in-depth comprehension of the factors that lead to consumer churn. As a result, this information may enable them to tailor targeted retention strategies and interventions. By effectively retaining consumers, government organizations can maintain a stable customer base, leading to sustained revenue and economic growth.

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] Banu, J. F., Neelakandan, S., Geetha, B. T., Selvalakshmi, V., Umadevi, A., & Martinson, E. O. (2022). Artificial Intelligence-based customer churn Prediction model for business markets. *Computational Intelligence and Neuroscience*, *2022*, 1–14. https://doi.org/10.1155/2022/1703696

[2] Baquar, Z. (2023, November 24). How to Predict Churn Risk with Customer Data in Python | Dev Genius. *Medium*. https://blog.devgenius.io/how-to-predict-customer-churn-risk-using-machine-learning-in-python-b11c09759491

[3] Bisoyi, A. (2023). Predictive Framework for Advanced Customer Churn Prediction using Machine Learning. *www.academia.edu*. https://www.academia.edu/110361825/Predictive_Framework_for_Advanced_Customer_Churn_Prediction_using_Machine_Learning

[4] Krishna, V., & Reddy, B. (2023). Machine learning approaches to predict customer churn in the telecommunications industry. *Irjet*. https://www.academia.edu/103798640/Machine_Learning_Approaches_to_Predict_Customer_Churn_in_Telecommunications_Industry

[5] Optimove. (2023, July 18). *Customer Churn Prediction & Prevention Model*. Optimove. https://www.optimove.com/resources/learning-center/customer-churn-prediction-and-prevention

[6] Prabadevi, B., Shalini, R., & Kavitha, B. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, *4*, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005

[7] proAIrokibul. (2022). *Churn-Prediction-And-Prevention/Model/model.ipynb at main · proAIrokibul/Churn-Prediction-And-Prevention*. GitHub. https://github.com/proAIrokibul/Churn-Prediction-And-Prevention/blob/main/Model/model.ipynb

[8] Takyar, A., & Takyar, A. (2023, September 8). *A guide to churn prediction using machine learning*. LeewayHertz - Software Development Company. https://www.leewayhertz.com/ai-and-ml-in-customer-churn-prediction/

[9] Willoughby, R., & Willoughby, R. (2023, December 7). Predicting customer churn using Python. *Data Science Blog*. https://nycdatascience.com/blog/student-works/predicting-customer-churn-using-python/