ACADEMIC CITY UNIVERSITY COLLEGE

END OF FIRST SEMESTER EXAMINATION – 2025/2026

D5203: STATISTICAL METHODS FOR DATA SCIENCE

Customer Churn Analysis in the Telecommunication Industry

Submitted By:
Jean-Luc Tonan Samuel Ahoyo

Roll Number:
2000250059

Lecturer:
Dr. Enoch Sakyi-Yeboah

Table of contents

Abstract

Customer churn represents a major challenge for telecommunications companies operating in highly competitive and subscription-based markets. The ability to accurately predict customer attrition is essential for revenue protection, customer lifetime value optimization, and strategic decision-making. This study applies statistical and machine learning classification techniques to predict customer churn using a telecommunications dataset comprising 7,032 observations and 30 encoded features.

The analysis begins with exploratory data analysis and data preprocessing, including handling missing values, encoding categorical variables, and assessing distributional properties of key numerical variables. Feature selection was performed using Mutual Information and Random Forest feature importance to identify the most influential predictors while preserving interpretability. Three supervised classification models were implemented: Logistic Regression, Random Forest, and Gradient Boosting. Model performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix analysis, and Precision–Recall AUC.

The results indicate that Logistic Regression achieved the highest overall accuracy (80.45%), while Gradient Boosting demonstrated superior recall (81.28%) and the highest Precision–Recall AUC (0.654), making it the most effective model for identifying customers at risk of churn. Given the asymmetric financial consequences associated with missed churners, recall emerged as a particularly important evaluation metric.

The findings highlight the importance of aligning predictive performance metrics with business objectives rather than relying solely on accuracy. By integrating statistical rigor, feature selection, and comprehensive model evaluation, this study provides a structured and interpretable predictive framework that supports data-driven customer retention strategies and informed decision-making within the telecommunications sector.

Introduction

In the contemporary digital economy, customer retention has become a critical determinant of organizational sustainability and profitability. Across industries such as telecommunications, banking, insurance, and fintech, firms increasingly operate in highly competitive markets where customers can easily switch providers. This phenomenon, commonly referred to as customer churn, represents the voluntary termination of a customer's relationship with a company. Churn directly impacts revenue stability, customer lifetime value, and long-term strategic growth. As markets become more saturated and customer acquisition costs rise, predicting and preventing churn has emerged as a central concern for businesses and policymakers alike.

Customer churn analysis gained prominence in the telecommunications industry during the 1990s, when deregulation and market liberalization intensified competition among service providers. Organizations quickly recognized that retaining existing customers was substantially less costly than acquiring new ones. Early analytical approaches relied primarily on descriptive statistics and classical econometric techniques, particularly logistic regression and survival analysis, to estimate the probability of customer attrition. These methods enabled firms to transition from reactive customer management strategies toward predictive and data-driven decision-making frameworks.

With advancements in computational capacity and the proliferation of large-scale customer datasets, churn modeling evolved to incorporate machine learning algorithms such as decision trees, support vector machines, and ensemble methods. Despite these developments, logistic regression remains widely used due to its probabilistic interpretation, theoretical grounding in maximum likelihood estimation, and interpretability of coefficients. Its statistical transparency makes it particularly suitable for structured evaluation, hypothesis testing, and policy-oriented analysis.

The rapid expansion of data science has further transformed churn prediction into a strategic financial instrument. Organizations now collect extensive behavioral, transactional, and demographic information that enables refined modeling of customer behavior. Statistical methods are no longer confined to estimating relationships; they are increasingly employed to optimize retention campaigns, personalize services, and guide strategic resource allocation. In subscription-based business models such as telecommunications services and digital platforms, even marginal reductions in churn rates can generate substantial improvements in profitability. Consequently, churn prediction has evolved from an operational tool into a strategic component of financial governance and competitive positioning.

Beyond corporate strategy, churn analytics provides valuable insights for regulators and policymakers. Telecommunications and fintech sectors represent foundational infrastructures in many economies, particularly in developing countries where mobile connectivity and digital financial services foster economic inclusion. Predictive modeling can assist regulators in monitoring market competitiveness, service quality, and consumer protection. Unusual churn patterns may indicate service deficiencies, pricing distortions, or structural

inefficiencies. Thus, statistical classification models contribute not only to firm-level optimization but also to evidence-based economic oversight.

Methodologically, churn prediction relies on supervised classification techniques designed to model a binary outcome (churn versus non-churn) based on explanatory variables such as contract type, tenure, billing characteristics, and service usage behavior. Logistic regression estimates the probability of attrition through a logit link function, allowing analysts to quantify the marginal effect of predictors on churn likelihood. However, as datasets expand in dimensionality and complexity, effective feature selection becomes essential to enhance model stability, interpretability, and predictive performance.

Rather than employing linear dimensionality reduction methods such as Principal Component Analysis, this study adopts feature selection techniques grounded in statistical dependency and ensemble learning. Mutual Information (MI) is utilized to measure the non-linear dependence between individual predictors and the churn outcome, thereby identifying variables with the strongest informational contribution. Complementarily, Random Forest feature importance is applied to rank predictors based on their contribution to impurity reduction within ensemble tree structures. These approaches enable the reduction of irrelevant or redundant variables while preserving interpretability and predictive relevance. Unlike latent component methods, feature importance and MI retain the original variable structure, facilitating clearer managerial and policy interpretation.

Recent developments in model diagnostics and evaluation metrics further strengthen churn analysis. Performance indicators such as accuracy, precision, recall, F1-score, and confusion matrices allow comprehensive assessment of classification performance. In churn modeling, recall assumes particular importance, as failing to identify high-risk customers may result in irreversible revenue losses. Consequently, evaluation metrics must be aligned with strategic objectives rather than relying solely on overall accuracy.

Despite widespread industrial adoption of churn prediction models, there remains a need for structured academic frameworks that integrate statistical assumptions, feature selection methodologies, and rigorous evaluation procedures. Many applied studies prioritize predictive accuracy without adequately addressing model interpretability, statistical validation, or variable relevance. A systematic statistical approach enhances robustness, transparency, and policy applicability.

Accordingly, this study applies statistical classification methods to predict customer churn within the telecommunications sector. By integrating exploratory data analysis, dependency-based feature selection through Mutual Information, ensemble-based feature ranking via Random Forest importance, logistic regression modeling, and comprehensive performance evaluation metrics, this research illustrates the progression of statistical methodologies from traditional econometric analysis to modern predictive analytics. The study contributes to the broader discourse on the evolving role of statistical modeling in corporate strategy, data-driven governance, and economic decision-making.

2. Problem Statement

Customer churn, defined as the voluntary termination of a customer's relationship with a service provider, remains a major challenge in competitive industries such as telecommunications, banking, and subscription-based digital services. Increased competition, technological innovation, and low switching barriers have made customer retention more difficult. Since retaining existing customers is generally more cost-effective than acquiring new ones, organizations increasingly rely on predictive analytics to anticipate and prevent churn. As a result, churn prediction has become both a research priority and a strategic business objective.

To address this issue, researchers have widely applied statistical and machine learning techniques. Logistic regression has traditionally provided an interpretable probabilistic framework for modeling customer behavior. More recently, machine learning models such as Random Forest and Gradient Boosting have demonstrated strong predictive performance due to their ability to capture nonlinear relationships and complex interactions. These approaches enable firms to proactively identify at-risk customers and implement targeted retention strategies.

Despite these advancements, important gaps remain in the literature. Many studies emphasize predictive accuracy without adequately considering statistical assumptions, diagnostic evaluation, or model interpretability. Additionally, performance assessment is often limited to overall accuracy, neglecting metrics such as precision, recall, and F1-score that better reflect the asymmetric costs associated with missed churners and unnecessary retention efforts.

Furthermore, although customer datasets frequently contain redundant or highly correlated variables, feature selection strategies are not always systematically incorporated into churn modeling frameworks. While some studies use dimensionality reduction techniques that transform variables into latent components, such approaches may reduce interpretability. Alternative techniques such as Mutual Information, which measures dependency between predictors and churn, and Random Forest feature importance, which ranks variables based on their predictive contribution, provide more transparent methods for identifying the most relevant features while preserving original variable meaning.

Finally, although churn prediction offers clear business benefits, its broader implications for policymakers and regulators remain underexplored. In sectors such as telecommunications and fintech, churn patterns may reflect service quality, pricing practices, or market competitiveness. Therefore, a statistically rigorous and interpretable predictive framework is needed to support both organizational decision-making and evidence-based policy formulation.

This study addresses these gaps by developing and comparing logistic regression, Random Forest, and Gradient Boosting models for predicting customer churn. By integrating exploratory data analysis, feature selection using Mutual Information and Random Forest importance, and comprehensive evaluation metrics, the study aims to provide a balanced and interpretable predictive approach that supports both business strategy and policy insights.

3. Objectives

3.1. General Objective

The general objective of this study is to develop and compare statistical and machine learning classification models for predicting customer churn in the telecommunications industry, in order to enhance customer retention strategies and support data-driven business and policy decision-making.

3.2. Specific Objectives

To achieve the general objective, the study will pursue the following specific objectives:

1. To perform exploratory data analysis (EDA) to understand the distribution, patterns, and relationships among customer attributes associated with churn.

2. To conduct data cleaning and preprocessing, including handling missing values, encoding categorical variables, and preparing the dataset for modeling.

3. To assess relevant statistical properties of the dataset and perform necessary diagnostic tests where applicable.

4. To apply feature selection techniques, particularly Mutual Information (MI) and Random Forest feature importance, to identify the most relevant predictors, reduce redundancy, and improve model efficiency while preserving interpretability.

5. To develop and train a logistic regression model as a baseline statistical classification approach.

6. To implement advanced machine learning models, including Random Forest and Gradient Boosting, for churn prediction.

7. To evaluate and compare model performance using accuracy, confusion matrix, precision, recall, and F1-score.

8. To interpret the comparative results in terms of business implications, predictive reliability, and policymaking relevance within the telecommunications sector.

4. Significance of the Study

Customer churn poses a significant challenge to organizations operating in competitive and subscription-based markets, particularly within the telecommunications industry. Retaining existing customers is generally more cost-effective than acquiring new ones, and high churn rates can lead to revenue instability, reduced profitability, and weakened market position. Therefore, developing reliable predictive models for identifying customers at risk of leaving is both a strategic and financial necessity. This study contributes to addressing this challenge by applying and comparing statistical and machine learning classification techniques to improve churn prediction accuracy and decision-making effectiveness.

From a business perspective, the study provides a structured analytical framework that supports data-driven customer retention strategies. By identifying high-risk customers in advance, organizations can implement targeted interventions such as personalized offers, service improvements, or pricing adjustments. Furthermore, comparing classical statistical models such as logistic regression with advanced ensemble methods like Random Forest and Gradient Boosting allows decision-makers to balance predictive performance with interpretability. While advanced models may enhance predictive power, logistic regression offers transparency and ease of interpretation, which are essential for managerial accountability and strategic planning.

Academically, this research strengthens the integration of statistical rigor and machine learning in churn modeling. Many applied studies emphasize predictive performance without systematically examining feature relevance or evaluation metrics aligned with business objectives. By incorporating exploratory data analysis, diagnostic evaluation, feature selection techniques such as Mutual Information and Random Forest feature importance, and comprehensive performance measures including precision, recall, and F1-score, this study adopts a balanced and interpretable predictive framework.

The study is also significant for policymakers and regulators, particularly in sectors such as telecommunications and fintech that contribute substantially to economic development. Churn analytics can provide insights into customer behavior, service quality, and market competitiveness. Unusual churn patterns may signal pricing inefficiencies, service deficiencies, or competitive instability. Thus, statistically grounded classification models support evidence-based regulatory decisions and consumer protection strategies.

Overall, the rationale for this study lies in its contribution to improving predictive reliability, enhancing financial sustainability, and strengthening the connection between statistical methodology and practical decision-making. By demonstrating the comparative strengths of classical and modern classification models, this research underscores the growing importance of data-driven analytics in contemporary business and policy environments.

5. Literature Reviews

5.1. THEORETICAL REVIEW

Customer Churn Concept: Definition, Importance, and Financial Impact

Customer churn refers to the phenomenon whereby customers discontinue their relationship with a service provider and switch to alternative competitors. This issue is particularly significant in industries characterized by subscription-based services, such as telecommunications, banking, and digital platforms, where switching costs are relatively low. In the telecommunications sector, customer churn has been identified as a major source of revenue loss and operational instability for service providers (Kaur & Gupta, 2018). Increasing competition and similar service offerings further intensify the challenge of customer retention (Dasgupta, 2020).

The importance of churn analysis has grown as organizations recognize that retaining existing customers is generally more cost-effective than acquiring new ones. Studies in customer relationship management suggest that customer retention strategies significantly enhance profitability and long-term financial performance (Jones & Sasser, 2020). Consequently, many organizations have adopted customer-centric approaches that emphasize understanding customer behavior and predicting potential churn to support strategic decision-making (Nguyen, 2022).

From a financial perspective, customer churn directly affects revenue stability, customer lifetime value, and marketing expenditure. High churn rates often lead to increased customer acquisition costs and reduced profitability for service providers (Ascarza et al., 2018). To mitigate these risks, companies increasingly rely on predictive analytics and statistical modeling techniques to identify customers likely to leave and implement targeted retention strategies. Predictive models have proven effective in enabling proactive interventions that improve customer satisfaction, loyalty, and overall business sustainability (Óskarsdóttir et al., 2017).

Overall, customer churn analysis represents both a strategic business concern and an important application of statistical methods in modern commerce. By understanding churn behavior through predictive analytics, organizations and policymakers can make informed decisions that enhance financial stability, market competitiveness, and consumer welfare.

Statistical Classification Methods

Statistical classification methods play a fundamental role in predictive analytics by enabling the categorization of observations into predefined classes based on explanatory variables. These methods are widely applied in data science, finance, healthcare, and telecommunications to support decision-making processes where binary or multiclass

outcomes are involved. In the context of customer churn prediction, classification models are used to estimate the likelihood that a customer will discontinue a service, allowing organizations to implement proactive retention strategies.

Among classical statistical approaches, logistic regression remains one of the most widely used classification techniques. Logistic regression models the probability of a binary outcome by estimating the relationship between a dependent variable and a set of independent variables through a logistic function. Its interpretability, probabilistic foundation, and relatively low computational complexity make it particularly suitable for applications where understanding the influence of predictors is as important as predictive accuracy (Hosmer, Lemeshow, & Sturdivant, 2013). In churn prediction, logistic regression enables analysts to quantify how factors such as service usage, customer tenure, and billing characteristics influence the probability of customer attrition.

Despite its strengths, logistic regression assumes linear relationships between predictors and the log-odds of the outcome and may struggle to capture complex nonlinear patterns in large datasets. This limitation has contributed to the growing adoption of machine learning classification methods, including ensemble techniques such as Random Forest and Gradient Boosting, which can model nonlinear relationships and interactions more effectively (Breiman, 2001). These approaches extend classical statistical modeling by improving predictive performance while maintaining the core objective of classification.

Nevertheless, statistical classification methods remain essential because they provide transparency, interpretability, and theoretical grounding that are often required for policy formulation, business accountability, and regulatory decision-making. While advanced machine learning models may achieve higher predictive accuracy, classical statistical models such as logistic regression offer clearer insights into the relationships between variables and outcomes, making them valuable for both predictive analytics and explanatory analysis.

Overall, statistical classification methods form the theoretical foundation of churn prediction and broader predictive analytics. Their evolution from classical probabilistic models to more complex machine learning techniques reflects the broader development of data science, where balancing predictive performance, interpretability, and practical applicability remains a central concern.

Machine Learning Approaches

The increasing availability of large-scale customer data and advances in computational capabilities have accelerated the adoption of machine learning techniques in predictive analytics. Unlike classical statistical models that rely on predefined assumptions about data structure, machine learning approaches are designed to learn complex patterns directly from data. These techniques are particularly valuable in customer churn prediction, where behavioral, transactional, and demographic variables may exhibit nonlinear relationships and interactions. Among the most widely used machine learning classification models are Random Forest and Gradient Boosting, both of which belong to the family of ensemble learning methods.

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve classification accuracy and reduce overfitting. The method operates by generating numerous decision trees from random subsets of training data and predictor variables, a process known as bootstrap aggregation or bagging. Each tree produces a classification outcome, and the final prediction is typically obtained through majority voting across all trees (Breiman, 2001).

One of the key advantages of Random Forest lies in its ability to handle high-dimensional data and complex nonlinear relationships without requiring strict statistical assumptions. It is also relatively robust to noise and multicollinearity, making it suitable for customer churn datasets that often include numerous correlated variables. Furthermore, Random Forest provides measures of feature importance, which can offer insights into the factors influencing customer behavior. However, despite its predictive strength, the model may lack interpretability compared to simpler statistical approaches such as logistic regression.

Gradient Boosting

Gradient Boosting is another ensemble learning technique that builds predictive models sequentially, where each new model focuses on correcting the errors made by previous models. Unlike Random Forest, which constructs independent trees simultaneously, Gradient Boosting develops trees iteratively to minimize prediction error through an optimization process based on gradient descent principles (Friedman, 2001).

This sequential learning process allows Gradient Boosting to capture subtle patterns in data and achieve high predictive performance, particularly in complex classification tasks such as customer churn prediction. The method is effective in handling nonlinear relationships, interaction effects, and heterogeneous datasets. Variants such as XGBoost and LightGBM have further enhanced computational efficiency and predictive accuracy, making Gradient Boosting one of the most powerful machine learning techniques in predictive analytics.

However, Gradient Boosting models can be computationally intensive and may require careful parameter tuning to prevent overfitting. Additionally, their complexity can reduce interpretability, which may limit their applicability in contexts where transparency and explainability are essential for decision-making.

Comparison with Classical Statistical Models

Machine learning classification techniques such as Random Forest and Gradient Boosting complement classical statistical approaches by improving predictive performance in complex datasets. While logistic regression provides interpretability and theoretical transparency, ensemble machine learning models often achieve higher accuracy by capturing nonlinear relationships and variable interactions. Consequently, contemporary churn prediction studies increasingly adopt a hybrid perspective that combines statistical interpretability with machine

learning predictive strength. This comparative approach enables organizations and policymakers to balance accuracy, interpretability, and practical applicability in predictive decision-making.

Dimensionality Reduction and Feature Selection Techniques: Mutual Information and Random Forest Importance

In predictive analytics, datasets often contain a large number of variables describing customer behavior, demographics, and transactional characteristics. While rich datasets can enhance predictive modeling, high-dimensional data may introduce challenges such as multicollinearity, redundancy among variables, increased computational complexity, and reduced model interpretability. To address these issues, feature selection techniques are commonly employed to simplify datasets while preserving the most informative predictors. Unlike transformation-based dimensionality reduction methods, feature selection approaches retain the original variables, thereby improving both efficiency and interpretability.

One widely used technique in statistical learning is **Mutual Information (MI)**. Mutual Information is an information-theoretic measure that quantifies the dependency between a predictor variable and the target variable. It captures both linear and nonlinear relationships by measuring how much knowing one variable reduces uncertainty about another. In the context of customer churn prediction, MI helps identify which customer attributes—such as tenure, contract type, service usage, or billing patterns—carry the greatest informational relevance for predicting churn. By ranking features according to their dependency with the outcome variable, irrelevant or weak predictors can be excluded, thereby reducing dimensionality while maintaining predictive strength.

Another effective approach is **Random Forest feature importance**, derived from ensemble tree-based models. Random Forest evaluates the importance of each feature based on its contribution to reducing impurity (e.g., Gini impurity) across decision trees. Features that consistently improve classification splits receive higher importance scores. This method is particularly valuable because it captures nonlinear interactions and complex relationships among variables. In churn prediction datasets, where customer behavior patterns may not follow strictly linear relationships, Random Forest importance provides a robust mechanism for identifying influential predictors.

Feature selection using MI and Random Forest importance offers several advantages. By retaining original variables rather than transforming them into latent components, these techniques preserve interpretability, an essential requirement for managerial decision-making and policy analysis. Moreover, reducing redundant or weak predictors enhances computational efficiency, mitigates overfitting risk, and improves model generalization performance.

Although feature selection methods improve interpretability and efficiency, they also require careful validation to ensure that important predictive information is not discarded. Nonetheless, MI and Random Forest importance remain powerful and widely adopted tools in modern predictive analytics. In customer churn modeling, these approaches contribute to

constructing more parsimonious, interpretable, and statistically robust classification models, thereby strengthening both predictive reliability and practical decision-making outcomes.

Evaluation Metrics

Evaluating the performance of classification models is a critical step in predictive analytics, as it determines how effectively a model distinguishes between different outcome categories. In customer churn prediction, classification models aim to identify customers who are likely to discontinue a service versus those who are likely to remain. Accurate evaluation is essential not only for model selection but also for guiding business decisions, customer retention strategies, and policy considerations.

One of the most commonly used evaluation tools is the **confusion matrix**, which summarizes classification outcomes by comparing predicted results with actual observations. The confusion matrix typically includes four components: true positives, true negatives, false positives, and false negatives. In churn prediction, these categories represent correctly identified churners, correctly identified non-churners, incorrectly predicted churners, and missed churners respectively. This framework provides a comprehensive view of model performance beyond simple accuracy measures.

Accuracy is often the first metric reported in classification studies. It represents the proportion of correctly classified observations relative to the total number of cases. While accuracy offers an overall measure of predictive performance, it may be misleading in datasets where class imbalance exists, which is common in churn prediction since non-churning customers often outnumber churners.

To address this limitation, additional metrics such as **precision** and **recall** are widely used. Precision measures the proportion of correctly predicted churn cases among all cases predicted as churn, reflecting the reliability of positive predictions. Recall, also known as sensitivity, measures the proportion of actual churn cases correctly identified by the model. In customer churn prediction, recall is often particularly important because failing to identify customers likely to leave can result in lost revenue and missed retention opportunities.

Another widely used metric is the **F1-score**, which represents the harmonic mean of precision and recall. This metric provides a balanced assessment of model performance, particularly when dealing with imbalanced datasets. The F1-score is useful when both false positives and false negatives carry significant consequences, as is often the case in customer churn analysis.

Overall, the use of multiple evaluation metrics provides a more comprehensive assessment of classification models than relying on accuracy alone. In predictive churn modeling, combining confusion matrix analysis with accuracy, precision, recall, and F1-score enables researchers and practitioners to evaluate models more effectively, ensuring that predictive performance aligns with business objectives and decision-making requirements. These evaluation measures are therefore essential for developing reliable and actionable predictive analytics frameworks.

5.2. Empirical Literature Review

Empirical studies on customer churn prediction have expanded significantly over the past two decades, particularly within the telecommunications and financial services sectors. Early research primarily relied on classical statistical techniques such as logistic regression and discriminant analysis to model customer attrition. For example, Verbeke et al. (2012) developed churn prediction models using advanced rule induction techniques and emphasized the importance of model interpretability in supporting managerial decision-making. Their findings suggested that while more complex models may improve predictive performance, transparency remains essential for practical implementation.

As data availability increased, researchers began incorporating more advanced data mining and machine learning techniques. Hadden et al. (2007) provided a comprehensive review of computer-assisted churn management and highlighted the growing application of classification algorithms in predicting customer turnover. Their study demonstrated that machine learning approaches could outperform traditional statistical models in certain contexts, particularly when dealing with large and complex datasets.

Ensemble learning methods have also gained considerable attention in empirical churn studies. Lemmens and Croux (2006) investigated the use of bagging and boosting techniques for churn prediction and found that ensemble methods significantly improved predictive accuracy compared to single decision tree models. Similarly, Moro, Cortez, and Rita (2014) applied data-driven modeling approaches to predict customer behavior in financial marketing campaigns and reported that advanced machine learning algorithms often achieved superior predictive performance.

More recent empirical research has increasingly focused on comparing classical statistical models with modern machine learning approaches. Several studies report that Random Forest and Gradient Boosting models frequently achieve higher classification accuracy due to their ability to capture nonlinear relationships and interaction effects. However, these studies also acknowledge trade-offs between predictive performance and interpretability, as ensemble methods tend to function as “black-box” models. This has raised concerns regarding transparency and explainability, particularly in contexts where managerial accountability and regulatory compliance are important.

Despite the extensive empirical work in churn prediction, certain limitations persist. Many studies concentrate primarily on improving predictive accuracy without systematically evaluating statistical assumptions or incorporating dimensionality reduction techniques. Furthermore, while evaluation metrics such as accuracy are commonly reported, fewer studies provide comprehensive analyses using precision, recall, and F1-score to address class imbalance issues. Additionally, limited attention has been given to the broader policy implications of churn analytics, particularly in sectors such as telecommunications and fintech that play critical roles in economic development.

Overall, empirical evidence confirms that both classical statistical methods and advanced machine learning models are effective tools for predicting customer churn. However, there remains a need for a structured comparative framework that integrates statistical rigor, dimensionality reduction, and comprehensive model evaluation. This study seeks to

contribute to the existing literature by addressing these limitations through a balanced comparison of logistic regression, Random Forest, and Gradient Boosting within a unified predictive framework.

5.3. Summary of Literature

As the body of work on churn indicates, it's clear that predictive analytics is becoming a necessity for highly competitive business models that operate on a subscription basis, e.g., telecommunications or finance. Churn represents a real business threat to business stability, customer lifetime value, and overall business sustainability, as scholars have noted.

Statistically speaking, churn prediction has traditionally been addressed via a series of classification techniques, with logistic regression as a key workhorse for disentangling probability distributions that are easily interpretable.

However, in recent years, machine learning has taken a new approach to churn prediction that has outperformed traditional statistical techniques. Ensemble techniques such as Random Forest or Gradient Boosting perform extremely well in disentangling nonlinear relationships in large datasets of customers. Feature selection techniques, including mutual information or feature importance in Random Forest, have also been effective in identifying key drivers of churn while reducing redundancy to create more efficient models that are still interpretable.

Overall, a variety of predictive models have been effective in improving business outcomes for companies that focus on customer retention strategies. However, as a body of work, it's clear that there are still a number of holes to be filled in the existing literature, including a lack of focus on statistical assumptions, a lack of emphasis on interpretability, a lack of a variety of performance metrics including precision, recall, or F1 score, as well as a lack of emphasis on feature selection in churn prediction models or a lack of emphasis on policy outcomes.

As a result, this work attempts to bridge this divide by comparing classical statistical models including logistic regression to state-of-the-art machine learning models including Random Forest or Gradient Boosting while also including feature selection techniques as well as a variety of performance metrics to create a tighter model that is more effective for business decisions as well as for informing business policies.

6. Methodology

6.1. Data Acquisition Process

Data Acquisition Process

This study adopts a quantitative research approach based on the use of secondary data obtained from the Kaggle platform. The dataset was selected due to its relevance to telecommunications churn prediction and its suitability for statistical and machine learning modeling. The use of publicly available data enhances research transparency and reproducibility.

Source of Data

The dataset used in this study is the Telco Customer Churn dataset, which contains information on 7,043 telecommunications customers described by 21 attributes. These variables include demographic characteristics, subscription details, billing information, and service usage patterns. The dependent variable, labeled *Churn*, is binary and indicates whether a customer discontinued the service.

6.2. Study Variables

The dataset used in this study consists of 21 variables describing customer characteristics, service usage patterns, billing information, and churn status. These variables are categorized into dependent and independent variables for the purpose of classification modeling.

Dependent Variable

The dependent variable in this study is **Churn**, which represents whether a customer has discontinued the telecommunications service. It is a binary categorical variable with two possible outcomes: *Yes* (customer left) and *No* (customer retained). This variable serves as the target outcome in the classification models developed in this research.

Independent Variables

The independent variables consist of demographic attributes, service subscription characteristics, and financial indicators that may influence customer churn behavior.

1. Demographic Variables

- **gender**: Indicates whether the customer is male or female.
- **SeniorCitizen**: Identifies whether the customer is a senior citizen.
- **Partner**: Indicates whether the customer has a partner.

- **Dependents**: Indicates whether the customer has dependents.

These variables capture customer profile characteristics that may influence service preferences and loyalty.

2. Service Subscription Variables

- **PhoneService**
- **MultipleLines**
- **InternetService**
- **OnlineSecurity**
- **OnlineBackup**
- **DeviceProtection**
- **TechSupport**
- **StreamingTV**
- **StreamingMovies**
- **Contract**
- **PaperlessBilling**
- **PaymentMethod**

These categorical variables describe the type of services subscribed to by customers, contractual agreements, and billing preferences. Service-related attributes are particularly relevant in churn prediction, as dissatisfaction with service offerings or contract type may increase the likelihood of attrition.

3. Account and Financial Variables

- **tenure**: Number of months the customer has remained with the company.
- **MonthlyCharges**: The monthly amount charged to the customer.
- **TotalCharges**: The total amount charged over the duration of the customer's relationship.

These numerical variables reflect customer engagement duration and financial contribution. Previous research suggests that tenure and billing characteristics are strong predictors of churn behavior, as longer customer relationships often indicate higher loyalty.

4. Identifier Variable

- **customerID**: Unique identifier assigned to each customer.

This variable is used solely for identification purposes and does not contribute to predictive modeling.

Overall, the combination of demographic, service-related, and financial variables provides a comprehensive representation of customer behavior within the telecommunications sector. These variables form the basis for developing statistical and machine learning classification models aimed at predicting customer churn.

6.3. Data Cleaning Concept

Data cleaning is an essential preprocessing step in statistics and machine learning that prepares raw data for sound and reliable modeling. The main objective of data cleaning is to enhance the quality of data by detecting and correcting inconsistencies, managing missing values, transforming variable formats, and ensuring that the entire data is appropriate for analysis. Effective data cleaning improves the reliability of statistical findings, enhances the accuracy of machine learning models, and reduces the likelihood of biased or misleading findings.

For instance, in this particular study, the customer churn data required appropriate preprocessing steps before developing any machine learning or statistical models. Firstly, missing or contradictory data values were critically evaluated to ensure that the data was accurate and reliable. Particular attention was given to numerical data values, which may be incorrectly formatted or missing, thus interfering with the entire process. Missing values in data sets should always be given utmost attention since missing values may influence the accuracy of statistical findings.

Second, categorical data such as service types, methods of payment, and categories of contracts were converted to numerical forms to make them compatible with statistical models and machine learning models. The methods used for encoding these data allow the models to work properly, including logistic regression, random forest, and gradient boosting models.

Third, numerical data such as tenure, monthly charges, and total charges were also reviewed to ensure consistency, absence of outliers, and other properties that could influence the models. Feature scaling was also performed where necessary to ensure that

numerical attributes were scaled properly to support models such as logistic regression, which is sensitive to differences in scales.

Lastly, non-informative attributes such as customer identification were removed to prevent the inclusion of 'noise' in the models. These processes ensured that the data is properly preprocessed, making it ready for feature selection models such as mutual information and random forest feature importance, as well as the actual models for classification.

Data cleaning is an essential process in the application of predictive analytics, enhancing the integrity of the data, providing robust statistical models, and making the results more interpretable for the models used in predicting churn.

6.4. Statistical Assumptions of the Models

The selection of appropriate classification models requires consideration of their underlying assumptions to ensure reliable and valid results. In this study, both classical statistical and machine learning approaches are employed, each characterized by different assumptions and theoretical foundations.

Logistic Regression Assumptions

Logistic regression, as a classical statistical classification technique, is based on several key assumptions. First, the dependent variable must be binary, which is satisfied in this study since the churn variable has two categories: *Yes* and *No*. Second, observations are assumed to be independent, meaning that each customer record represents a distinct and unrelated case. This assumption is appropriate given that each row in the dataset corresponds to a different customer.

Another important assumption is the linear relationship between the independent variables and the log-odds of the dependent variable. While logistic regression does not require a linear relationship between predictors and the outcome itself, it assumes linearity in the logit transformation. Additionally, the model assumes the absence of severe multicollinearity among independent variables. High multicollinearity can distort coefficient estimates and reduce interpretability. Finally, logistic regression assumes that there are no extreme outliers that unduly influence model estimates.

Machine Learning Model Assumptions

In contrast, machine learning models such as Random Forest and Gradient Boosting make fewer strict statistical assumptions about data distribution or linearity. These ensemble methods do not require normality of predictors, linear relationships, or independence in the same strict sense as classical models. They are capable of capturing nonlinear relationships and complex interactions among variables.

However, while they are less assumption-driven, proper data preparation remains essential to avoid overfitting and ensure generalizable performance. These models rely on sufficient sample size and representative data to achieve reliable predictive outcomes.

Overall, acknowledging these assumptions ensures that the selected models are applied appropriately and that their results are interpreted within the correct methodological framework.

6.5. Model Diagnostic Tests and Evaluation

For the evaluation of the classification models developed in this research, a set of diagnostic evaluation metrics was used. Model assessment is an essential component of the predictive analytics process, as it determines the degree to which the model is able to identify the set of customers likely to churn, while the predictions made by the model are accurate enough for the purpose of decision-making. The evaluation metrics used for the purpose of this research are based on the overall predictive accuracy and the ability of the model to make accurate predictions.

One of the diagnostic tests used for the purpose of this research was the confusion matrix, which provides an overall idea regarding the outcome of the classification process by comparing the predicted result with the actual result. The test provides an overall idea regarding the predictions made by the model, such as the number of true positives, false positives, false negatives, and true negatives.

For the purpose of this research, a number of quantitative evaluation metrics were used, derived from the confusion matrix. Accuracy is the ratio of the number of accurate predictions made by the model relative to the total number of instances, while the overall evaluation is made possible by the use of this test. However, considering the imbalance in the dataset, the use of other evaluation metrics was considered essential for the purpose of this research.

Precision evaluates the ratio of successfully predicted customer churns over the total customer churns predicted by the model, which relates to the accuracy of positive predictions made by the model. Recall, also referred to as sensitivity, evaluates the ratio of actual customer churns successfully predicted by the model.

Recall is of critical importance in customer churn prediction models since failure to identify actual customer churns would lead to loss of income for the organization.

The F1-score, which is calculated by combining precision and recall in their harmonic mean, was also used for evaluating the model's performance in order to provide a comprehensive assessment of the model's performance in customer churn prediction. This measure is advantageous in dealing with unbalanced datasets since it considers both false positive and false negative rates simultaneously.

These diagnostic metrics were used for evaluating the performance of the logistic regression, Random Forest, and Gradient Boosting models in order to ensure comprehensive assessment and draw informed conclusions regarding the best approach for customer churn prediction modeling.

7. Results

7.1 Exploratory Data Analysis (EDA)

Dataset Overview

The dataset contains **7,043 observations** and **21 variables**, representing customer demographic information, service usage characteristics, billing details, and churn status.

Metric	Value
Number of observations	7,043
Number of variables	21

Target Variable Distribution

The target variable **Churn** is binary and distributed as follows:

Churn Status	Count
No	5,174
Yes	1,869

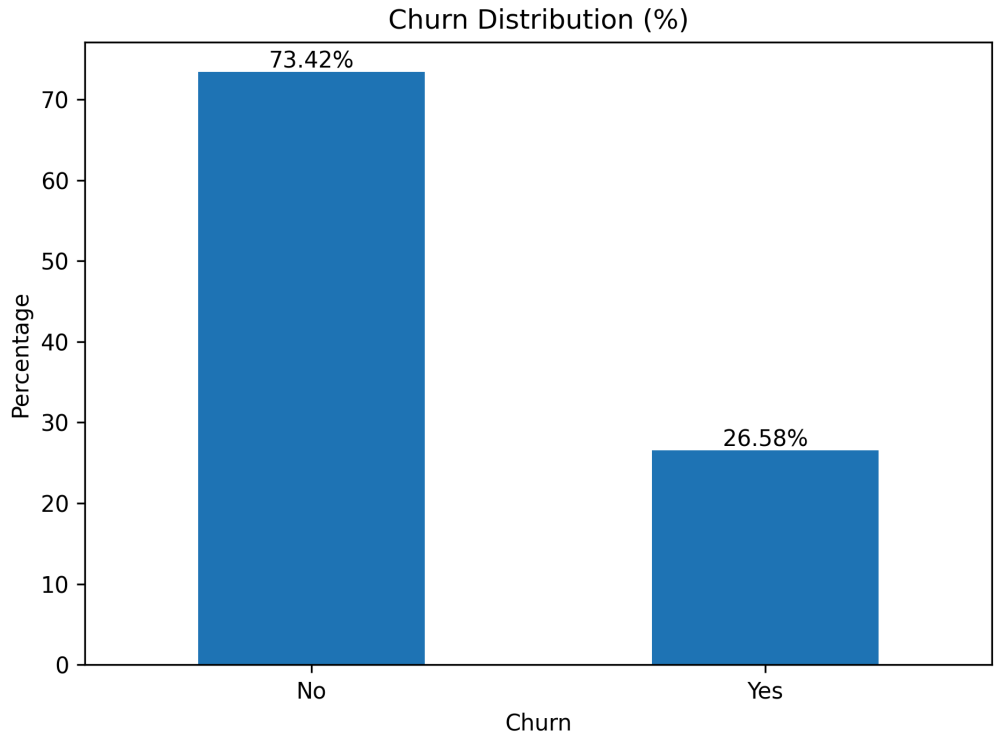

Figure 1: Customer churn distribution

This corresponds to approximately **73.5% non-churners and 26.5% churners.**

This indicates a moderately imbalanced dataset, where the majority class consists of retained customers.

7.2 Data Cleaning and Preprocessing

The dataset was preprocessed prior to modeling to ensure consistency and analytical reliability. The variable **TotalCharges**, originally stored as an object type, was converted to numeric format. Observations containing missing values were identified and removed, as they represented a small proportion of the dataset and did not materially affect sample size.

The identifier variable **customerID** was excluded from modeling due to its non-predictive nature. Categorical variables were encoded into numerical representations to ensure compatibility with classification algorithms, including logistic regression, Random Forest, and Gradient Boosting.

These preprocessing steps ensured that the dataset was complete and suitable for predictive analysis.

Key relationships between features and the target variable

Keys Numerical x churn

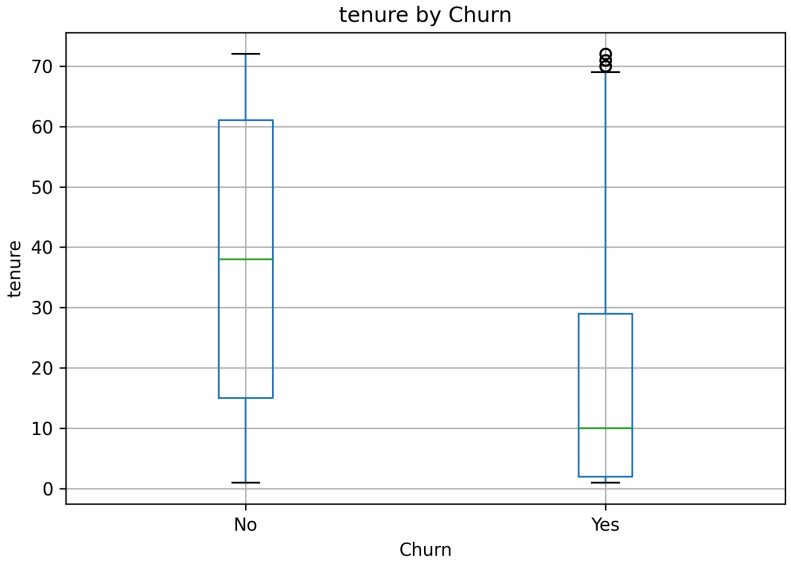

Figure 2: Tenure by Churn box plot

Customers who churn tend to have shorter tenure compared to retained customers, suggesting that customer attrition is more common among recently acquired subscribers.

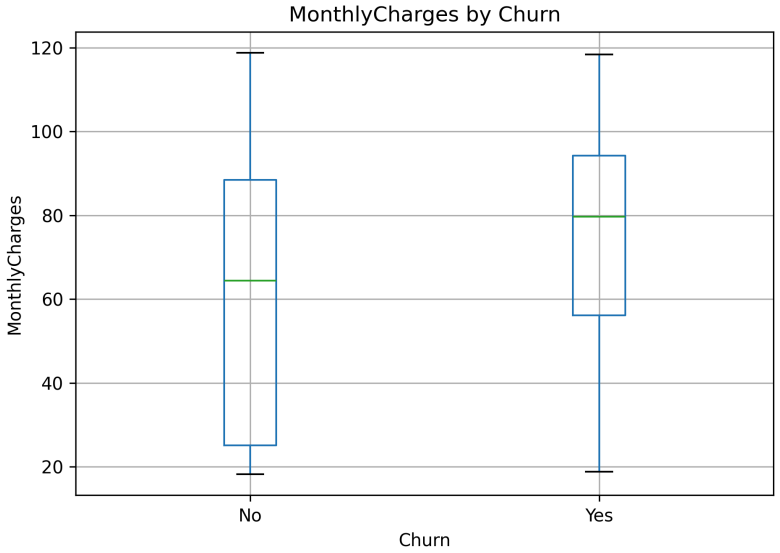

Figure 3: MonthlyCharges by Churn box plot

Higher monthly charges appear associated with increased churn probability, indicating that pricing may influence customer retention.

Keys Categoricals by Churn countplot

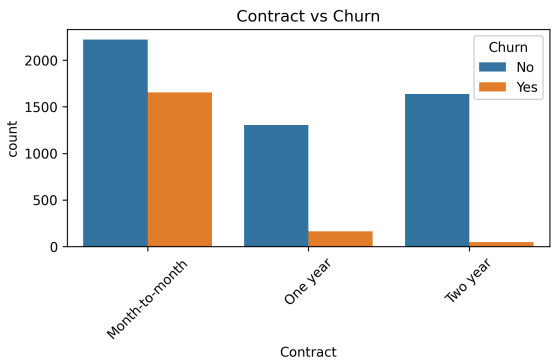

Figure 4: Contract by Churn

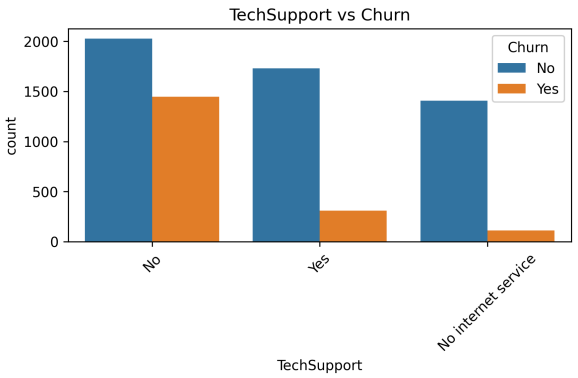

Figure 5: TechSupport by Churn

Customers on month-to-month contracts and those lacking technical support show higher churn tendencies.

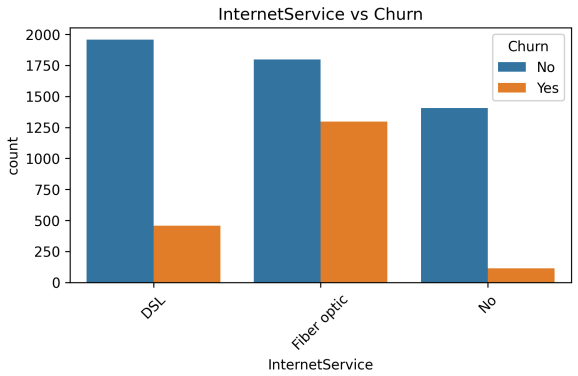

Figure 6: InternetService by Churn

Customers on Fiber Optic service are more likely to churn

7.3 Normality Testing

To assess the distributional properties of key numerical variables, normality tests were conducted using the **Shapiro-Wilk test**, complemented by graphical inspection through **Q–Q plots** and **P–P plots**.

Variable	Shapiro-Wilk Statistic	p-Value
Tenure	0.9034	< 0.001
MonthlyCharges	0.9221	< 0.001
TotalCharges	0.8622	< 0.001

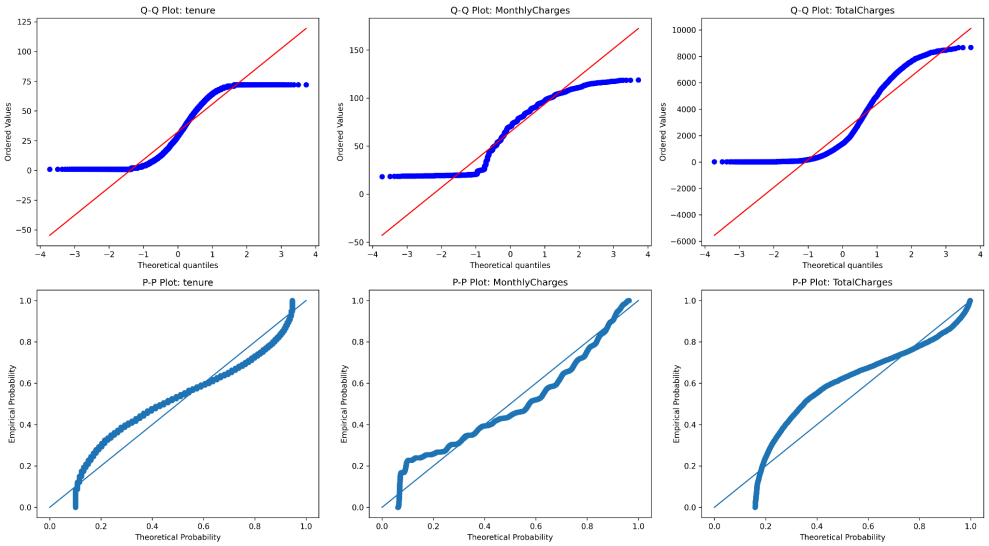

Figure 7: Normality Plots

For all variables, the p-values are below 0.05, leading to rejection of the null hypothesis of normality. This indicates that the distributions of tenure, MonthlyCharges, and TotalCharges significantly deviate from a normal distribution.

Graphical analysis through Q–Q and P–P plots further confirmed these findings. The plotted points showed noticeable departures from the diagonal reference line, particularly in the tails of the distributions. These deviations reflect skewness and heterogeneity in customer tenure and billing characteristics, which are common in telecommunications datasets.

7.4 Feature Engineering and Selection

One-Hot Encoding

Categorical variables were transformed into numerical format using **one-hot encoding** to ensure compatibility with classification algorithms. After encoding, the dataset expanded from its original structure to a matrix of:

- **7,032 observations**
- **30 encoded features**

This transformation allowed categorical service and billing attributes to be incorporated into statistical and machine learning models without introducing ordinal bias.

Feature Selection

To reduce dimensionality and identify the most informative predictors, two complementary feature selection techniques were applied:

- **Random Forest feature importance**
- **Mutual Information (MI)**

Both methods ranked variables according to their predictive relevance for churn classification.

Random Forest Feature Importance

Random Forest importance identified billing, service type, and tenure-related variables as strong predictors. Among the highest-ranked features were:

- InternetService_No
- PaymentMethod_Electronic check
- InternetService_Fiber optic
- MonthlyCharges
- Tenure
- TotalCharges

These results indicate that service subscription type, payment method, and customer duration are key determinants of churn behavior.

Mutual Information

Mutual Information analysis similarly highlighted:

- Tenure
- Contract_Two year
- MonthlyCharges
- InternetService_Fiber optic
- TotalCharges

The consistency between Random Forest and Mutual Information rankings reinforces the robustness of these predictors.

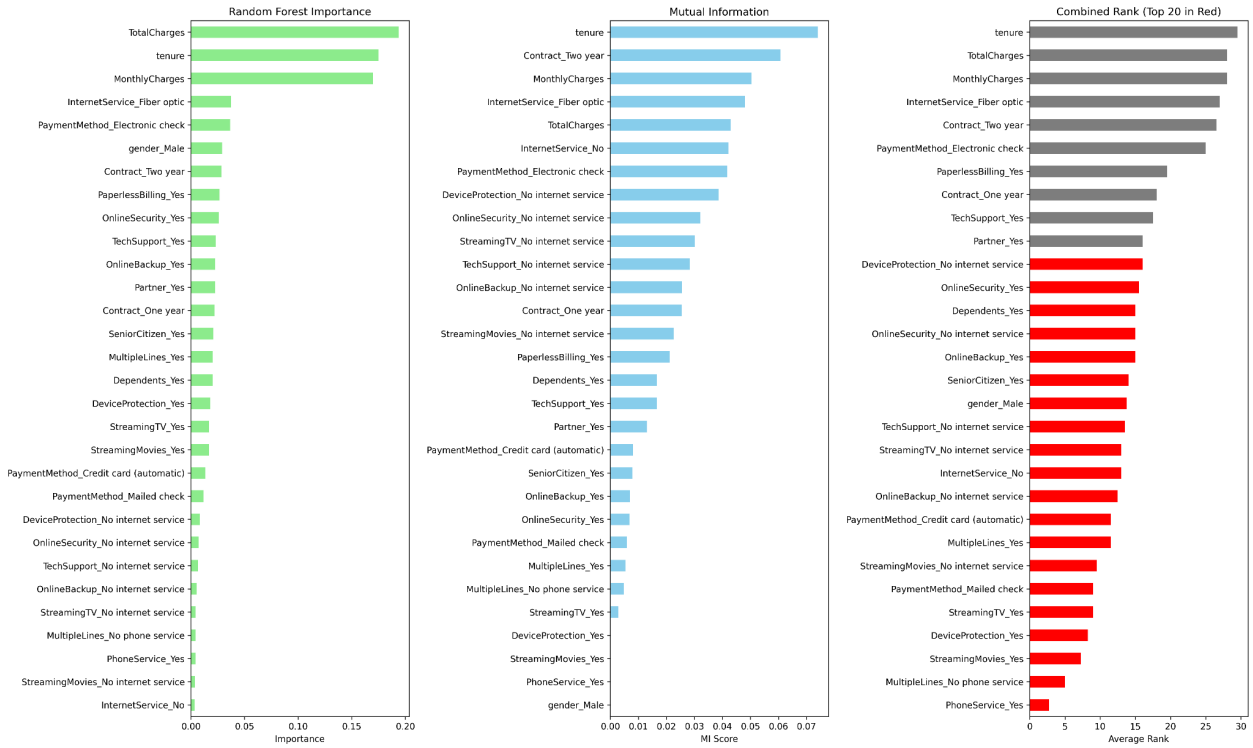

Figure 8: Feature Importance RF x MI

Both feature selection techniques confirm that:

- Contract type
- Internet service type
- Billing characteristics
- Customer tenure

are the most influential drivers of churn. By focusing on these predictors, dimensionality was effectively reduced while preserving interpretability and predictive strength.

7.5 Train-Test Split

The dataset was divided into training and testing subsets to evaluate model generalization performance. The training set was used for model development, while the test set was reserved for out-of-sample evaluation to prevent data leakage and overfitting.

7.6 Model Implementation

Three supervised classification models were implemented:

1. Logistic Regression (baseline statistical model)
2. Random Forest (ensemble tree-based model)
3. Gradient Boosting (boosting-based ensemble model)

Model performance was evaluated using:

- Accuracy
- Confusion Matrix
- Precision
- Recall
- F1-score

These metrics provide a comprehensive assessment of classification effectiveness, particularly in the presence of class imbalance.

7.7 Model Performance Comparison

To evaluate predictive effectiveness, three supervised classification models were implemented: **Logistic Regression (LR)**, **Random Forest (RF)**, and **Gradient Boosting (GB)**. Model performance was assessed on the test set using accuracy, precision, recall, F1-score, confusion matrix analysis, and Precision–Recall (PR) curve performance.

Logistic Regression

The logistic regression model achieved a test accuracy of 80.45%, the highest among the three models.

Performance Metrics:

- Precision: 0.6505
- Recall: 0.5722
- F1-score: 0.6088

- PR-AUC: 0.621

Logistic regression demonstrates balanced performance, with relatively strong precision and moderate recall. The model correctly identifies a substantial portion of churners while maintaining a controlled false positive rate. Its probabilistic nature and interpretability make it a reliable baseline model for churn prediction.

Random Forest

The Random Forest model achieved an accuracy of 77.40%.

Performance Metrics:

- Precision: 0.5952
- Recall: 0.4679
- F1-score: 0.5240
- PR-AUC: 0.598

Although Random Forest captures nonlinear interactions, its recall is comparatively low, indicating a higher number of missed churners. In churn prediction contexts, this may reduce its practical effectiveness, particularly when identifying at-risk customers is a priority.

Gradient Boosting

The Gradient Boosting model achieved an accuracy of 74.13%.

Performance Metrics:

- Precision: 0.5084
- Recall: 0.8128
- F1-score: 0.6255
- PR-AUC: 0.654

Gradient Boosting demonstrates the highest recall (81.28%) and the strongest PR-AUC (0.654), indicating superior ability to detect churners across classification thresholds. Although its precision is lower resulting in more false positives the model minimizes false negatives, which is critical in churn management where missed churners represent lost revenue.

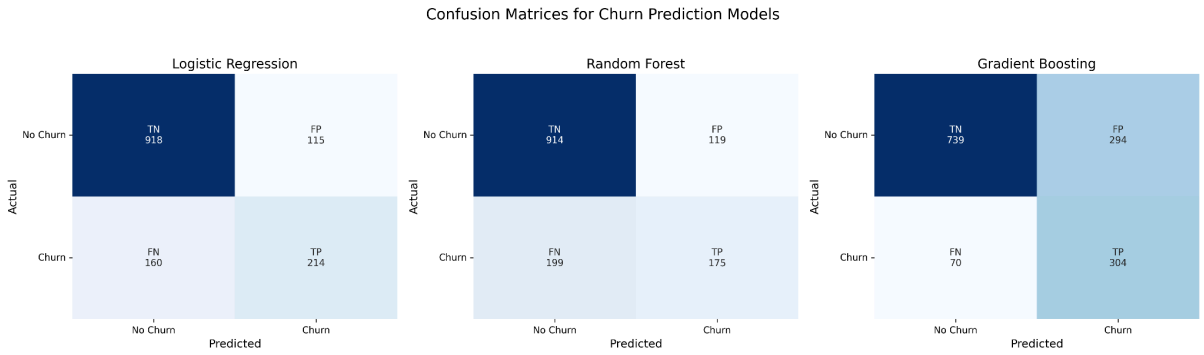

Figure 9: Confusion Matrix (Logistic Regression, Random Forest, Gradient Boosting

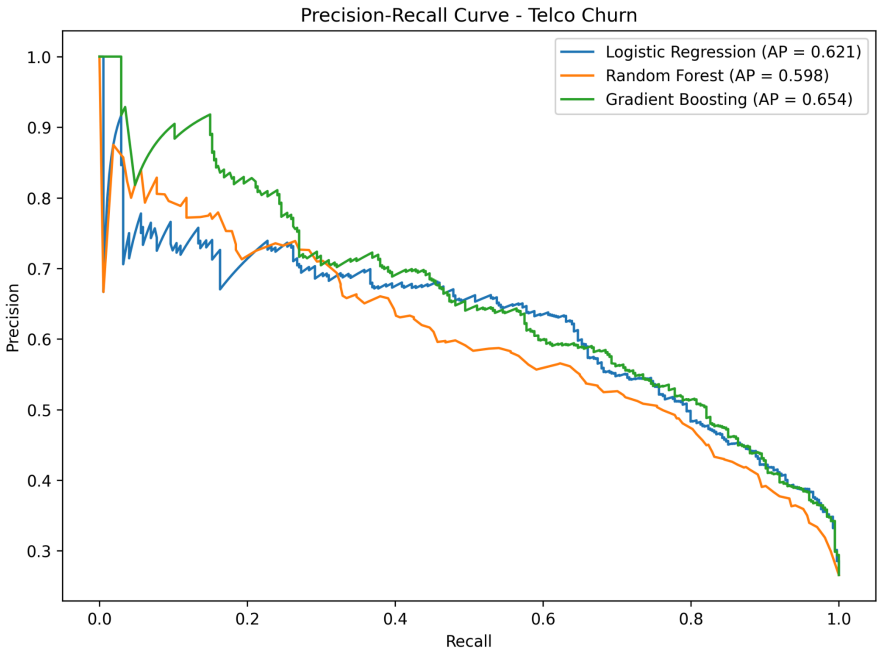

Figure 10: PR-Curve models

While Logistic Regression achieves the highest overall accuracy, Gradient Boosting outperforms the other models in recall, F1-score, and PR-AUC. Since churn prediction involves asymmetric costs—where failing to identify a churner results in revenue loss—recall becomes a critical evaluation metric. From a business perspective, Gradient Boosting provides the most effective detection of at-risk customers, despite generating more false positives.

7.8 Most Important Metric in Customer Churn Scenario

In a telecommunications churn context, **recall remains a critical metric**, but its importance must be interpreted in business terms rather than clinical risk. A false negative in churn prediction represents a customer who is at risk of leaving but is not identified by the model. This results in lost revenue, reduced customer lifetime value, and potentially increased acquisition costs to replace that customer.

Maximizing recall ensures that a higher proportion of at-risk customers are correctly identified and targeted with retention strategies. However, unlike medical diagnosis, false positives in churn prediction carry a financial cost rather than a life-threatening consequence. A false positive corresponds to offering incentives or retention efforts to customers who would not have churned, thereby increasing marketing expenditure.

Therefore, in telecom churn modeling, the most important metric depends on business priorities. If the objective is to minimize revenue loss, recall becomes the primary focus. If retention campaigns are costly, precision must also be considered to avoid excessive false positives. In practice, metrics such as **F1-score** and **Precision–Recall AUC** provide a balanced evaluation, aligning predictive performance with financial impact.

In this study, Gradient Boosting demonstrates the highest recall and PR-AUC, indicating stronger effectiveness in identifying customers at risk of churn.

8. Discussion, Summary, Conclusion and Recommendations

8.1 Discussion

This study examined customer churn prediction in the telecommunications sector using three supervised classification models: Logistic Regression, Random Forest, and Gradient Boosting. The objective was to evaluate predictive performance while balancing statistical rigor and interpretability.

The results indicate that Logistic Regression achieved the highest overall accuracy (80.45%), demonstrating strong baseline performance and interpretability. However, Gradient Boosting achieved the highest recall (81.28%), F1-score (0.6255), and Precision–Recall AUC (0.654), indicating superior capability in identifying customers at risk of churn.

In churn prediction, false negatives represent lost customers and lost revenue. Therefore, recall becomes particularly important. The Gradient Boosting model, although producing more false positives, minimized missed churners. This trade-off is strategically acceptable in

many telecommunications contexts, where retaining at-risk customers often outweighs the cost of unnecessary retention offers.

Feature selection analysis using Mutual Information and Random Forest importance identified tenure, contract type, billing characteristics, and internet service type as the most influential predictors. These findings align with existing literature, confirming that long-term contracts and longer customer tenure significantly reduce churn probability.

8.2 Summary

This study applied statistical and machine learning techniques to predict customer churn in a telecommunications dataset containing 7,032 cleaned observations and 30 encoded features.

The analysis included:

- Exploratory Data Analysis (EDA)
- Data cleaning and preprocessing
- Normality testing
- Feature selection using Mutual Information and Random Forest importance
- Model development using Logistic Regression, Random Forest, and Gradient Boosting
- Model evaluation using accuracy, precision, recall, F1-score, confusion matrix, and PR-AUC

Comparative results demonstrated that Gradient Boosting provides the strongest churn detection capability, while Logistic Regression offers higher interpretability and strong overall accuracy.

8.3 Conclusion

Customer churn prediction plays a critical role in revenue protection and strategic decision-making within competitive telecommunications markets. This study confirms that advanced ensemble methods, particularly Gradient Boosting, outperform classical models in detecting at-risk customers.

However, model selection should not rely solely on accuracy. In churn prediction, evaluation metrics such as recall and Precision–Recall AUC provide more meaningful insight into financial implications. The results demonstrate that combining statistical rigor, feature selection, and comprehensive evaluation metrics leads to more reliable and actionable predictive frameworks.

The study highlights the importance of aligning model performance metrics with business objectives rather than relying on conventional accuracy measures alone.

8.4 Recommendations

Based on the findings, the following recommendations:

1 - Focus retention strategies on customers with:

- Short tenure
- Month-to-month contracts
- Higher monthly charges
- Fiber optic internet service

2 - For regulatory bodies, monitor churn trends as indicators of service quality and market competitiveness.

9. References and Appendix

9.1 References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

9.2 Appendix

(See Figures 1 to 10)

10. Real-Time Customer Churn Prediction System

10.1 System Extension to Real-Time Deployment

Beyond the statistical analysis and comparative modeling conducted in this study, a real-time customer churn prediction system was implemented to demonstrate the practical applicability of the developed model in a production-like environment.

The deployed system enables end users to input customer characteristics through a web-based interface and instantly receive a churn probability estimate along with a categorized risk level. This extension bridges the gap between theoretical predictive modeling and operational decision support.

Github link: https://github.com/samuelOndata/customer-telecom-churn-prediction

10.2 System Architecture

The real-time system follows a modular, production-oriented architecture composed of:

- **Streamlit Web Interface** for user interaction
- **Serialized Scikit-Learn Pipeline** containing preprocessing and the Gradient Boosting classifier
- **PostgreSQL Database** for logging prediction results
- **Docker and Docker Compose** for containerized deployment

The trained machine learning pipeline integrates feature scaling, categorical encoding, and classification into a single serialized artifact (`churn_pipeline.pkl`). This ensures consistency between training and deployment environments and eliminates feature mismatch risks.

When a user submits customer attributes (e.g., tenure, contract type, monthly charges, service features), the application:

1. Loads the serialized pipeline
2. Applies preprocessing automatically
3. Computes churn probability using `predict_proba()`

4. Assigns a risk category (Low, Medium, High)
5. Stores the input data and prediction results in PostgreSQL

10.3 Risk Classification Strategy

Predicted probabilities are translated into actionable business categories:

- **High Risk (≥ 70%)** – Immediate retention action recommended
- **Medium Risk (40%–69%)** – Monitoring and targeted offers suggested
- **Low Risk (< 40%)** – Stable customer

This segmentation supports strategic customer relationship management by prioritizing intervention resources toward the most vulnerable customers.