FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 6, June, 2025, pp 80 – 89
DOI: https://doi.org/10.33003/fjs-2025-0906-3556

INTEGRATING RANDOM FOREST AND LOGISTIC REGRESSION TO FORECAST CUSTOMER ATTRITION IN THE TELECOM SECTOR

***Owoade, Ayoade A., Bakare, Olugbenga I. and Ogunsanwo, Gbenga O.**

Computer Science Department, Tai Solarin University of Education, Ijebu Ode

*Corresponding authors' email: owoadeaa@tasued.edu.ng

ABSTRACT

Since customers are the foundation of any successful business, companies must prioritize making sure they are satisfied. However, due to increased corporate competition, the importance of customers' and marketing tactics more informed conduct in the past few years, client attrition is a significant problem and is acknowledged as one of the top concerns among businesses. Organizations must take a number of measures to address the problems with churn brought on by the services they offer. Customer attrition strategies are essential in the fiercely competitive and rapidly changing telecom industry. Utilizing machine learning methods, assess the possibility that a client will leave a firm. This research uses logistic regression, random forest, and big data to predict customer attrition in the telecom sector. A large-scale logistic regression analysis has been used to assess the probability of churn as a function of a variable set or customer attribute. Similarly, based on how close a feature is to customers in each class, random forest is employed to ascertain if or not a customer churns. This research makes use of information from the Kaggle website to forecast and examine churn. According to the results of the study show that 0.84 percent is the area under the curve., and the forecast precision rates for consumer churn using linear regression and random forest are 0.80 and 0.79 percent, respectively.

Keywords: Random Forest, Machine Learning, Logistic Regression, and Customer Attrition

INTRODUCTION

Since the telecommunications industry is one of the biggest in the world, technological developments and an expanding number of operators have made it more competitive. Telecommunications companies have developed a variety of methods to make substantial profits in order to thrive in this ruthless industry. In order to improve client retention, companies must reduce the likelihood of customer churn, which is the transfer of clients from one service provider to another (Prabadevi et al., 2023). Customer turnover is considered a significant problem in service businesses as the level of competition increases intense (Zhao et al., 2021). Furthermore, machine learning approaches are becoming more and more effective in predicting this customer attrition, according to Srinivasan et al. (2023) and Karamollaoglu et al. (2021). Estimating the amount of clients that are presently utilizing a business and providing solutions to lower large attritions is the basic concept underlying forecasting client attrition within the telecom sector. It is now vital to calculate churners before they discontinue a telecommunications service because of the fierce rivalry among enterprises. Developing prediction methods in addition to churn prediction is even more critical given the vital role that the telecom industry plays. The significance of user retention in this industry is only partially supported by research (Shields, 2021). Shields (2021) asserts that a 1% increase in customer retention movement might potentially lead to a 5% increase in the total number of firm shares. In the telecom industry, the yearly rate of customer attrition was 27%, with a monthly ratio of 2.2%, according to Wagh et al. (2023).

However, maintaining clients in the telecommunication company has become a nightmare because of the emergence of competing services (Wagh et al., 2023). An inventive data mining technique to detect customer attrition was proposed by Ahmad et al. (2019) utilizing machine learning methods such as NN (Neural Networks) and SVM (Support Vector Machine). The results showed how much more accurate machine learning algorithms are in forecasting customer attrition. Sana et al. (2022) adopted a machine learning model to address the problem of customer attrition in major telco businesses. Based on the volume, velocity, and variety of the data, they showed how big data significantly enhances the process of determining client attrition. In 2020, Li and Zhou investigated the issue of client attrition in big data platforms. The characteristics of the social network analysis application enhance the results of assessing attrition in the telecom industry. (Ahmad and others, 2019).

At between 20 and 40 percent, the churn rate in the telecommunications sector has been the greatest of any other industry (Ribeiro et al., 2023). This essentially affects the company's finances because five to 10 times as much is spent to gain a new user as it does to retain a current one (Petropoulos et al., 2022). Today's businesses want to build trusting relationships with their clients (Mandal, 2023). Therefore, it is now thought that the most effective marketing approach is to hold onto existing consumers, or, to put it another way, to address customer attrition (Shobana et al., 2023). Although machine learning approaches have improved prediction accuracy when applied to huge, highly dimensional, nonlinear datasets, they are still perceived as challenging to apply in real-world scenarios (Aghaabbasi and Chalermpong, 2023). For example, ANN, a well-liked machine algorithm technique, are used to solve the churn prediction problem (Saghir et al., 2019). Supervised learning techniques called SVM analyze data and identify trends (Sarker, 2021). Regression and classification analysis are their main applications. Decision trees (DT) are another machine learning application, however they are not very good at identifying intricate and non-linear correlations between characteristics.

Neural networks are expected to outperform decision trees in comparison, according to Lu et al. (2022). However, a Decision Tree's accuracy level regarding the problem of customer turnover may be higher depending on the data type (Zhao, 2023). Comparably, the Naive Bayes classifier has demonstrated superior prediction rates compared to other popular algorithms such as Decision Tree (Ahmad et al. 2019) and has yielded positive findings for the telecom sector's churn prediction challenge (Wagh et al., 2023). Since The random forest method can complete the work, and the logistic regression algorithm is employed to calculate the risk of attrition without any prior knowledge of the variables set or client attributes of classification, the dissemination of data, both algorithms were chosen for this study. It is useful to compare the two algorithms in order to predict customer churn and address the factors that affect client retention. The algorithms Random Forest and Logistic Regression outperform others based on many benchmark data sets.

Models of machine algorithms were compared by Karamollaoğlu et al. (2021). The study found that neural networks outperform logistic regression and decision trees by a little margin. A comparison research on forecasting customer churn discovered that the accuracy levels of decision trees and BPN were 94%, SVM was 93%, and logistic regression was 86% conducted by Lu et al. (2022). SVM approaches are believed to have better predictive performance and have been extensively studied for predicting customer attrition (Nguyen et al., 2022). According to Wibawa et al. (2019), Naive Bayes evaluates the probability that a user will remain with their present service provider or choose to use a different one.

The book written by Fredrick Winslow "The Principles of Scientific Management" popularized early 1900s scientific management practices, is credited with helping to establish big data (Dar, 2022). More modern data analytics and dashboards are preceded by data visualization (Michele et al., 2019). In the future, database technology will be needed to make sense of enormous volumes of data management, data gathering, Data analysis and extraction were employed (Dash et al., 2019). (Ku-Mahamud and Basim Alwan, 2020) "Big data" explains the event and examination of "large" data collections. In the same context, numerous scholars have asserted that technological advancements are what drive and advance data processing and analysis.

Favaretto et al. (2020) provided a comprehensive definition and description of the methods used to manipulate geocoded data; they are more helpful for analysis than for strategy development and even less helpful for decision-making. Furthermore, Saini and Bansal (2023) consider geomarketing to be an effective marketing tactic that aids decision-makers in resolving certain pressing concerns.

As big data becomes more popular, a worldwide strategy or roadmap that helps IT divisions not just implement big data efforts but also maximize their usage to accomplish corporate objectives is becoming more and more necessary. It looks at two basic types of data: transactional data, which is utilized to transmit purchasing data as well as structured data, often referred to as relational information, which is the compilation of logs of transactions that a business may generate. However, incidental or non-transactional data are formed in reaction to a transaction; they would be considered disorganized data if they lacked structure.

Data is an intellectual property or intangible good. In the 1800s, data administration was described as the creation, implementation, and oversight of plans, policies, programs, and procedures that regulate, safeguard, provide, and increase the value of data and information assets by the Data Management Body of Knowledge (DAMA-DMBOK). Over the past 20 years, the big data age has expanded significantly across numerous industries. Big data analytics has transformed the contemporary marketing process (Islam, 2024). Furthermore, Gartner predicts that big data will soon be accepted as the norm (Dash et al., 2019). The customer has taken the lead, and the main factor influencing the transaction represents the possible worth of the huge data on consumer behavior. The early 2000s saw the rise in popularity of big data. In order to effectively extract value from very large volumes of diverse data, big data technologies are a new breed of architectures and technologies that allow high-velocity data gathering, discovery, and/or analysis (Batko & Ślęzak, 2022).

Roberts (2023) defines big data as 3V, which stands for volume, velocity, and variety. According to Jabeen (2020), Big data technologies are used to economically derive useful information from enormous volumes of varied data. A new breed of technologies and architectures that allow for high-velocity acquisition (streaming data), discovery, and/or analysis, up to petabyte volumes (Berisha et al., 2022). Data quantities that fall within the terabyte, petabyte, exabyte, and zettabyte categories form the foundation of the continuum of data, regardless of how big or small it is. While data variety can be divided into organized, unstructured, and semi-structured categories, data velocity can be divided into batch, real-time, and flow categories.

Cavlak and Cop (2021) define big data as an extremely high volume of unstructured information that comes from a wide range of sources, such as websites, surveys, social media, chat forums, Twitter trends, Facebook, scholarly publications, and more. Its quick inclusion in many formats enables tools for processing and analyzing database administration. Data analytics is the study of current data to examine the corporate environment, including customer behavior, and to investigate marketing trends and potential customers. The largest issue is figuring out how to sort through the enormous amount of data collected from multiple sources in order to uncover all the hidden information. Conventional market research sometimes includes surveys, focus groups, and interviews conducted at malls. Big Data looks at what people say about their behavior in the past, present, and future. In addition to monitoring people's actions, it's also a good idea to predict future trends. This is helpful because of Big Data's capacity to manage massive volumes of unstructured, real-time data (Udeh et al., 2024).

MATERIALS AND METHODS

In a corporate context, predicting customer churn is used to describe a company's attempts to retain customers that are more prone to stop their use of its services. By classifying which consumers are likely to leave and which won't, this study lowers the churn rate. Since it can be challenging to obtain new customers, keeping existing ones is essential.

Attrition can be decreased by carefully examining the previous actions of the most significant clients. Potential churning customers can be identified by appropriately assessing the substantial amount of customer data that is controlled. Operators have a number of alternatives for anticipating and preventing customer churn by analyzing the feasible data in different ways. Figure 1 illustrates the procedures used for the suggested system.

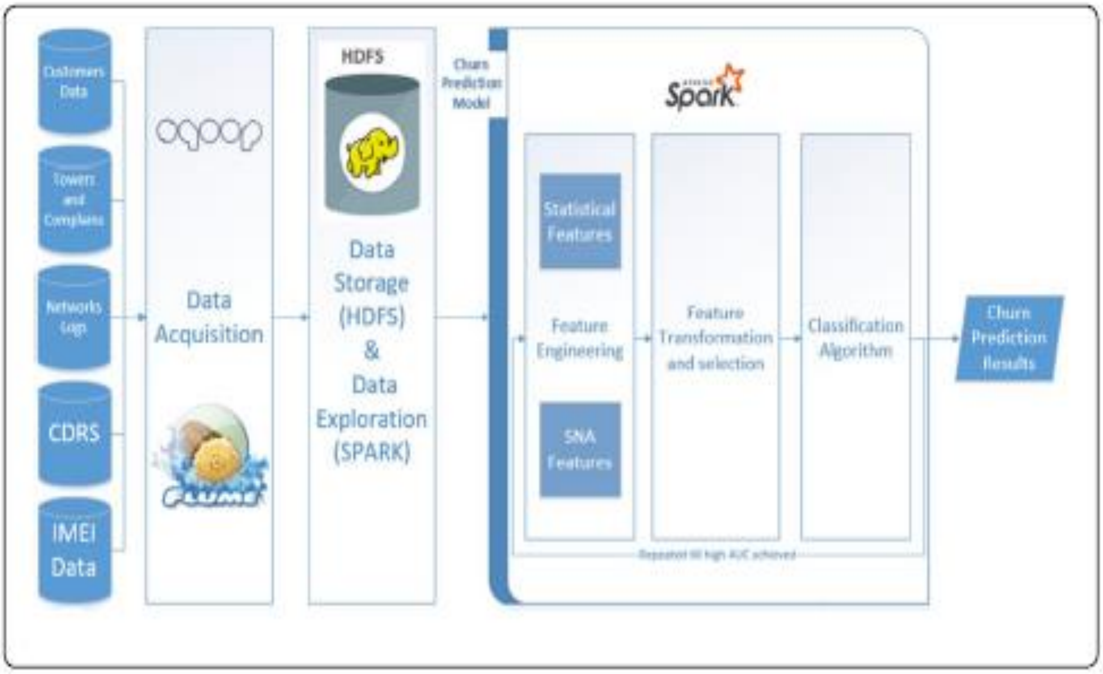


Figure 1: Proposed churn prediction system architecture (Source: Ahmad et al., 2019)

Procedures for The Suggested System

Data Gathering

The information that can be analyzed in a telecommunication dataset was gathered and used to create predictions.

Data preprocessing

Three steps constitute preprocessing of data: feature selection, data transformation, and data cleaning. The actions listed below are utilized: To better fit the selected models, two explanatory variables in the data transformation procedure is capable of being transformed from binary to binomial. Imputation, or dealing with missing data, is a step in the data cleaning process. Since missing data cannot be handled by some of the chosen techniques; values that are unavailable can be translated using the median, mean, or zero. Nonetheless, it is better to use statistically calculated values to fill in the missing data. The used data set contained missing values for a few numerical variables and two categories. Among the most important components which could impact the model's performance prior to instruction is chosen.

Data Preparation

Data preparation's primary objectives revolve around speed up data processing and enhance data quality. To arrive at a definitive conclusion, the data must be prepared, imputation of missing values, discretization of numerical variables, feature selection of the most valuable variables, switching between discrete value sets, and variable generation are all examples of data preparation techniques. Imputation is the process of using an estimate from the final values to replace the missing values with the whole set of data. The basis for generating fresh variables generated by data is transformation and discretization. Two new factors were developed in order to measure the voice and transformation in data use. Prior to reviewing the data, it needs to be cleaned and prepared so that the right results can be obtained from it. In order to avoid producing inaccurate results, data must be accurate in order to eliminate unnecessary information and mistakes. The attrition assessment of the telecom data in this study aims to find possible clients who may switch service providers. The outcome demonstrates how likely that every client will depart. Using logistic regression, the churn analysis is carried out. Instead of using a continuous output variable, a statistical technique known as logistic regression employs a categorical one. For logistic regression, the range of possible outcomes is restricted to zero and one.

Data Prediction

Given the company's emphasis with the end result, it is critical to present the outcomes in an intelligible virtual format so that it can provide the necessary forecasts and generate revenue. Several elements work together to achieve the same objective.

Data Visualization Tools

Using visualization tools is the best way to communicate findings. Important patterns that would be missed if only statistics were considered can be discovered by visually representing data. Jupyter is one of the data visualization tools used in this study. Reports can be produced using Microsoft's Jupyter business analytics application. Data from the research is currently been cleansed, and the findings have been added to a prediction file, which will be utilized to graphically depict the appearance of the data and importance.

Dataset Analysis

The work predicts and analyzes churn using datasets from the Kaggle website. Kaggle's open-source online dataset served as the inspiration for the dataset. The dataset initially had 21 columns and 7043 rows. ML tournaments are held on the Kaggle website and community. Rivalry machine learning might be the best method for them to explain and refine their skills. Additionally, users can compete to overcome the challenges of data science. For this dataset, pre-processing procedures include:

i. Adding null values to the blank spaces in the total charges column;
ii. lowering the percentage of null values in the total charges column, which has 15% missing data; and
iii. transforming the data into a float format. Afterwards, the categorical column is changed to tenure, separated are

the numerical and category columns, as well as the dropout and non-dropout clients.

iv. and the values for SenioCitizen are changed to 0 for No and 1 for Yes.

Model Construction

Many methods had been suggested for forecasting the loss of clients in the telecom company. This study will make use of Logistic Regression and Random Forest to evaluate the suggested feature subsets.

Random Forest

To predict client attrition, the Random Forest machine learning technology will be employed. It is an amalgam of various decision trees. It uses the bagging methodology to get results and is a decision tree ensemble learning technique for classification and regression issues. Weak learners band together to create stronger learners in group learning. Bootstrap Aggregation, often known as bagging, is employed to reduce the Decision Tree's variation. It is an ensemble approach to machine learning that generates accurate results by combining predictions from different machine learning algorithms. Probst et al. (2019) claim that Random Forest's default hyper-parameters are quite effective at avoiding overfitting and yield good results. In Random Forest, the output that was consistent across all decision trees was used to determine the final projected class. During the training phase, it constructs an enormous quantity of decision trees and shows the class as the mode or the average prediction for every tree. Because Random Forest is fast and can handle missing and unbalanced data, it will be used in this study to forecast customer attrition (Qin et al., 2022).

Logistic Regression

This is the best regression analysis model to use when there is a variable that is binary-dependent. Logistic regression has been used to assess the likelihood of customer turnover based on a set of customer behaviors or characteristics (Sultan et al., 2021). Jain et al. (2020) claim that logistic regression is also utilized to calculate the probability that a customer will depart. Despite being used in a variety of settings, such as the context of pediatric ADHD (Mooney et al., 2022), logistic regression has also been used in customer analysis. Martínez et al. (2020) forecasted partially faulty clients in a retail setting using logistic regression. Multinomial regression has been used to predict the client's future profitability based on their demographic information and past book club purchases (Sleiman et al. 2022). A logistic regression can only employ one dependent variable. Logistic regression employs maximum likelihood estimation following the conversion of the dependent into a logistic variable (Idriss et al., 2023). The logistic regression mathematical formulas are:

$$P(b) = 1|\alpha_i, \ldots\ldots, \alpha_m) = F(b)$$

$$F(b) = \frac{1}{(1+e^{-b})}$$

$$b = \beta_0 + \beta_1\alpha_1 + \beta_2\alpha_2 + \cdots + \beta_m\alpha_m$$

Where β_0 is a constant, b is every individual e target variable, and y is a binary label class one or zero, $\alpha_1 . \alpha_2, \ldots . , +\alpha_m$ is the variables of predictor for every customer e from which α is to be predicted.

After the client datasets are evaluated to produce the regression equations, the process of evaluating each customer in the dataset is finished. A customer could end up churning if their p value is higher than a predetermined value. A threshold value will be used to divide the logistic regression model's continuous probability outcomes into two categories. In this study, churners and non-churners will be distinguished by the threshold value. Typically, 0.5 is chosen as this cutoff value.

Implementation of the system

The telecom operator uses a big data platform to build a churn prediction system. The Hortonworks data platform, or HDP, has become more well-known since it is a free and open-source framework. The Apache 2.0 license governs its use. Numerous tools and open-source software are part of this data platform. These tools are used in conjunction with freely available software. Each HDP tool group is classified based on a particular field of expertise, such as governance integration, operations, security, data access, and management of data. The application of machine learning in large data systems for telecom customer churn prediction will be provided in this research, which combines software and hardware resources.

System Requirements

For the software to run correctly, the aforementioned operating system and hardware specifications are essential:
The necessary pieces of hardware:

i. 32 kb RAM capacity.
ii. 2 GB hard disk capacity.

Voltage stabilizer

This facilitates the computer system's ability to control the voltage required to avoid system electrical destruction.

Software Requirements

Microsoft Disk Operating System (Windows)

The graphical operating system Microsoft Windows was developed and made available. It is commonly referred as Win or Windows. It provides a way to access the Internet, save files, run programs, play games, and watch videos.

Phyton 3.9

Python programming language is a versatile language. This programming language can be used to develop desktop and web-applications. Additionally, Python is utilized to create complex mathematical and scientific applications. Python is designed with essential features to enhance data visualization and analysis.

Evaluation of the Results

Four metrics will be used in this study to assess the forecasts' accuracy. Accuracy, F-score, precision, and recall are the four metrics. The ratio of correctly identified cases to all correctly and mistakenly classified cases is known as precision.
The following is an explanation of the precision equations:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

The proportion of correctly categorized cases to all correctly classified and unclassified instances is known as the recall. Recall is represented mathematically by the following equation:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The F-score, which is thought to be a useful measure of the link between precision and recall metrics, is obtained by combining them. Here's an illustration of it.

$$F\ score = \frac{2^{Precision\ c * Recall\ c}}{Precision\ c + Recall\ c}$$

In a similar vein, accuracy gives the percentage of all predictions that were computed properly.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + False\ Positve + True\ Negative}$$

Discussion of Findings

This section examines and explores the use of big machine learning data approaches, specifically logistic regression and random forest, to estimate customer attrition in the telecom industry. The effectiveness of the logistic regression and random forest algorithms in reducing churn rates in the telecom industry is also covered in this chapter.

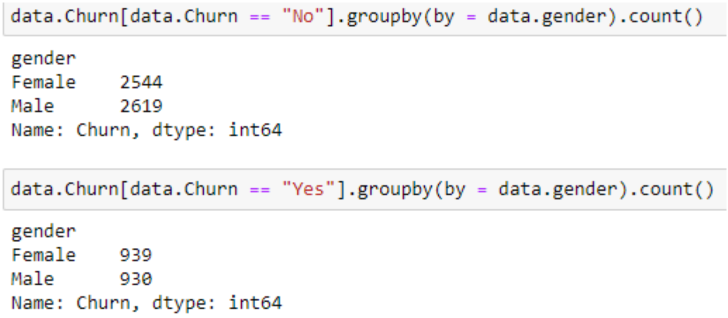
The study was carried out using the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. It provided a standard design and principles for data miners. This methodology consists of five steps: data collection, pre-processing, preparation, prediction, and visualization tools. Chapter 3 on System Design covered the data collection phase.

Pre-Processing of Data

Careful preparation and a deep understanding of the data were required to conduct machine learning research. The CRISP-DM process's second stage is all about gathering data, analyzing it, assessing its quality, and making inferences from it. The original telecom customer data was obtained via the Kaggle website. A thorough statistical and visual analysis of the data was conducted using Python.

Dataset

The telecom customer dataset comprises 21 variables with information from 7043 entries. A boolean variable known as the closed target variable reveals whether a customer was kept or churned. The descriptive analysis of the turnover rate by gender is displayed in Figure 2.



```
data.Churn[data.Churn == "No"].groupby(by = data.gender).count()

gender
Female    2544
Male      2619
Name: Churn, dtype: int64

data.Churn[data.Churn == "Yes"].groupby(by = data.gender).count()

gender
Female    939
Male      930
Name: Churn, dtype: int64
```

Figure 2: Descriptive analysis of churn rates by gender

The telecom industry's churn rate distribution is depicted in Figure 3. 73.37% of consumers did not churn or stop utilizing the company's services, however 26.58% of customers did, according the statistical analysis of the telecom dataset.

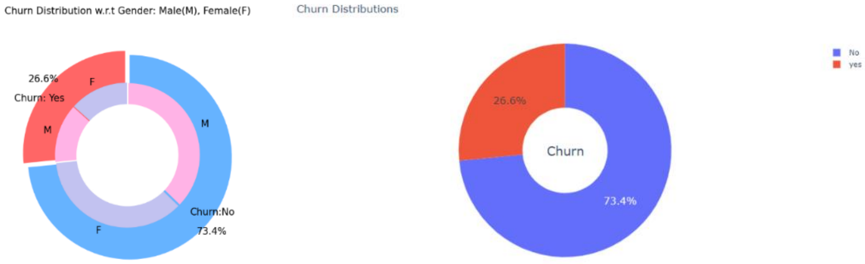


Figure 3: Distribution of churn rates

Correlation Analysis

The correlation analysis of the data used is shown in Figure 4 in order to investigate the link between the independent and dependent variables as well as to identify multicollinearity among the independent characteristics. Since this includes both continuous and categorical data, the "Spearman" correlation technique was used to determine correlation.

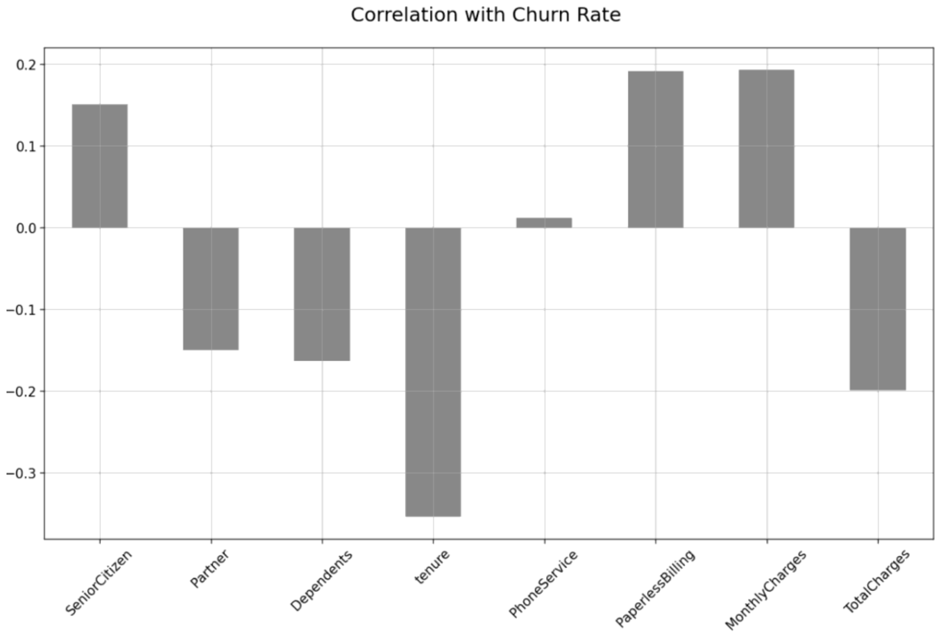
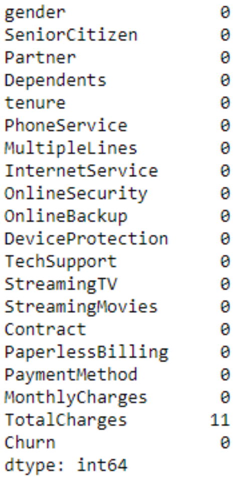


Figure 4: Correlation bar of the variables

The correlation bar in figure 4 showed that the variables "MonthlyCharges" and "TotalCharges" had the strongest link with the churn rate. Its connection with the target variable is weak.

Outlier Analysis

Check Figure 5 for any data outliers."TotalCharges" was the variable that was analyzed. Analysis was used to establish whether the outliers belonged to the minority class because of the extreme imbalance in the data. Since doing so may cause information loss, no outlier that is presented in a minority class can be removed. The output was created with the turnover rate in mind to give information on outliers.



```
gender               0
SeniorCitizen        0
Partner              0
Dependents           0
tenure               0
PhoneService         0
MultipleLines        0
InternetService      0
OnlineSecurity       0
OnlineBackup         0
DeviceProtection     0
TechSupport          0
StreamingTV          0
StreamingMovies      0
Contract             0
PaperlessBilling     0
PaymentMethod        0
MonthlyCharges       0
TotalCharges        11
Churn                0
dtype: int64
```

Figure 5: The outlier of the dataset

Eleven entries lack the Total Charges, according to the investigation. Based on the data collection from the previous phase, many data pre-processing processes were carried out to produce the final dataset for the experiment. During the pre-processing step, missing values were fixed, the data were encoded, normalized, features were selected, features were extracted, the data were standardized, and lastly, the data were separated.

Handling Missing Values

When the descriptive analysis was done on the telecom customer dataset, no missing values were discovered. It was observed that the telecom customer dataset contained 7043 records with the value "True". All of those records from the dataset were visualized using the Python drop command. Figure 6 provides a detailed summary of the dataset that was used.

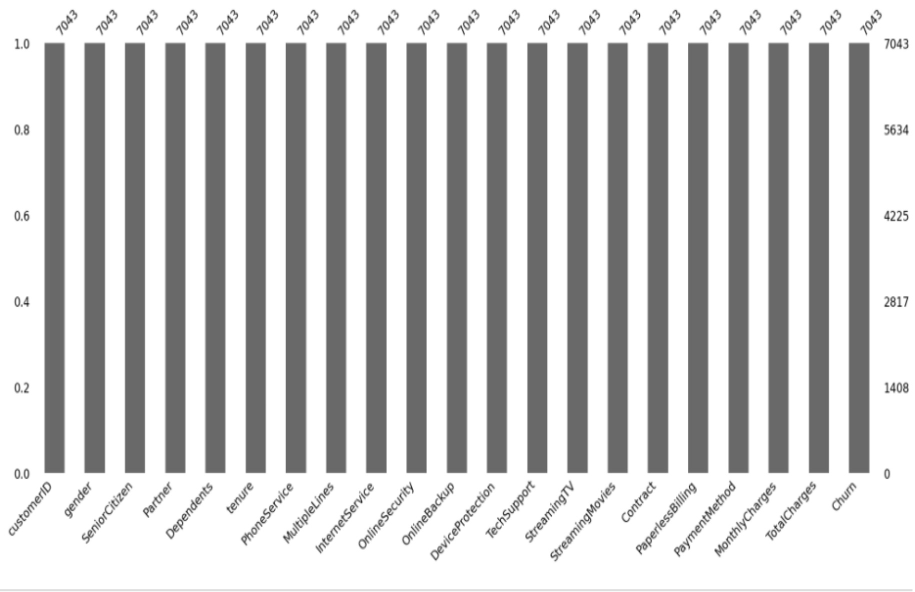


Figure 6: The overview of the dataset

Modelling

The Logistic Regression method and the sklearn.linear_model class were used to generate the Logistic Regression model in Python. Sklearn, often known as Scikit Learn, is an open-source machine learning package for Python. There are many different supervised and unsupervised learning methods available.

The Random Forest was built using the Python sklearn.ensemble class. The Random Forest algorithm is based on ensemble learning. Ensemble learning uses a variety of machine learning models to enhance dataset predictions. Table 1 and Figures 7 and 8 show the classification results for the logistic regression and random forest approaches. Precision, recall, and f1-score are used to quantify accuracy.

Table 1: Classification outcomes

Model	Recall	Precision	F1 Score
Logistic Regression	0.559614	0.658171	0.605009
Random Forest	0.508010	0.650574	0.570560

The classification results for the logistic regression utilizing recall, precision, and F1-score to assess accuracy are displayed in Figure 7.

Table 1 shows that the dataset's accuracy value for the logistic regression is greater than the F1 and recall score. Recall and F1 are less significant than accuracy.

The random forest classification outcomes are shown in Figure 8, with recall, precision, and f1-score indicating accuracy.

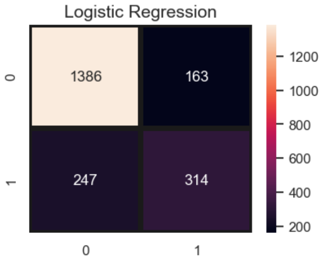


Figure 7: Logistic Regression for the telecom dataset

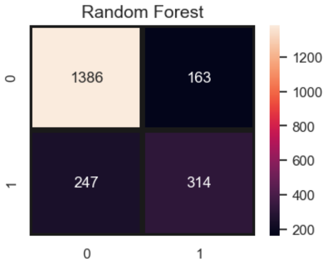


Figure 8: Random forest for the telecom dataset

Comparison of Accuracy of Logistic Regression and Random Forest Dataset of Algorithms

Table 2 shows the comparison of the two models used in this research work.

Table 2: A comparison of the accuracy of the models

Model	Accuracy
Logistic Regression	0.805676
Random Forest	0.796672

It is clear from Table 2 that the logistic regression's accuracy has outperformed the random forest classifier's accuracy.

RESULTS AND DISCUSSION

Predicting customer attrition in the telecom sector was the main emphasis of the study's findings. The personality of the customer determines a company's capacity to enhance customer service and inspire the telecom industry's overall performance. The study's expected results show that the classification accuracy of the proposed method has improved. Thus, it can be concluded that the logistic regression classifier's accuracy has exceeded the random forest classifier's.

The objective of this study was to compare the effectiveness of Logistic Regression and Random Forest models in forecasting customer attrition within the telecom sector. Following extensive data pre-processing, including addressing missing values, encoding, normalization, and feature engineering, both models were trained and evaluated on the prepared dataset.

Classification Performance

Table 1 provides a detailed breakdown of the classification performance for both models, focusing on Recall, Precision, and F1-Score.

The Logistic Regression model exhibited a Recall of approximately 0.560, indicating it correctly identified about 56% of actual churners. Its Precision stood at roughly 0.658, meaning that about 65.8% of customers predicted to churn by the model actually did. The F1-Score for Logistic Regression was 0.605, representing a balanced measure of its precision and recall.

In contrast, the Random Forest model showed a slightly lower Recall of approximately 0.508, identifying about 50.8% of actual churners. Its Precision was comparable to Logistic Regression at around 0.651. The F1-Score for Random Forest was 0.571, also slightly lower than that of Logistic Regression.

While the document emphasizes accuracy as the primary metric, the F1-Score and Recall are crucial for churn prediction. A high Recall is vital for telecom companies as it minimizes false negatives, ensuring that fewer customers at risk of churning are missed, allowing for proactive intervention. Conversely, high Precision reduces false positives, preventing the wasteful allocation of resources on customers who are not actually at risk. The Logistic Regression model demonstrates a better balance between these metrics in this specific context.

Overall Model Accuracy

Table 2 presents a direct comparison of the overall accuracy achieved by both models. As shown, the Logistic Regression model achieved an accuracy of 0.805676, slightly outperforming the Random Forest model, which recorded an accuracy of 0.796672. This indicates that Logistic Regression correctly classified a marginally higher percentage of customers (both churners and non-churners) compared to Random Forest.

Discussion

The results suggest that, for this particular telecom customer dataset, Logistic Regression was marginally more effective than Random Forest in overall churn prediction accuracy. This finding is notable because Random Forest, as an ensemble method, often excels in capturing complex, non-linear relationships within data, which are typically prevalent in customer behavior. However, Logistic Regression, despite its simpler, linear assumptions, performed commendably. This could imply several factors:

i. Linear Separability: The underlying relationship between the features and customer churn in this specific dataset might be predominantly linear or close to linear. In such cases, Logistic Regression can perform very well and sometimes even surpass more complex models.

ii. Data Characteristics: The extensive pre-processing, including normalization and feature engineering, might have transformed the data in a way that made it more amenable to Logistic Regression. Well-engineered features can significantly boost the performance of simpler models.

iii. Hyperparameter Tuning: While both models were presumably optimized, slight differences in hyperparameter tuning efficacy could contribute to the observed performance gap.

iv. Interpretability vs. Complexity: Logistic Regression offers greater interpretability, allowing businesses to understand the direction and magnitude of the impact of various factors on churn probability. While Random Forest is powerful, its "black box" nature can make it harder to extract actionable insights directly from the model's structure. In scenarios where interpretability is highly valued alongside predictive power, Logistic Regression gains an advantage.

Despite the marginal difference in accuracy, both models demonstrate reasonable performance in identifying customer churn. The choice between them for deployment in a real-world scenario would depend on additional considerations, such as the specific business goals (e.g., maximizing identified churners vs. minimizing false alarms), computational resources, and the need for model interpretability.

Findings

The study, focused on forecasting customer attrition in the telecom sector, revealed the following key findings:

i. Data Preparation is Crucial: The initial dataset of 7043 records required meticulous pre-processing, including addressing 11 missing 'TotalCharges' entries, along with standard steps like encoding, normalization, and feature engineering, to prepare it for model training.

ii. Both Models Show Promise: Both Logistic Regression and Random Forest models were found to be effective in predicting customer churn.

iii. Logistic Regression Slightly Outperforms Random Forest:

a. Accuracy: Logistic Regression achieved a higher overall accuracy (0.805676) compared to Random Forest (0.796672).

b. Recall: Logistic Regression also showed a better recall (0.559614 vs. 0.508010), meaning it was more successful at identifying actual churners.

c. F1-Score: The F1-Score for Logistic Regression (0.605009) was slightly better than Random Forest (0.570560), indicating a better balance between precision and recall.

iv. Implications: The superior performance of Logistic Regression, despite its simpler nature, suggests that the underlying relationships in this specific telecom churn dataset might be effectively captured by linear models, or that the extensive pre-processing made the data more conducive to such models.

CONCLUSION

Machine learning and big data analysis greatly simplify the telecom industry's churn forecasting procedure. This work combines big data with logistic regression and random forest to predict customer attrition for the telecom sector using machine learning on a big data platform. According to the survey's statistical results, 26.58% of consumers stopped using the telecom company's services, while 73.37% of consumers did not. The results of the survey also showed that monthly and total costs have a major impact on the telecom industry's churn rate. Additionally, the results of the study showed that the logistic regression classifier outperformed the random forest classifier in terms of accuracy.

Contributions to Knowledge

This study contributed to our understanding of the telecom sector and consumer attrition. By reviewing the ideas of telecommunications firm services, customer behavior, decision-making process, and how these concepts relate to one another, the study also made a conceptual contribution to the body of knowledge. The study's machine learning findings show how telecom companies perceive the risk variables that lead to customer attrition. As a result, this research has increased the current understanding. The study's findings demonstrate that a telecom company's revenue is significantly impacted by customer attrition, with monthly and total charges showing the strongest correlation with the churn rate.

REFERENCES

Prabadevi, B., Shalini, R., & Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, *4*.

Zhao, H., Yao, X., Liu, Z., & Yang, Q. (2021). Impact of pricing and product information on consumer buying behavior with customer satisfaction in a mediating role. *Frontiers in Psychology*, *12*(1). frontiersin.

Srinivasan, R., Rajeswari, D., & Elangovan, G. (2023, January 1). *Customer Churn Prediction Using Machine Learning Approaches*. IEEE Xplore.

Karamollaoğlu, H., Yücedağ, İ., & Doğru, İ. A. (2021, September 1). *Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis*. IEEE Xplore.

Shields, K. (2021). Chapter 3: Managing a Customer Service Team. *Ecampusontario.pressbooks.pub*, *3*(5).

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2023). Customer Churn Prediction in Telecom Sector using Machine Learning Techniques. *Results in Control and Optimization*, *14*, 100342.

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1).

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. (2022). A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. *PLOS ONE*, *17*(12)

Ribeiro, H., Barbosa, B., Moreira, A. C., & Rodrigues, R. G. (2023). Determinants of churn in telecommunication services: a systematic literature review. *Management Review Quarterly*.

Petropoulos, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3). sciencedirect.

Mandal, P. C. (2023). Engaging Customers and Managing Customer Relationships. *Journal of Business Ecosystems*, *4*(1), 1–14.

Shobana, J., Gangadhar, Ch., Arora, R. K., Renjith, P. N., Bamini, J., & Chincholkar, Y. devidas. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, *27*

Aghaabbasi, M., & Chalermpong, S. (2023). Machine learning techniques for evaluating the nonlinear link between built-environment characteristics and travel behaviors: A systematic review. *Travel Behaviour and Society*, *33*.

Saghir, M., Bibi, Z., Bashir, S., & Khan, F. H. (2019). Churn Prediction using Neural Network based Individual and Ensemble Models. *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 1–21. Springer. https://doi.org/10.1007/s42979-021-00592-x

Lu, Y. L., Zhelavskaya, I. S., & Wang, C. (2022). Neural Decision Tree: A New Tool for Building Forecast Models for Plasmasphere Dynamics. *Earth and Space Science*, *9*(7).

Zhao, M., Zeng, Q., Chang, M., Tong, Q., & Su, J. (2021). A Prediction Model of Customer Churn considering Customer Value: An Empirical Research of Telecom Industry in China. *Discrete Dynamics in Nature and Society*, *2021*, 1–12. hindawi. https://doi.org/10.1155/2021/7160527

Nguyen, N. Y., Tran, L. V., & Dao, S. V. T. (2022). Churn prediction in telecommunication industry using kernel Support Vector Machines. *PLOS ONE*, *17*(5).

Wibawa, A. P., Kurniawan, A. C., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. *International Journal of Recent Contributions from Engineering, Science & IT (IJES)*, *7*(2), 91.

Dar, S. A. (2022). The Relevance of Taylor's Scientific Management in the Modern Era. *Journal of Psychology and Political Science (JPPS) ISSN 2799-1024*, *2*(06), 1–6.

Michele, P., Fallucchi, F., & De Luca, E. W. (2019). Create Dashboards and Data Story with the Data & Analytics Frameworks. *Metadata and Semantic Research*, 272–283.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data*, *6*(1), 1–25.

Ku-Mahamud, K. R. & Basim Alwan, H. (2020). Big data: definition, characteristics, life cycle, applications, and challenges. *IOP Conference Series: Materials Science and Engineering*, *769*, 012007. https://doi.org/10.1088/1757-899x/769/1/012007

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE*, *15*(2),

Islam, Md. A. (2024). Impact of Big Data Analytics on Digital Marketing: Academic Review. *Journal of Electrical Systems*, *20*(5s), 786–820. https://doi.org/10.52783/jes.2327

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: Management, analysis and future prospects. *Journal of Big Data*, *6*(1), 1–25. springer.

Batko, K., & Ślęzak, A. (2022). The Use of Big Data Analytics in Healthcare. *Journal of Big Data*, *9*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8733917/

Roberts, S. (2023). *What are the 3V's of Big Data?* Www.theknowledgeacademy.com. https://www.theknowledgeacademy.com/blog/big-data-3v/

Jabeen, H. (2020). Chapter 3 Big Data Outlook, Tools, and Architectures. *Lecture Notes in Computer Science*, 35–55. https://doi.org/10.1007/978-3-030-53199-7_3

Berisha, B., Meziu, E., & Shabani, I. (2022). Big data analytics in Cloud computing: an overview. *Journal of Cloud Computing*, *11*(1).

Cavlak, N., & Cop, R. (2021). The Role of Big Data in Digital Marketing. *Advances in Marketing, Customer Relationship Management, and E-Services*, 16–33. https://doi.org/10.4018/978-1-7998-8003-5.ch002

Udeh, C. A., Orieno, O. H., Daraojimba, O. D., Ndubuisi, N. L., & Oriekhoe, O. I. (2024). Big Data Analytics: A Review of Its Transformative Role in Modern Business Intelligence. *Computer Science & IT Research Journal*, *5*(1), 219–236.

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3).

Qin, Z., Liu, Y., & Zhang, T. (2022). Research on Early Warning of Customer Churn Based on Random Forest. *Journal on Artificial Intelligence*, *4*(3), 143–154.

Sultan, A., Sałabun, W., Faizi, S., & Ismail, M. (2021). Hesitant Fuzzy Linear Regression Model for Decision Making. *Symmetry*, *13*(10), 1846.

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. *Procedia Computer Science*, *167*, 101–112.

Mooney, M. A., Neighbor, C., Karalunas, S., Dieckmann, N. F., Nikolas, M., Nousen, E., Tipsord, J., Song, X., & Nigg, J. T. (2022). Prediction of Attention-Deficit/Hyperactivity Disorder Diagnosis Using Brief, Low-Cost Clinical Measures: A Competitive Model Evaluation. *Clinical Psychological Science*, 216770262211202. https://doi.org/10.1177/21677026221120236

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, *281*(3), 588–596.

Sleiman, R., Mazyad, A., Hamad, M., Tran, K.-P., & Thomassey, S. (2022). Forecasting Sales Profiles of Products in an Exceptional Context: COVID-19 Pandemic. *International Journal of Computational Intelligence Systems*, *15*(1), 99.

Idriss, I. A., Cheng, W., & Hailu, Y. (2023). Weighted Maximum Likelihood Technique for Logistic Regression. *Open Journal of Statistics*, *13*(06), 803–821.