# Customer Churn Prediction in Telecom Industry through Applied Machine Learning Approaches

MSc Research Project
Research in Computing

Ananya Kachawa
Student ID: X21136751

School of Computing
National College of Ireland

Supervisor: Mr. Taimur Hafees

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student** Ananya Kachawa **Name:**

**Student ID:** x21136751

**Programme:** Master of Science in Data Analytics Information **Year:** 2023-24

**Module:** Research in Computing (Final Project)

**Supervisor:** Mr. Taimur Hafees

**Submission Due Date:** 14/12/2023

**Project Title:** Customer Churn Prediction in Telecom Industry through Applied Machine Learning Approaches

**Word Count:** 6425 **Page Count:** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Ananya Kachawa

**Date:** 14/12/2023

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Customer Churn Prediction in Telecom Industry through Applied Machine Learning Approaches

Ananya Kachawa
X21136751

## Abstract

The objective of this research is to develop and assess predictive models for customer attrition in the telecommunications industry through the utilization of sophisticated data analysis and machine learning techniques. The objective is to recognize key components that impact customers and give recommendations to assist mobile companies in progressing their approach to customer retention. Exploratory data analysis was conducted on a variety of data sets to understand the characteristics and behaviours of mobile phone customers. An assessment was conducted to decide the viability of machine learning models such as logistic regression, random forest, and gradient boosting in predicting customer churn.
The study also sought to examine the basic components that impact customer inclinations and behaviours. A comprehensive performance assessment of the diverse models was performed, centring on F1 score, accuracy, accuracy, recall, and fatigue expectation. The discoveries give practical counsel for telcos to diminish customer disarray and construct client loyalty.
***Keywords***: Customer Churn, Predictive Modelling, Customer Retention, Machine Learning, Telecommunication, Random Forest, Gradient Boosting, Logistic Regression.

# 1 Introduction

In today's highly competitive business environment, recognizing and foreseeing customer errors may be a major jump for organizations operating in different areas. Client churn has a quick and noteworthy impact on an organization's income and development. The subject of this paper is “Customer Churn Prediction”. The aim is to utilize progressed data science strategies to address these common business issues. The utilization of predictive models crosses industry boundaries and incorporates a comprehensive set of distinctive segments. To analyse client data and anticipate turnover, studios utilize several machine learning algorithms, such as logistic regression, random forest, and gradient boosting. To distinguish the finest approach to foresee steady loss, these models were compared utilizing exactness, exactness, review, and F1 score as measures. This investigation recognizes itself by completely looking at customer behaviour designs and recognizing key affecting variables, in this manner giving valuable experiences to businesses.

## 1.1 Background

The concept of customer conversion has gotten to be exceptionally vital within the cutting-edge business environment. As the industry gets to be more competitive and markets become more competitive, the company depends intensely on holding its existing clients. Abandonment has negative money-related results, including decreased sales, expanded promoting costs to obtain clients, and lost clients. Propels in machine learning and data analytics have made better approaches to get it and foresee customer behaviour (Labhsetwar, 2020). This considers the employment of these innovative advancements to look at client churn. By coordinating heterogeneous datasets and actualizing numerous predictive models, this study points to revealing key designs and choices that impact shoppers. Choosing to discontinue a service or product. As a result of this setting, it is conceivable to examine churn predictions, which are scholastically critical within the data-driven business world of the 21st century.

## 1.2 Research Aim and Objectives

The primary objective of this study is to create and assess a predictive model of client diversifying within the mobile sector utilizing advanced data analysis and machine learning strategies. The point of this study is to distinguish the key components that impact client churn and give down-to-earth advice to broadcast communications companies to move forward with their approach to client retention.

**Research Objectives**

- To conduct comprehensive exploratory data analysis on multiple datasets to understand the patterns and characteristics of telecom customers, and to prepare the data for modelling through cleaning and feature engineering.
- To develop various machine learning models, including Logistic Regression, Random Forest, and Gradient Boosting, and evaluate their performance in predicting customer churn.
- To identify and analyse the most significant features influencing customer churn, providing insights into customer behaviour and preferences.
- To compare the performance of different models in terms of accuracy, precision, recall, and F1-score, determining the most effective approach for churn prediction.
- To derive actionable insights from the model results and suggest strategies for telecom companies to reduce churn rates, enhance customer satisfaction, and increase customer loyalty.

**Research Questions**

- What are the key factors contributing to customer churn in the telecommunications industry?
- How effectively can machine learning models predict customer churn, and which model performs best in this context?
- What insights can be derived from the data regarding customer preferences and behaviours that lead to churn?
- How can the findings be translated into actionable strategies for telecom companies to reduce churn rates?

**Research Novelty**

- Incorporation of advanced machine learning techniques, such as ensemble methods and deep learning, to improve prediction accuracy.
- A comparative analysis of models across different datasets to understand the generalizability of the models.

## 1.3 Research Scope

In an effort to develop a model that can be applied to a broad spectrum of industries, this research endeavours to develop a generalizable customer attrition prediction system. The process entails the examination of diverse datasets in order to encompass a wide range of consumer behaviours and patterns. The study centres on the implementation and evaluation of various machine learning algorithms in order to ascertain the most effective and precise predictive model.

## 1.4 Significance of the Research

The potential transformative effect of this study on business approaches to client retention methodologies is critical. With a more total understanding of what contributes to client churn, organizations can take steps to address these issues, in this manner expanding client satisfaction and loyalty. The information and system coming about from this research will be a profitable asset for policy creators, permitting them to define the foremost fitting and compelling maintenance techniques (Tékouabou et al., 2022). In expansion, this consideration contributes to investigation by presenting new methods and discoveries within the areas of consumer analytics and predictive modelling. These commitments have vital suggestions for future inquiries about and usage in diverse business settings.

# 2 Related Work

## 2.1 "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform"

A research paper titled "Customer Churn Expectation" within the telecommunications industry utilizing machine learning on a huge database" has been published. Much obliged to its interesting application of huge information and machine learning procedures, it is picking up attention within the field of client churn forecast, particularly within the mobile sector. This article is consistent with the current drift of utilizing advanced analytics to anticipate customer behaviour, a subject that's still prevalent in modern communication investigate (Ahmad, Jafar, and Aljoumaa, 2019). Incorporating viewpoints of social network analysis (SNA) into churn forecast models represents an imaginative strategy and represents a developing mindfulness in academic research of the significance of social elements within the examination of consumer behaviour. The premise of this inquiry is the utilization of machine learning algorithms to analyse a huge set of data given by the telecommunications company SyriaTel. This aspect of the investigation is steady with a progressing scholarly discussion that emphasizes the significance of comprehensive and different information sets to progress the precision of predictive models.

The inquiry utilizes different tree-based calculations such as decision trees, random forests, GBM, and xgboost. These calculations contribute altogether to the body of knowledge on the adequacy of these approaches in predicting client attrition. Most imperatively, the high execution of the XGBOOST calculation, which comes to an amazing AUC value of 93.3%, sets an unused benchmark within the field of attrition forecast. By integrating technical and methodological choices and utilizing big data stages, the investigate will touch on the most recent improvements in data science and analysis.

The consideration of SNA highlights, which greatly improved the show and execution, reflects the current scholarly approach to consolidating social and common behavioural components into predictive examinations. The critical thing is that the paper's procedure for understanding the information vulnerability issue, a methodology that does not include tree algorithms or sparse analysis, bargains with this problem. This approach is reliable with current investigative approaches that point to addressing big information boundaries within the field of machine learning applications. This writing survey contributes significantly to the zone of customer determining within the broadcast communications industry. By coordinating intelligent machine learning strategies, huge data analytics, and SNA capabilities, this unused system makes a comprehensive and successful show for anticipating client errors. The study's strategies and come about will make a noteworthy commitment to a quickly developing field and set a standard for future investigation.

## 2.2 "Predictive Framework for Advanced Customer Churn Prediction using Machine Learning"

An academic paper composed by Jena, Bisoyi, and Tripathy (2020) gives a comprehensive investigation of the application of machine learning algorithms to anticipate attrition within the mobile sector. An imperative finding of the study is the assessment of decision trees and random forest algorithms in comparison. Comparing the two models, it was found that the Random Forest demonstration performed way better on a few performance indicators (accuracy, precision, recall, F-1 score, and phi-coefficient). In specific, higher gains (27.20%), precision (6.30%), and recall (4.85%) than utilizing the Random Forest show. To assess these models, a comprehensive assessment was performed utilizing different measurements and procedures such as confusion matrix and ROC examination. The confusion matrix gives a comprehensive summary of the extent of true positives, false positives, true negatives, and false negatives, and the predictive capabilities of each show can moreover be assessed. To perform a comprehensive assessment of the model's performance, the region under the convex hull (AUCH) was calculated.

The random forest model showed a higher AUCH, illustrating a stronger classification performance. A critical region of research is the creation of client interfacing that progresses client churn forecasting. This hands-on utilization outlines the practical regard of utilizing various characteristics to better predict individual turnover. The interface goes past a theoretical study and serves as verification to meet noteworthy examinations and hone needs within the industry. This paper gives basic enlargements to the extent of steady loss desires in broadcast communications, especially concerning the execution and comparison of machine learning procedures in a mechanical setting. The revelations allow basic bits of information for broadcast communications companies looking to execute data-driven methods to hold clients and dodge churn.

## 2.3 "Developing a Customer Leak Detection Model Using Machine Learning Techniques"

Calatayud Coquillat (2020) conducted a comprehensive analysis of machine learning (ML) in the communications industry for customer attrition forecasting. This study is notable for its comprehensive methodology that examines the effectiveness of different machine learning techniques in identifying consumer misinformation. The paper begins by explaining the concepts of machine learning (ML) and the classification of supervised and unsupervised learning. This highlights the importance of proper data preparation, with special attention to handling binary or linear variables and discarding incomplete observations. This concept is important in laying the foundation for machine learning modelling.

The main strength of the study is its extensive exploratory data analysis. Research uses graphical visualization tools to increase understanding of data and patterns in consumer behaviour. This methodological approach not only facilitates the interpretation of the results of ML models but also places them in the proper context, increasing the relevance and applicability of the research. The main focus of this study is to compare and implement various algorithms such as Naïve Bayes Classifier, Decision Tree, Bagging, Random Forest, Xgboost, and Support Vector Machine (SVM) in radial and linear regression. The research evaluates these models by examining performance metrics, including area under the curve (AUC), sensitivity, specificity, and accuracy. Benchmarking is of great value as it shows the strengths and weaknesses of all algorithms when predicting customer misunderstanding. This study presents an innovative method for examining the economic consequences of turnover forecasts. The model is evaluated in a cost-sensitive manner that outperforms traditional equity metrics by considering the financial impact of inaccurate forecasts. Most importantly, this element ensures that machine learning models are aligned with real business scenarios, showing the financial impact on customers. In this article, some solutions to the problem of data uncertainty in regression prediction models were explored. Techniques such as SMOTE (Synthetic Minority Oversampling Technique) are used to generate balanced synthetic data. A major concern with attrition forecasts is that they distort the majority of the class. This method is important to reduce this problem. The use of SMOTE will improve research methods and provide a set of examples to guide training and assessment.

The study concludes with an exhaustive evaluation of the ML models, encompassing an assessment of their economic feasibility and efficacy in a practical situation. The study's comprehensive methodology, which includes an in-depth examination of the data, the implementation of diverse machine learning techniques, and the integration of a cost sensitive assessment, establishes it as a noteworthy advancement in the discipline. This not only contributes to the comprehension of customer churn but also facilitates the development of more practical and cost-effective machine learning applications in the business sector.

## 2.4 "Performance Evaluation of Different Machine Learning Methods Applied on Churn Database"

A study conducted by Rodríguez Suarez in 2022 examines the efficacy of machine learning methods in the telecommunications industry with respect to predicting customer attrition. The principal objective of this study is to assess the efficacy of two distinct machine learning approaches—Strong Forest (RF) and Extreme Gradient Boosting (XGB)—particularly when applied to unbalanced data. By utilizing a dataset obtained from a telecommunications company, this paper underscores the criticality of selecting an appropriate data collection design and analysis method. The selected dataset, which is accessible via Kaggle, is dedicated to the prediction of customer behaviour with a specific emphasis on churn or customer attrition. Through an investigation of the two machine learning methods, the article conducts an exhaustive analysis of the operation of RF and XGB. RF generates and integrates a large number of low-performance models (Decision Trees) to improve overall performance. XGB is an iterative evolutionary model that consistently enhances itself in response to previous iterations. Both methodologies are renowned for their resilience when confronted with unbalanced datasets. A variety of performance indicators—including Sensitivity, Specificity, Precision, F1 Score, and Geometric Mean—are employed in the research to assess the efficacy of these methods. These indicators facilitate the identification of the most optimal attrition prediction model. A comprehensive analysis is conducted on the tendencies of these indicators across various parameter values of RF and XGB.

The enhancement of efficacy resulting from model customization is a noteworthy discovery of the research. The study consists of generating and evaluating a grid of RF and XGB parameters with an emphasis on the F1 Score. The optimal model was chosen from a pool of 77,000 models for XGB and 1200 models for RF as a result of this procedure. Performance is enhanced for RF when the parameters 'mtry' and 'ntree' are increased, according to the study. Performance, however, reaches its maximum at 'ntree' = 175, after which no additional enhancements are observed. It was determined that maximum depth values greater than three and minimum offspring weights of zero were optimal in XGB. The quantification of the performance enhancement induced by customization revealed that RF experienced a 3.51% improvement, whereas XGB witnessed a 6.21% improvement. XGB outperformed RF by 4.5%, attaining an F1 test score of 0.8850575, in contrast to the top-performing RF model which attained 0.8463146. XGB demonstrated superior accuracy, whereas RF demonstrated greater sensitivity.

Based on the research findings, it can be concluded that attrition prediction is a proficiency shared by both RF and XGB, with XGB marginally surpassing RF. On the contrary, RF has been observed to exhibit superior performance in mitigating False Negatives, a critical factor in situations where customer retention is of utmost importance. Additional client segmentation and retention strategies could be created utilizing other ML techniques, such as clustering, according to the paper. This research paper presents astute comparisons and pragmatic assessments of machine learning methodologies employed in the prediction of customer attrition. It effectively showcases the potential of these approaches to enhance business outcomes within the telecommunications industry.

## 2.5 "A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation"

The scholarly article authored by Zhang, Moro, and Ramos (2022) centres on the construction of an attrition prediction model that is efficacious within the telecommunications industry. The research study constructs an attrition prediction model by employing Fisher discriminant equations and logistic regression analysis on data obtained from three prominent Chinese telecommunications companies. The significance of this investigate is that it can be connected to the mobile media division, where it straightforwardly impacts consumer performance. Due to the strong competition and accessibility within the industry, companies are continuously trying to find successful ways to build client loyalty. This prerequisite is met by a consideration that proposes a model that precisely predicts the likelihood of client conversion at a rate of

93.94%. This demonstration is way better than the basic Fisher and equation, coming to 75% accuracy.

The unmistakable highlight of this study is its client segmentation strategy. This approach uses case analysis to understand the business attributes of versatile customers and creates attrition prediction models custom-fitted to these individual customer sections. This strategy of approach isn't as it were inventive but moreover in line with the industry drift towards targeted promoting procedures. The accuracy of the calculated regression model created in this study is exceptionally critical. The adequacy of the model lies in its capacity to predict sales in reaction to client complaints about service, giving telecommunication firms with data that can move forward their approach to client retention.

Research centres more on the viability of estimating client retention than on client securing. This shows the significance of churn forecast models to guide client retention programs. Utilizing discriminant investigation and Fisher logistic regression within the setting of mobile client churn, this paper fills a gap in the existing literature. Past considerations have centred on analysis and cluster analysis, recognizing this consideration for its imaginative methodological system. It moreover gives data directors the instruments to precisely anticipate client behaviour and optimize client maintenance procedures, resulting in cost savings and expanded benefits.

According to the research and suggestions, telecommunications companies ought to centre on diminishing settled and local costs and improving service quality over the whole extend of telecommunications products. These enhancements will have a positive impact on customer retention. This consideration makes a significant contribution to the zone of customer churn expectation within the versatile segment. In expansion to giving dependable, data-driven models that demonstrate forecast accuracy, the author advised telecommunication firms on vital client retention measures, utilizing advanced information analytics procedures to solve trade issues.

| Research Citation | ML Models Used | Findings with Numerical Values |
|---|---|---|
| Ahmad, Jafar, and Aljoumaa, 2019 | Decision Tree, Random Forest, GBM, XGBOOST | XGBOOST achieved an AUC value of 93.3% |
| Bisoyi, and Tripathy, 2020 | Decision Tree, Random Forest | Random Forest outperformed Decision Tree in accuracy (6.30% improvement), precision (27.20% improvement), and recall (4.85% improvement) |
| Calatayud Coquillat, 2020 | Naïve Bayes Classifier, Decision Tree, Bagging, Random Forest, Xgboost, SVM (linear and radial) | Various performance metrics<br>were evaluated, with higher |

| | | |
|---|---|---|
| | | accuracies. |
| Rodríguez Suarez, 2022 | Extreme Gradient Boosting (XGB), Random Forest (RF) | XGB outperformed RF with a<br>4.5% higher F1 Score (XGB: 0.8850575, RF: 0.8463146) |
| Zhang, Moro, and Ramos, 2022 | Fisher Discriminant Equations,<br>Logistic Regression | Logistic Regression model had a prediction accuracy of 93.94%, higher than Fisher's discriminant equations (75%) |

# 3 Research Methodology

The objective of this research is to analyse customer attrition across multiple telecommunication datasets to identify trends and factors that impact customer retention. The methodology is based on techniques that are fundamental to data science and include data processing, modelling, and analysis. Adopting this empirical methodology guarantees a rigorous and scientific procedure for comprehending the intricacies of customer attrition.

## 3.1 Data Collection

The datasets for this study were sourced from Kaggle, renowned for its extensive collection of datasets for diverse machine learning tasks. The datasets include:

1. **Telco Customer Churn Data:** Offering insights into customer profiles of a telecom company, including churn status.
2. **BigML Telecommunication Data:** Providing detailed information on customer usage and account statuses, along with churn details.
3. **Churn Train Data:** This dataset gives further insights into customer behaviours and their churn status.

Each dataset brings a unique angle to customer behaviours and churn, making them ideal for a comparative analysis framework.

## 3.2 Data Preprocessing

Data preprocessing was the foundational step:

- Loading Data: The Python library `pandas` was utilized for loading and handling the datasets.
- Initial Exploration: An initial examination of the datasets helped understand their structure, using functions like `.head()` for a preliminary overview.
- Missing Value Analysis: The presence of missing values was checked in each dataset using `isnull().sum()`. All datasets were found to be complete with no missing values.

- Duplicate Records Check: Each dataset was checked for duplicate entries to ensure data integrity.
- Data Standardization: The 'Churn' column across the datasets was standardized to maintain consistency, facilitating comparative analysis.
- Numerical Data Conversion: Certain columns, such as 'TotalCharges' in the Telco dataset, were converted into numeric types for appropriate quantitative analysis.

## 3.3 Exploratory Data Analysis (EDA)

A thorough EDA was conducted to identify patterns and gather insights:

**Churn Distribution Visualization**: Bar plots were used to display the churn distribution in each dataset.

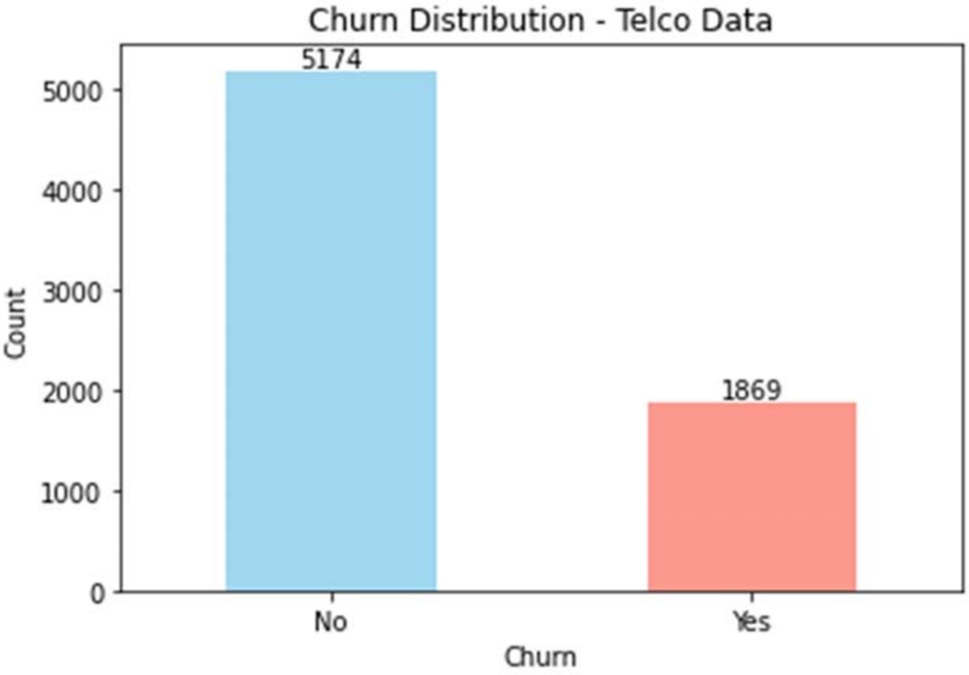


**Figure 1: Churn Distribution for Telco Data (Source: Self-Created)**

The customer churn count is displayed in the above visualization. In case of the Telco Data, the count of “Not Churn” customers are significantly higher than the “Churn” customers.

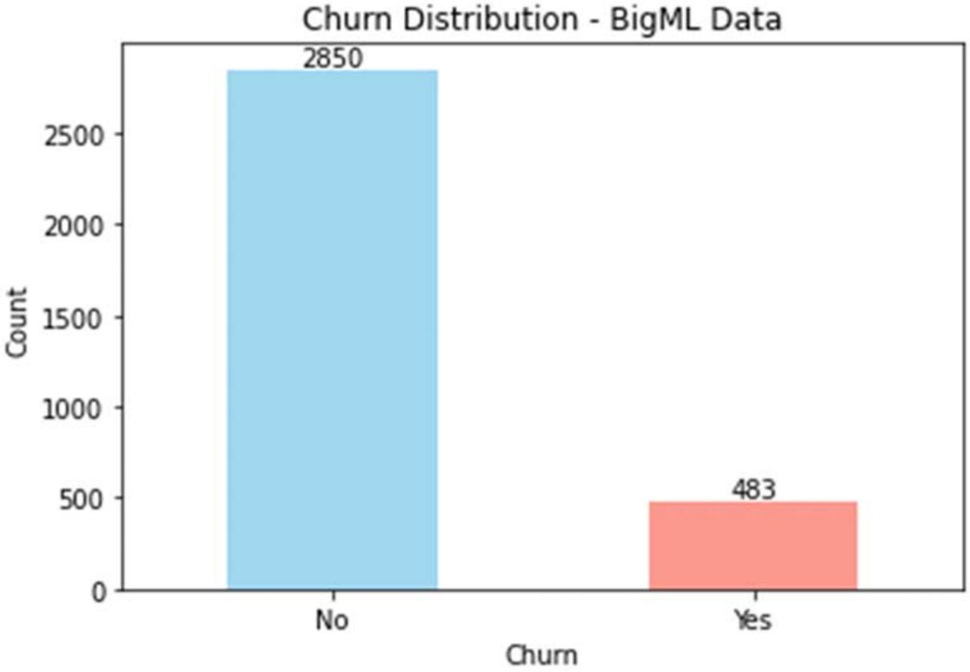


**Figure 2: Churn Count by BigML Data (Source: Self-Created)**

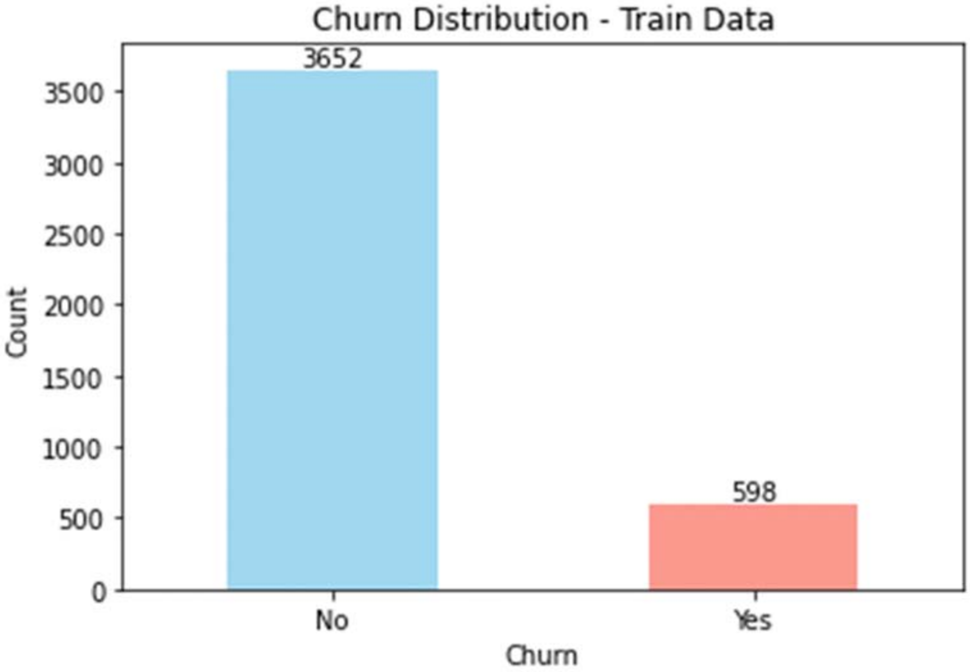


**Figure 3: Churn Distribution by Train Data (Source: Self-Created)**

By interpreting and analysing the distribution of the churn variables in each dataset, it can be said that the number of churn customers are significantly higher than the non-churn customers.

**Numerical Feature Analysis**: Histograms analysed the distribution of key numerical features like 'tenure', 'total day minutes', and 'total evening minutes'.

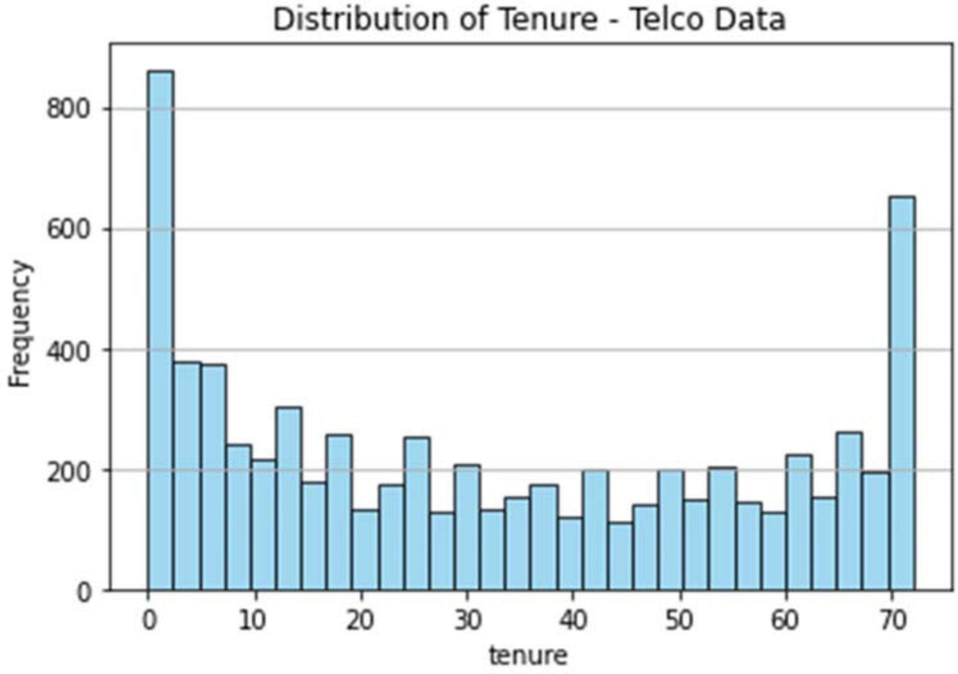


**Figure 4: Distribution of Tenure-Telco Data (Source: Self-Created)**

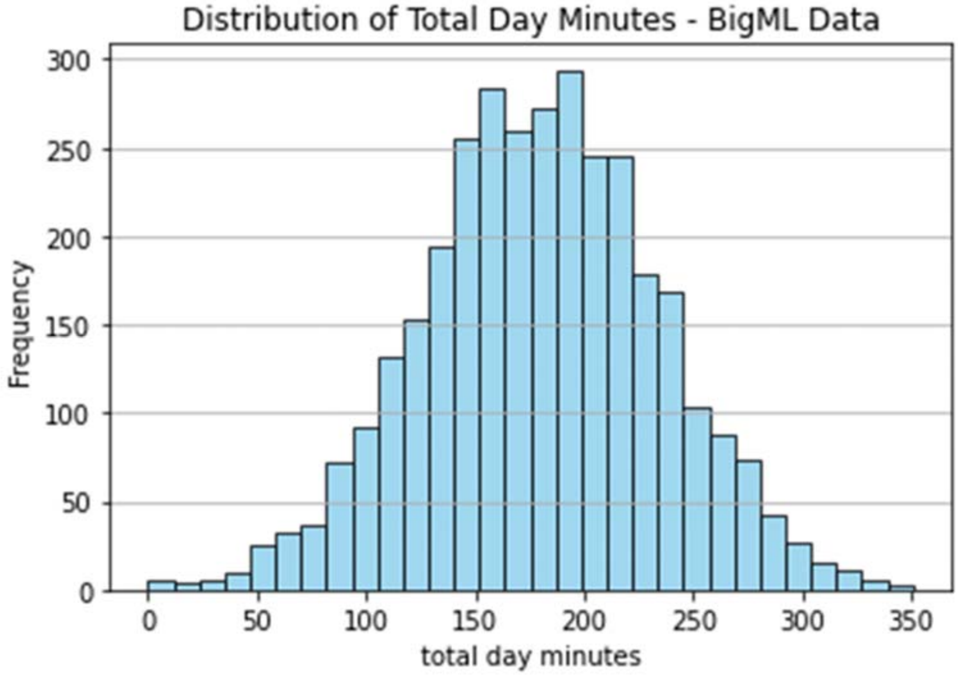


**Figure 5: Distribution of Total Day Minutes-BigML Data (Source: Self-Created)**

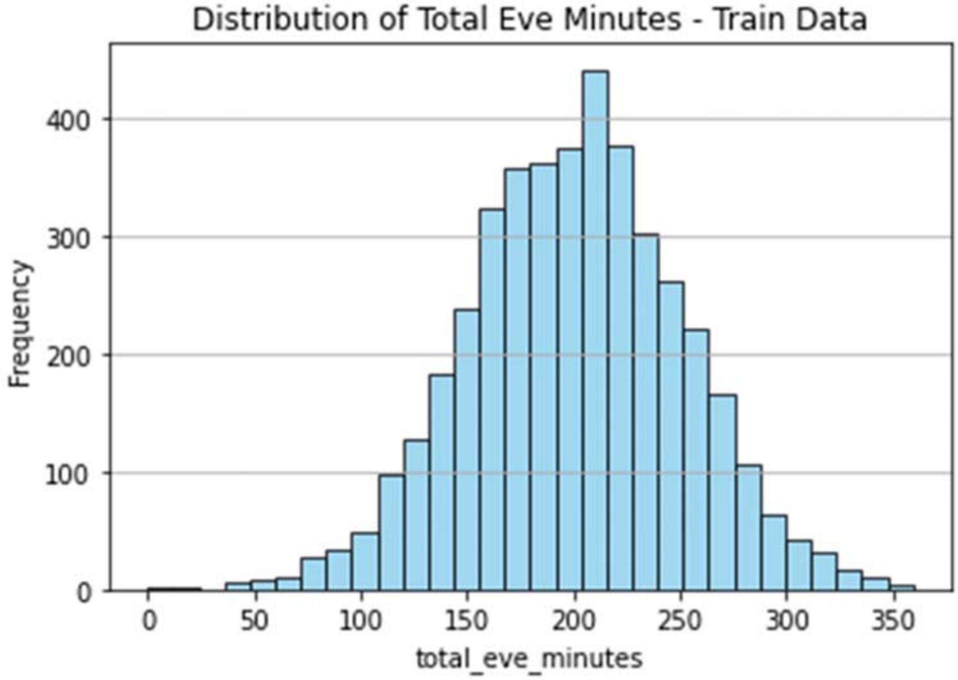


**Figure 6: Distribution of Total Eve Minutes (Source: Self-Created)**

**Categorical Data Analysis**: Churn rates against categorical variables like 'Contract' type, 'International Plan', and 'Voice Mail Plan' were examined using bar plots.

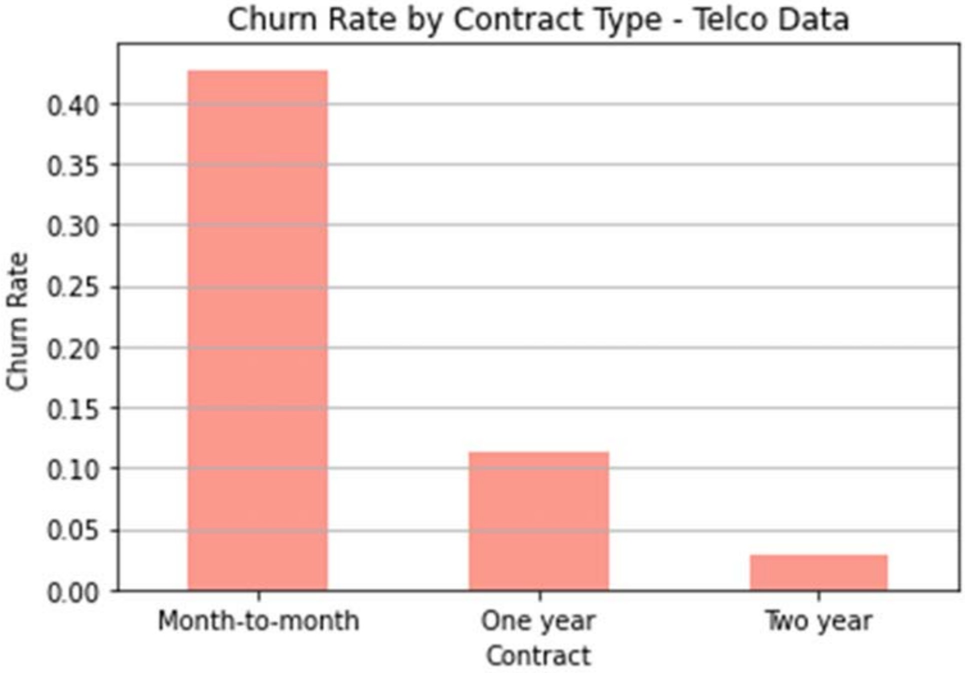


**Figure 7: Churn Rate by Contract Type-Telco Data (Source: Self-Created)**

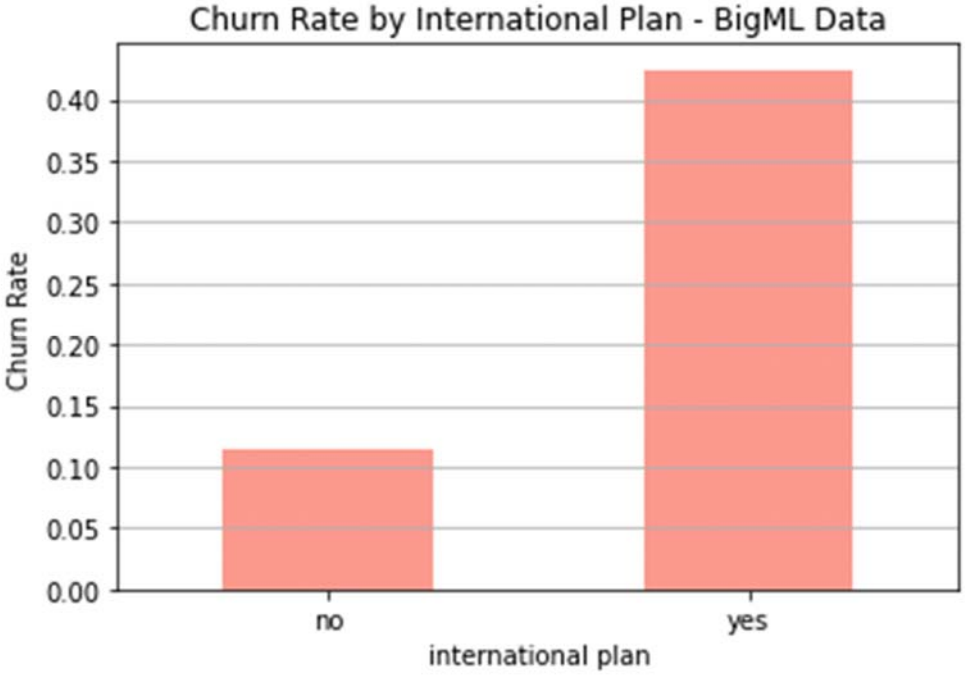


**Figure 8: Churn Rate by International Plan-BigML Data (Source: Self-Created)**

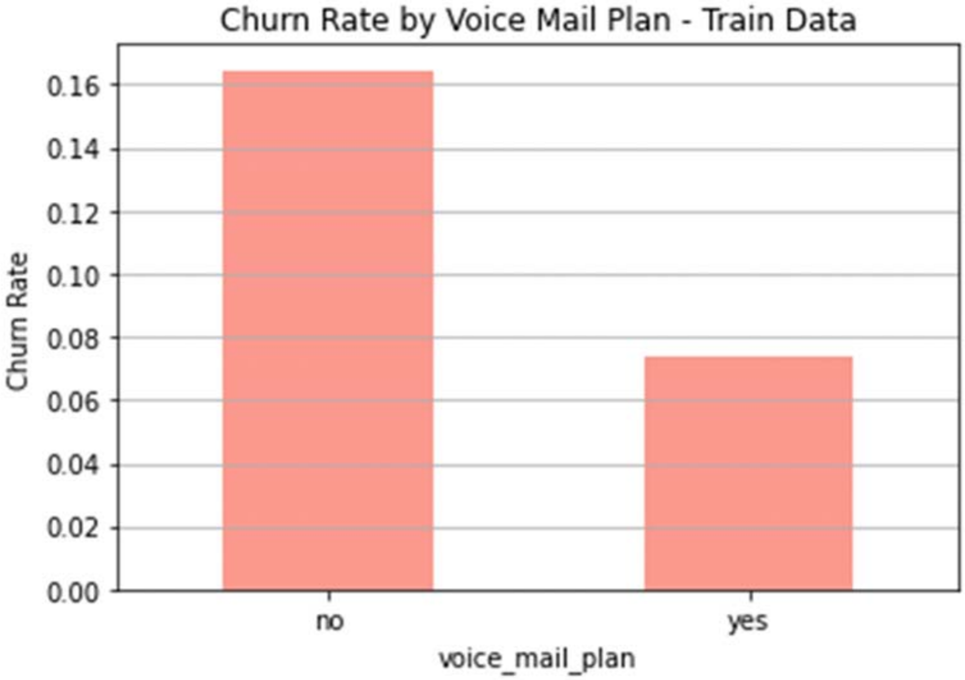


**Figure 9: Churn Rate by Voice Mail Plan-Train Data (Source: Self-Created)**

**Correlation Matrix**: A heatmap was generated for the Telco dataset to understand correlations among different features.

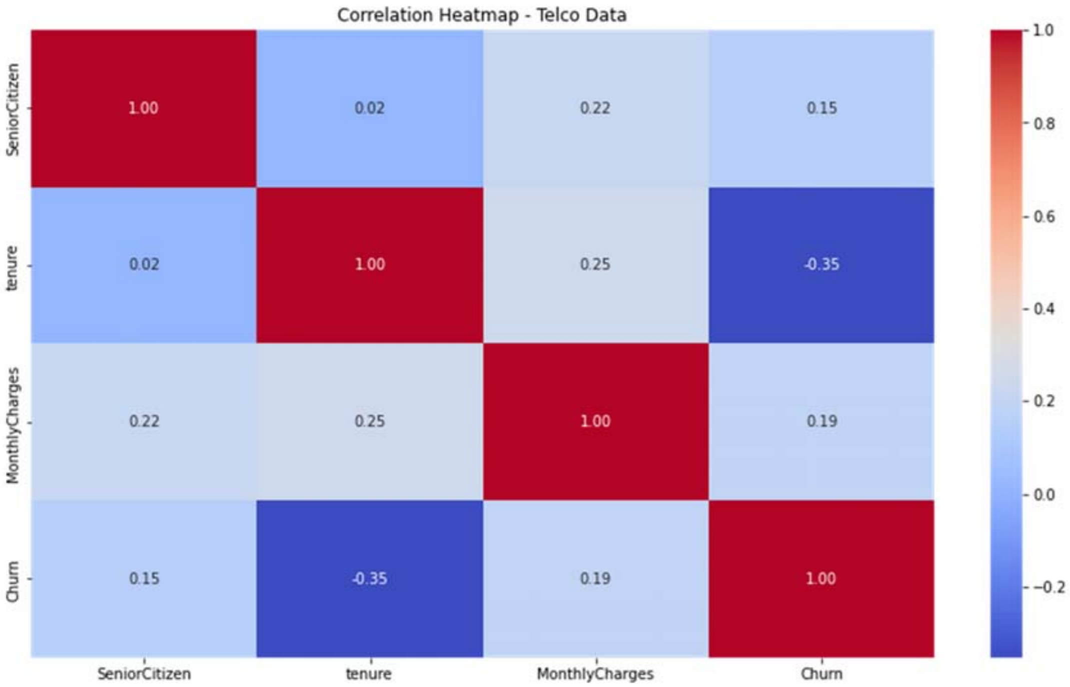


**Figure 10: Correlation Matrix (Source: Self-Created)**

As per the above correlation matrix, it can be said that the "Tenure" and "Monthly Charges" are two most corelated variables with the churn.

Additional Visualization Techniques: Box plots, count plots, and other visual tools were employed to delve deeper into specific feature analyses.

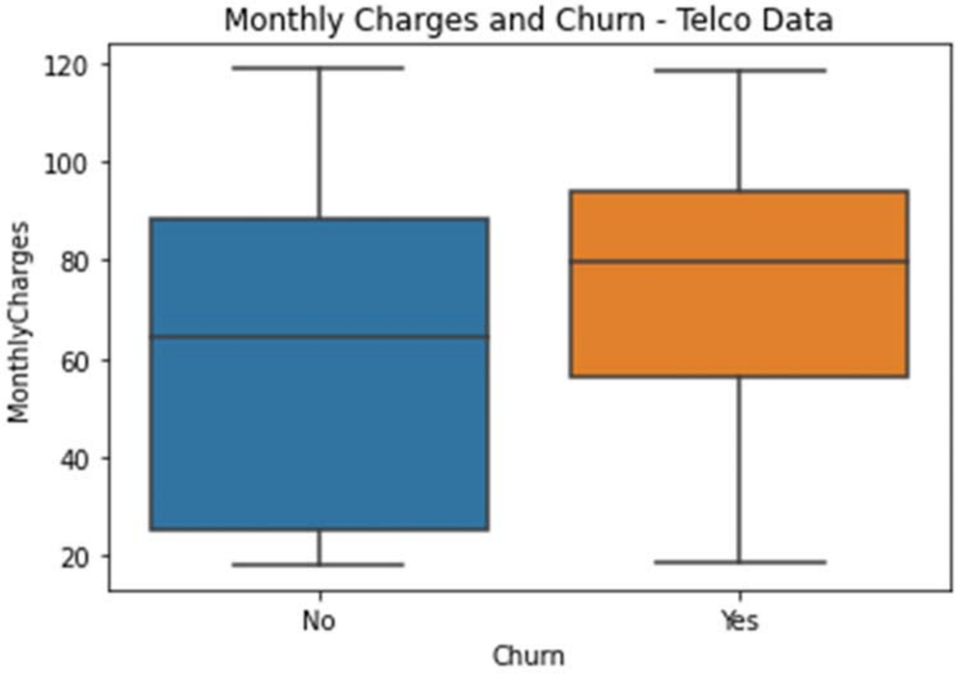


**Figure 11: Monthly Charges and Churn (Source: Self-Created)**

## 3.4 Data Transformation for Modelling

For model development, the datasets underwent transformation:

- Feature Engineering: Categorical and numerical columns were identified. Categorical columns were transformed via one-hot encoding, and numerical columns were standardized.
- Missing Value Handling: Missing values in 'TotalCharges' were imputed using mean values.
- Pipeline Development: A `ColumnTransformer` pipeline was established for systematic preprocessing.

## 3.5 Model Development and Evaluation

The processed Telco dataset was divided into training and testing sets, and three distinct machine learning models were applied:

1. *Logistic Regression*
2. *Random Forest Classifier*
3. *Gradient Boosting Classifier*

Each model's performance was assessed on the test set, focusing on precision, recall, f1score, and accuracy to evaluate the models' effectiveness in predicting customer churn.

## 3.6 Statistical Techniques

The study employed a range of statistical methods:

- Descriptive Statistics: To summarize the data and identify underlying patterns.
- Correlation Analysis: Employed to gauge the strength and direction of relationships between variables.

## 3.7 Ethical Considerations

The prioritization of ethical considerations, specifically data privacy, was observed. Anonymity was applied to the publicly accessible datasets to assure confidentiality and ethical data management.

## 3.8 Limitations and Recommendations for Future Research

The research, while insightful, faces certain limitations:

- Dataset Limitations: The datasets are limited in scope to particular geographic areas and temporal intervals.
- Model Selection: The preference for simpler models due to their interpretability potentially resulted in a reduction in predictive accuracy.

Subsequent avenues of inquiry may encompass the implementation of more intricate modelling methodologies, the incorporation of external data sources to augment the depth of understanding, and the investigation of time-dependent dynamics in customer attrition.

The current strategy presents an efficient and viable strategy to examine client misconceptions within the telecommunications industry (Dalli, 2022). This study contributes altogether to the understanding of the determinants of client churn through thorough sample assessment, exploratory analysis, and cautious data handling.

# 4 Design Specification

The framework for this examination is data visualization, which utilizes a combination of machine learning and statistical methods planned to explore consumer misconceptions within the telecommunications industry. The fundamental system comprises three primary parts: exploratory data analysis (EDA), predictive modelling, and data handling.

## 4.1 Data Preprocessing

In this step, the raw data is cleaned and converted into an organized way that is appropriate for examination. Major duties incorporate taking care of missing values, applying special coding to crude factors, and rectifying numerical conditions to plan for normalization (AlShourbaji et al., 2023). Python pipeline execution utilizing 'pandas' for data manipulation and 'ColumnTransformer" from sklearn for automated pre-processing. This guarantees consistency and reproducibility.

## 4.2 Exploratory Data Analysis (EDA)

This step uses a few statistical and visualization methods to uncover designs and perceptions within the information. Visual tools such as histograms, bar charts, and heat maps were utilized to understand feature distributions and relationships. For this matplotlib and seaborn Python libraries are emphatically backed.

## 4.3 Predictive Modelling

Three machine learning models were included within the design: logistic regression, random forest classifier, and gradient boosting classifier. The determination of these models is based on their capacity to perform binary classification tasks. sklearn library is utilized within the implementation to facilitate model training and evaluation (Chaudhary et al., 2022). Individual models are assessed employing a comprehensive set of measurements, including accuracy, recall, f1 score, and precision, to decide predictive performance. The general design is organized flexibly and extensively, permitting the integration of additional data sources or diverse modelling strategies at afterward stages of work. The plan prerequisites meet the criteria for conducting a comprehensive, repeatable, and insightful study of client billing within the telecommunications industry.

# 5 Implementation

An arrangement of systematic steps was utilized to run the client churn analytics program, which resulted in a careful study of churn patterns and the creation of predictive models. The outputs of this implementation are adjusted datasets, developed models, and visualizations (Jeyakarthic and Venkatesh, 2020). The primary programming language utilized was Python, which features a huge collection of data science libraries. In specific, to this extent, few libraries are utilized such as pandas, scikit-learn, matplotlib, ocean, etc.

## 5.1 Data Preprocessing

The preparatory phase was devoted to the analysis of the data set. This process includes cleaning and modifying data to optimize the format for analysis and modelling (Almufadi and Qamar, 2022). The main tasks of this step are:

- **Handling Missing Values:** Total Charges contains a missing value. The columns of the Telco data set were found and determined using the mean strategy.
- **Encoding Categorical Variables:** The process of one-hot encoding was utilized to convert categorical variables into a modelling-compatible format.
- **Scaling Numerical Features:** In order to establish a consistent measure, attributes such as 'tenure,''monthly charges,' and 'total charges' were standardized.

The Python library pandas was instrumental in data manipulation, while sklearn's ColumnTransformer was used for streamlined preprocessing.

## 5.2 Exploratory Data Analysis (EDA)

The EDA phase was crucial for gaining insights into the datasets and understanding underlying patterns. This stage produced a variety of outputs, including:

- **Distribution Visualizations:** Histograms and bar plots were created to analyse the distribution of various features and the churn rate.
- **Correlation Analysis:** Heatmaps were utilized to understand the relationship between different variables.
- **Feature-Specific Analysis:** Box plots and count plots provided deeper insights into specific attributes.

The libraries matplotlib and seaborn played a vital role in generating these visualizations, offering an extensive range of plotting options.

## 5.3 Predictive Modelling

The culmination of the implementation was the development of predictive models. Three models were chosen:

- **Logistic Regression:** A baseline model for binary classification tasks.
- **Random Forest Classifier:** An ensemble model known for its high accuracy and robustness against overfitting.
- **Gradient Boosting Classifier:** An advanced ensemble technique that combines weak learners to form a strong predictive model.

Before each model was trained, the Telco dataset was pre-processed. The procedure entailed dividing the dataset into distinct training and testing sets, followed by the application and evaluation of the models on the testing set (Faritha Banu et al., 2022). The assessment centred on key performance indicators such as accuracy, precision, recall, and the F1 score in order to thoroughly examine the models' capabilities.

Sklearn's modelling and metrics modules, which offer an all-encompassing collection of tools for model training, prediction, and evaluation, were utilized extensively during this phase.

## 5.4 Tools and Languages

The implementation placed significant reliance on Python as a result of its robust data science libraries and efficient operation. Essential tools and libraries comprised of:

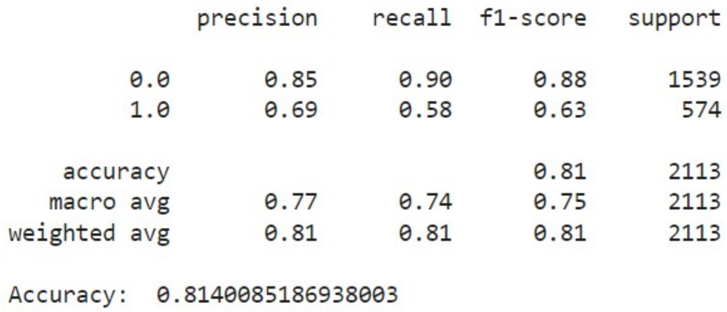
- **Python:** For overall programming and data manipulation.
- **Pandas:** For data loading, cleaning, and transformation.
- **Scikit-learn:** For implementing machine learning models, data preprocessing, and model evaluation.
- **Matplotlib and Seaborn:** For data visualization and generating insightful plots.

# 6 Evaluation

In this section, an in-depth analysis of the results of three machine learning models, logistic regression, gradient boosting classifier, and random forest classifier, were provided. In this study, the evaluation focuses on the predictive validity of these classifiers. In this study, the performance metrics like accuracy, precision, F-1 score, and recall were utilized. The implications of these findings from academic and viable viewpoints were reviewed.

## 6.1 Experiment 1: Logistic Regression Model

**Performance Metrics**



```
              precision    recall  f1-score   support

         0.0       0.85      0.90      0.88      1539
         1.0       0.69      0.58      0.63       574

    accuracy                           0.81      2113
   macro avg       0.77      0.74      0.75      2113
weighted avg       0.81      0.81      0.81      2113

Accuracy:  0.8140085186938003
```

**Figure 12: Logistic Regression Model Performance (Source: Self-Created)**

**Analysis**

The Logistic Regression model demonstrated a satisfactory overall accuracy score of 0.81.

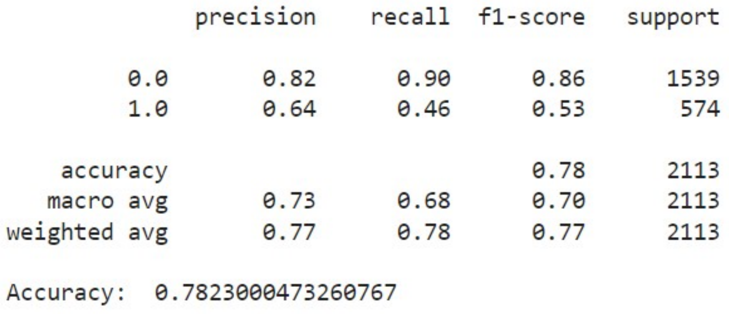
Nevertheless, its efficacy in forecasting the churn class (Class 1) was merely average (AlShourbaji et al., 2022). This is apparent from the precision and recall metrics, which suggest that although the model's churn predictions are reasonably dependable, it fails to identify a considerable quantity of churn instances (low recall).

**Implications**

This underscores the constraints of Logistic Regression when applied to datasets containing intricate patterns, as viewed through an academic lens. From a practical standpoint, this implies that although the model is valuable for baseline forecasting, placing exclusive reliance on it could result in overlooked prospects for identifying prospective attrition customers.

## 6.2 Experiment 2: Random Forest Classifier

**Performance Metrics**



```
              precision    recall  f1-score   support

         0.0       0.82      0.90      0.86      1539
         1.0       0.64      0.46      0.53       574

    accuracy                           0.78      2113
   macro avg       0.73      0.68      0.70      2113
weighted avg       0.77      0.78      0.77      2113

Accuracy:  0.7823000473260767
```

**Figure 13: Random Forest Model Performance (Source: Self-Created)**

**Analysis**

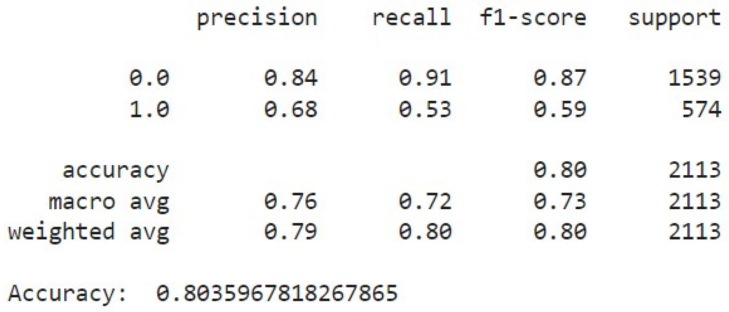
The reliability of the Random Forest model was inferior to that of the Logistic Regression model. The model exhibited moderate precision and recall in its churn prediction, suggesting difficulties in accurately identifying instances of churn (Singh et al., 2023). The reduced recall indicates that a significant number of genuine churn cases evaded detection.

**Implications**

This exemplifies the academic intricacy of decision trees within random forests when confronted with nuanced datasets. Practically speaking, this indicates that although Logistic Regression offers a more comprehensive analysis than Random Forest, the former's ability to predict attrition remains constrained, particularly with regard to recall.

## 6.3 Experiment 3: Gradient Boosting Classifier

**Performance Metrics**



```
              precision    recall  f1-score   support

         0.0       0.84      0.91      0.87      1539
         1.0       0.68      0.53      0.59       574

    accuracy                           0.80      2113
   macro avg       0.76      0.72      0.73      2113
weighted avg       0.79      0.80      0.80      2113

Accuracy:  0.8035967818267865
```

**Figure 14: GB Model Performance (Source: Self-Created)**

**Analysis**

The accuracy of the Gradient Boosting model was marginally inferior to that of Logistic Regression, but superior to that of Random Forest (Mirabdolbaghi & Amiri, 2022). The model exhibited a more equitable distribution of precision and recall in contrast to Random Forest, indicating a more proportionate methodology for churn case identification.

**Implications**

This underscores the efficacy of ensemble techniques such as Gradient Boosting when applied to intricate datasets in an academic setting. In practice, it proposes an enhanced model for attrition prediction that is more dependable; however, it also highlights the need for further refinement, specifically in the area of recall enhancement.

## 6.4 Comparative Analysis and Visual Representation

A more careful examination of the precision and recall metrics reveals the subtle variations in model performance, despite the fact that the aforementioned results indicate Logistic Regression to be marginally superior in terms of overall performance.

**Overall Implications:**

- **Academic Perspective:** The findings enhance comprehension regarding model selection in the field of predictive analytics, specifically in the context of forecasting customer attrition. This highlights the significance of assessing various performance metrics in addition to accuracy.
- **Practical Perspective:** These insights aid practitioners in the process of selecting the most suitable model in accordance with the business objective, which may be the reduction of false positives (recall) or false negatives (precision).

## 6.5 Discussion

Centring on gradient boosting, random forest, and logistic regression models, the results of this study give insightful comparisons with past scholarly research. Despite its complexity, the logistic regression showed a high accuracy of 81.4%. This is often reliable with the findings of Zhang, Moro, and Ramos (2022), who found that logistic regression outflanked distinctive models. In any case, the precision of this study is exceptionally low compared to the 93.94% precision recorded within the research, demonstrating ways to move forward aspects of feature engineering or model optimization.

In differentiation, the Random Forest model performed ineffectively in this consider, accomplishing 78.2% accuracy (Agasti and Satpathy, 2022). In differentiation, Bisoyi and Tripathy (2020) found that the random forest model appeared significant improvement over the decision tree model. The low results observed in this research may be due to the essential characteristics of the datasets utilized or the need to optimize the parameterization.

The study and results of the adjusted work of Gradient Boosting in terms of accuracy and review are reliable with the discoveries of Ahmad, Jafar, and Aljoumaa (2019), who found that a comparable strategy of Gradient Boosting called XGBOOST accomplished high performance. In any case, the AUC esteem of 93.3% accomplished by XGBOOST was not replicated within the display examination, recommending that advanced tests with gradient boosting parameters may be required to optimize the model.

### *6.5.1 Critical Analysis*

The study and design are exceptionally great, but it can be progressed by presenting more advanced synthesis strategies, such as xgboost. This is proposed by Ahmad, Jafar, and Aljoumaa (2019) and Rodríguez Suarez (2022).

Improvements in feature selection and engineering can optimize model performance. Current feature sets may not incorporate all the highlights required to make accurate forecasts. Algorithm optimization, particularly Random Forest and gradient boosting, may require assistance in examination to achieve results that coordinate the high levels of accuracy detailed in past studies.

### *6.5.2 Suggested Improvements*

Incorporating advanced optimization strategies such as XGBOOST can improve the predictive capabilities of the model.

By implementing state-of-the-art technology and the capacity to progress domain knowledge, models will incredibly progress their capacity to distinguish complex designs.

A stronger approach to parameter optimization, utilizing procedures such as random search or grid search, can improve model fit.

### *6.5.3 Contextualizing Findings*

Setting these results within the broader system of earlier research also uncovers the complexity and unpredictability of regression forecasts. It moreover shows the significance of nonstop testing including distinctive models and strategies to deal with particular clashes emerging from diverse information sets and trade situations.

# 7 Conclusion and Future Work

This study aims to answer key questions about the components that contribute to client churn in the mobile segment and to assess the performance of diverse machine learning models in anticipating churn. We effectively accomplished our objective by completely analysing the data and building models such as logistic regression, random forest, and gradient boosting. Most results show that even though logistic regression was utilized as a reliable basis, Gradient Boosting and Irregular Forest created more complex observations, but with drawbacks in terms of accuracy and recall.

The primary suggestion of the research for the telecommunications industry is to supply a system for recognizing and addressing the causes of customer disarray. In any case, there are limitations to the study, such as the depth of engineering and the breadth of the sample. Further investigation is needed to investigate outfit strategies and deep learning strategies to increase forecast accuracy. Further studies utilizing bigger datasets can be conducted to assess the model. Ability to generalize. Potential commercial executions incorporate the creation of predictive tools that empower telcos to recognize and oversee clients at risk of churn, in this manner expanding client maintenance methodologies and by and large proficiency. Additional fields of study that could prove fruitful include the investigation of temporal dimensions in customer behaviour and the incorporation of real-time data analysis.

# References

Agasti, B.R. and Satpathy, S., 2022. Hybrid ML Classification Approach for Customer Churn Prediction in Telecom Industry. Mathematical Statistician and Engineering Applications, 71(4), pp.10359-10368.

Ahmad, A.K., Jafar, A. and Aljoumaa, K., 2019. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), pp.1-24.

Almufadi, N. and Qamar, A.M., 2022. Deep Convolutional Neural Network Based Churn Prediction for Telecommunication Industry. Comput. Syst. Sci. Eng., 43(3), pp.1255-1270.

Al-Shourbaji, I., Helian, N., Sun, Y., Alshathri, S. and Abd Elaziz, M., 2022. Boosting ant colony optimization with reptile search algorithm for churn prediction. Mathematics, 10(7), p.1031.

AlShourbaji, I., Helian, N., Sun, Y., Hussien, A.G., Abualigah, L. and Elnaim, B., 2023. An efficient churn prediction model using gradient boosting machine and metaheuristic optimization. Scientific Reports, 13(1), p.14441.

Calatayud Coquillat, M., 2020. Developing a customer leak detection model using machine learning techniques (Doctoral dissertation, Universitat Politècnica de València).

Chaudhary, M., Gaur, L., Jhanjhi, N.Z., Masud, M. and Aljahdali, S., 2022. Envisaging Employee Churn Using MCDM and Machine Learning. Intelligent Automation & Soft Computing, 33(2).

Dalli, A., 2022. Impact of hyperparameters on Deep Learning model for customer churn prediction in telecommunication sector. Mathematical Problems in Engineering, 2022, pp.1-11.

Faritha Banu, J., Neelakandan, S., Geetha, B.T., Selvalakshmi, V., Umadevi, A. and Martinson, E.O., 2022. Artificial intelligence based customer churn prediction model for business markets. Computational Intelligence and Neuroscience, 2022.

Jena, D.K., Bisoyi, A. and Tripathy, A., 2020. Predictive Framework for Advanced Customer Churn Prediction using Machine Learning. International Journal of Computer Applications, 975, p.8887.

Jeyakarthic, M. and Venkatesh, S., 2020. An effective customer churn prediction model using adaptive gain with back propagation neural network in cloud computing environment. Journal of Research on the Lepidoptera, 51(1), pp.386-399.

Labhsetwar, S.R., 2020. Predictive analysis of customer churn in telecom industry using supervised learning. ICTACT Journal on Soft Computing, 10(2), pp.2054-2060.

Rodríguez Suarez, L.A., 2022. Performance evaluation of different machine learning methods applied on churn database (Master's thesis, Universitat Politècnica de Catalunya).

Sina Mirabdolbaghi, S.M. and Amiri, B., 2022. Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions. Discrete Dynamics in Nature and Society, 2022.

Singh, P.P., Anik, F.I., Senapati, R., Sinha, A., Sakib, N. and Hossain, E., 2023. Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. Data Science and Management.

Tékouabou, S.C., Gherghina, Ş.C., Toulni, H., Mata, P.N. and Martins, J.M., 2022. Towards explainable machine learning for bank churn prediction using data balancing and ensemblebased methods. Mathematics, 10(14), p.2379.

Zhang, T., Moro, S. and Ramos, R.F., 2022. A data-driven approach to improve customer churn prediction based on telecom customer segmentation. Future Internet, 14(3), p.94.