Interdisciplinary Journal of the African Alliance For Research, Advocacy & Innovation
ISSN (O) : 3093-4664
Vol.2, Issue 4 | Oct–Dec 2026
www.ijaarai.org

Predicting Customer Churn in Telecommunications: A Controlled Comparison of Machine Learning Models Using Mutual Information Feature Selection

Author(s): Ernest Edem Dzisah

Abstract

Customer churn significantly impacts telecommunications revenue and acquisition costs, but machine learning can proactively identify at-risk customers using historical data. This study compares four supervised algorithms for churn prediction: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

Using a dataset of 7,032 customer records, the study applied a stratified 80:20 train-test split. Preprocessing included median imputation and standardization for numerical features, and most-frequent imputation with one-hot encoding for categorical data. Mutual Information feature selection, using a fixed random state of 42, retained the ten highest-ranked features. All models shared this identical pipeline to ensure a controlled comparison.

Performance was evaluated on an independent holdout test set using accuracy, precision, recall, F1-score, ROC AUC, and confusion matrices, alongside five-fold stratified cross-validation on the training data. On the holdout set, Logistic Regression achieved the highest accuracy (78.82%), while Gradient Boosting recorded 78.61% accuracy and the highest ROC AUC (83.60%), slightly outperforming Logistic Regression (83.39%).

During cross-validation, Gradient Boosting maintained a superior mean ROC AUC of 84.48% ± 0.65%, compared to 84.29% ± 0.52% for Logistic Regression. However, Gradient Boosting identified only 51.60% of actual churn cases in the holdout set, underscoring the ongoing difficulty of accurately capturing leaving customers.

Ultimately, the findings demonstrate that predictive performance can be closely matched across different algorithms. Reliable churn prediction requires a holistic approach using multiple evaluation metrics rather than relying on accuracy alone.

Keywords: *customer churn, telecommunications, machine learning, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Mutual Information, feature selection, predictive modeling*

1. Introduction

Customer churn is a major concern for telecommunications companies because customers can move between service providers with relatively little difficulty. Competition among providers gives customers access to different pricing plans, service packages, network options, and contract arrangements. When a customer terminates a service or moves to another provider, the company loses current revenue and may also incur additional costs when trying to replace that customer. For this reason, identifying customers who are likely to leave has become an important application of predictive analytics in the telecommunications sector. Previous research has examined churn as a

https://doi.org/10.64261/hfpwc474

classification problem in which customer characteristics and service information are used to estimate the likelihood of future attrition (Wei & Chiu, 2002; Hung et al., 2006; Verbeke et al., 2012).

The growth of customer relationship management systems has made larger amounts of customer information available for analysis. Telecommunications datasets can contain demographic information, service subscriptions, contract details, payment methods, tenure, and patterns of service use. These variables provide opportunities to identify relationships between customer characteristics and churn behaviour. Machine learning methods are particularly useful in this setting because they can learn patterns from historical customer records and use those patterns to classify new customers according to their predicted churn status. Research in the telecommunications domain has consequently moved from traditional statistical approaches toward a wider range of machine learning techniques, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, and boosting methods. A recent systematic review covering studies published between 2010 and 2024 identified 199 eligible studies and reported extensive use of both classical machine learning and more recent deep learning approaches for telecom churn prediction (Jain et al., 2026).

Model comparison is therefore a recurring part of churn prediction research. Different classifiers can produce different results when applied to the same prediction problem. Logistic Regression provides a relatively simple statistical baseline, while Decision Trees can represent nonlinear decision rules. Random Forest combines multiple decision trees, and Gradient Boosting builds an ensemble by successively improving the errors made by previous trees. These differences make comparative experiments useful when the models are trained and evaluated under the same conditions. Earlier studies have compared several classification techniques for telecom churn prediction, often together with different preprocessing, feature selection, class balancing, or parameter optimization procedures (Coussement & Van den Poel, 2008; Burez & Van den Poel, 2009; Edwine et al., 2022).

Feature selection is another important part of the prediction process. Telecommunications datasets may contain a mixture of demographic, service, contract, and billing variables, and not every available variable contributes equally to the prediction task. Selecting informative variables can reduce the number of features presented to a model and may also simplify the resulting prediction process. Mutual Information is one established method for measuring the dependence between variables and a target variable. It has been applied in telecom churn studies as a feature selection technique. For example, Edwine et al. (2022) used Mutual Information to identify important features before comparing Random Forest, Support Vector Machine, and K Nearest Neighbour models. Their experimental design also included class balancing and hyperparameter optimization, showing how several methodological choices can influence the final results.

These methodological differences create an important problem when results from separate studies are compared directly. Reported performance can depend on the dataset, class distribution, preprocessing procedure, feature selection method, sampling strategy, model configuration, and evaluation protocol. A model that reports a high accuracy in one study may not produce the same result on another dataset or under another experimental design. The recent systematic review by Jain et al. (2026) highlights this variation across the telecom churn literature and identifies class imbalance, high dimensional data, evaluation measures, interpretability, and differences in modelling approaches as continuing issues in the field.

https://doi.org/10.64261/hfpwc474

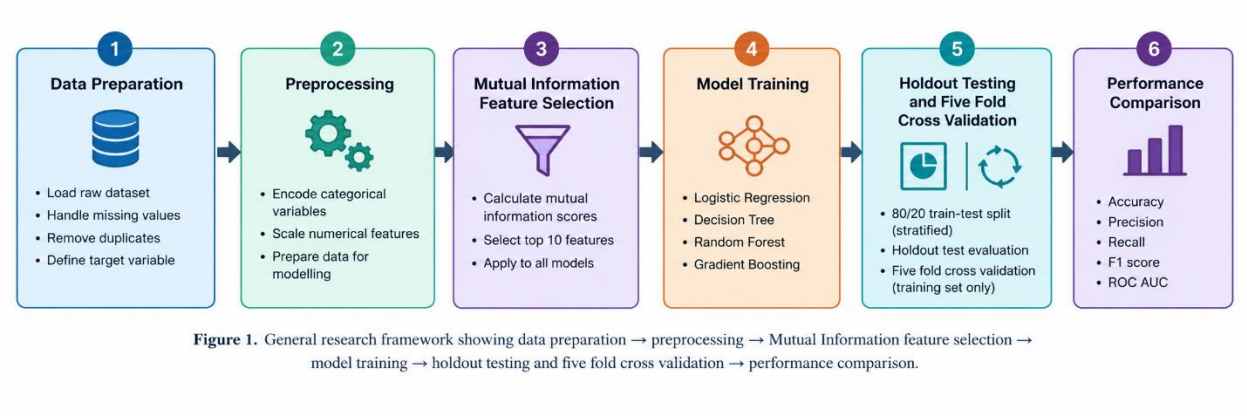


Figure 1. General research framework showing data preparation → preprocessing → Mutual Information feature selection → model training → holdout testing and five fold cross validation → performance comparison.

Figure 1: General research framework data preparation → preprocessing → Mutual Information feature selection → model training → holdout testing and five fold cross validation → performance comparison

Although previous studies have provided extensive evidence on customer churn prediction, there remains value in controlled comparisons in which several commonly used classifiers are evaluated using the same dataset and experimental procedure. Such a design reduces the influence of methodological differences when interpreting model performance. It also allows the comparison to focus on the behaviour of the selected algorithms rather than on differences between separate datasets or research protocols.

This study addresses that need by comparing Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting for telecommunications customer churn prediction under a common experimental framework. The study uses the same preprocessing procedure and Mutual Information based feature selection approach for all four models. The data are divided into training and holdout test sets using stratified sampling. Model performance is assessed on the heldout test set using accuracy, precision, recall, F1 score, and ROC AUC. Five fold stratified cross validation is also applied to the training data to examine the consistency of model performance.

The primary research question is:

How do Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting perform for telecommunications customer churn prediction when evaluated using a common preprocessing and Mutual Information based feature selection pipeline?

Three supporting questions are also examined:

1. How do the four models differ in accuracy, precision, recall, F1 score, and ROC AUC on the heldout test set?

https://doi.org/10.64261/hfpwc474

2. How stable is model performance across five fold stratified cross validation on the training data?
3. Which of the evaluated models provides the strongest predictive performance under the experimental conditions used in this study?

The general objective of this study is to empirically compare selected supervised machine learning models for telecommunications customer churn prediction using a consistent preprocessing, feature selection, training, and evaluation framework. The specific objectives are to prepare and preprocess the customer churn dataset, apply Mutual Information based feature selection, train and compare the four selected classifiers, evaluate their performance using multiple classification measures, examine performance stability through five fold stratified cross validation, and identify the strongest performing model within the experimental setup.

The study makes three main contributions. First, it provides a controlled comparison of four commonly used classification models using the same dataset and experimental conditions. Second, it applies Mutual Information based feature selection consistently across the compared models with a fixed random state, rather than changing the feature selection procedure between experiments. Third, it combines holdout test evaluation with five fold stratified cross validation, providing two complementary views of model performance. The study does not attempt to establish that one algorithm is universally superior. Instead, it reports the differences observed under the stated experimental conditions and discusses their practical and methodological limitations.

Table 1. Research questions and corresponding evaluation approach

Research question	Evaluation approach	Data used	Main measures
RQ1. How do Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting differ in predictive performance on the held out test set?	Comparative holdout test evaluation	1,407 test observations	Accuracy, precision, recall, F1 score, ROC AUC, confusion matrix
RQ2. How stable is model performance across five fold stratified cross validation?	Five fold stratified cross validation	5,625 training observations	Mean and standard deviation of accuracy, weighted precision, weighted recall, weighted F1 score, and ROC AUC
RQ3. Which of the evaluated models provides the strongest predictive	Comparative assessment of holdout and cross validation results	Training and holdout test data	ROC AUC, accuracy, precision, recall, F1 score, and

https://doi.org/10.64261/hfpwc474

performance under the experimental conditions?			performance consistency

Table note: *All four models were evaluated using the same preprocessing procedure and Mutual Information based feature selection pipeline. The holdout test set was kept separate from model development and cross validation.*

The remainder of this paper is organized as follows. Section 2 reviews previous research on telecommunications customer churn prediction, machine learning models, feature selection, class imbalance, and model evaluation. Section 3 describes the dataset, preprocessing procedure, feature selection method, model configurations, training strategy, and evaluation protocol. Section 4 presents the experimental results. Section 5 discusses the findings in relation to previous research and the objectives of the study. Section 6 concludes the paper by summarizing the main findings, limitations, and directions for further research.

2. Literature Review

2.1 Customer Churn Prediction in Telecommunications

Customer churn prediction has been studied for many years in the telecommunications sector. Early research treated churn as a data mining and classification problem in which information about customer behaviour and service use was used to identify customers who were likely to leave. Wei and Chiu (2002), for example, developed a data mining approach for converting telecommunications call detail information into churn prediction variables. Hung et al. (2006) later examined the application of data mining methods to telecom churn management and demonstrated the value of predictive models for identifying customers at risk of leaving.

The importance of churn prediction has remained relevant as telecommunications markets have become more competitive. Customer records now commonly contain demographic information, contract details, billing information, service subscriptions, and measures of customer tenure. These variables provide several possible signals of customer attrition. Recent research continues to describe churn prediction as an important problem because retaining existing customers is generally more desirable for service providers than losing customers and replacing them.

The nature of available data has also changed the scope of churn research. Earlier studies often relied on call records and traditional customer information, whereas more recent studies have examined larger combinations of demographic, account, service, and usage variables. A 2026 systematic review of telecom churn research identified studies using benchmark datasets such as Orange, Cell2cell, and IBM or Kaggle telecom data, as well as data obtained directly from telecommunications operators. The review included 199 studies selected from 299 candidates identified through searches of Scopus, ACM Digital Library, and IEEE Xplore.

2.2 Machine Learning Models for Churn Prediction

Machine learning has become a common approach to customer churn prediction because churn can be formulated as a supervised classification problem. A model is trained using historical customer records for which the churn outcome is known and is then used to classify customers whose future status is unknown. Different algorithms approach this task in different ways, and their performance can vary according to the structure of the data and the experimental conditions.

Logistic Regression is frequently used as a baseline because it provides a relatively simple approach to binary classification. Decision Trees use a sequence of decision rules to divide observations into

groups. Random Forest extends the tree based approach by combining predictions from multiple trees, while Gradient Boosting builds an ensemble sequentially by giving greater attention to observations that were not adequately classified by earlier learners. Random Forests and boosting methods have received considerable attention in predictive modelling because they can represent nonlinear relationships and interactions among variables (Breiman, 2001; Friedman, 2001).

Telecom churn research has evaluated a broad range of models. Coussement and Van den Poel (2008) examined Support Vector Machines for churn prediction and compared different parameter selection approaches. Verbeke et al. (2012) studied churn prediction from a profit perspective and demonstrated that the choice of model can be considered alongside the economic consequences of customer retention decisions. More recent work has expanded the range of compared methods to include ensemble methods, neural networks, and other advanced models. The 2026 systematic review by Jain, Khunteta, and Sharma reports that Logistic Regression, Decision Trees, Support Vector Machines, Random Forests, and Gradient Boosting remain among the commonly examined classical methods, alongside newer deep learning approaches.

https://doi.org/10.64261/hfpwc474

Table 2: Previous studies comparing machine learning models for telecommunications customer churn prediction

Study	Dataset	Models	Preprocessing or balancing method	Evaluation measures	Main finding
Wei & Chiu (2002)	Mobile telecommunications customer records from Taiwan, including contractual information and call patterns	Decision Tree and a multi classifier combination approach	Call pattern changes and contractual data; class imbalance addressed through a multi classifier combination approach	Predictive effectiveness and lift	Recent call pattern information provided useful signals for identifying potential churners. (ScienceDirect)
Hung et al. (2006)	Wireless telecommunications customer data from Taiwan	Decision Tree, Neural Network, and other data mining techniques	Customer demographics, billing information, contract and service status, call records, and service change information	Predictive accuracy and propensity to churn	Decision Tree and Neural Network approaches produced useful churn prediction models using multiple customer information sources. (ScienceDirect)
Verbeke et al. (2012)	Eleven real world telecommunications datasets from operators in different countries	Multiple classification techniques	Comparative modelling using different customer and telecom variables; analysis included profit based evaluation	Statistical performance measures and profit based measures	Several classifiers produced comparable performance, while profit based evaluation provided a different basis for selecting models than statistical measures alone. (ScienceDirect)
Edwine et al. (2022)	Two telecom datasets with 3,333 and 7,043 observations	Random Forest, Support Vector Machine, K Nearest Neighbour	Mutual Information feature selection, controlled ratio undersampling, and hyperparameter	Accuracy, recall, precision, AUC, F1 score, MAE	The optimized Random Forest using Grid Search and low ratio undersampling produced the strongest results in

https://doi.org/10.64261/hfpwc474

			optimization using Grid Search, Random Search, and Genetic Algorithms		their experimental setting. (DOI)
Sana et al. (2022)	Multiple telecommunications churn datasets	Several machine learning classifiers	Data transformation, feature selection, and parameter optimization	AUC, precision, recall, F measure	The study showed that combining data transformation and feature selection can affect the predictive performance of churn models. (PLOS)
Wagh et al. (2024)	Telecommunications customer churn data	Multiple machine learning techniques	Data preparation and comparative machine learning modelling	Classification performance measures	The study demonstrated continued use of comparative machine learning approaches for telecom churn prediction. (DOI)
Present study	7,032 processed telecommunications customer records	Logistic Regression, Decision Tree, Random Forest, Gradient Boosting	80:20 stratified split, numerical and categorical preprocessing, Mutual Information feature selection, top 10 features; no resampling	Accuracy, precision, recall, F1 score, ROC AUC, confusion matrix, five fold stratified cross validation	Provides a controlled comparison in which the same preprocessing, feature selection, data split, and evaluation framework are applied to all four models.

https://doi.org/10.64261/hfpwc474

Note. *The studies use different datasets, preprocessing procedures, feature selection methods, class balancing strategies, model configurations, and evaluation protocols. Their reported performance values should therefore not be interpreted as directly comparable numerical benchmarks. The purpose of the table is to show methodological differences across the literature and position the controlled experimental design of the present study.*

The continued use of several different algorithms suggests that there is no single model that can be assumed to perform best across all churn datasets. Reported performance depends on the characteristics of the dataset and the procedures used before and during model training. This makes direct comparison across independent studies difficult. A controlled experiment in which the same dataset, preprocessing procedure, feature selection method, and evaluation strategy are applied to each classifier provides a more consistent basis for comparison.

2.3 Feature Selection in Customer Churn Prediction

Feature selection is an important consideration when customer datasets contain many available variables. Some variables may provide useful information about churn, while others may contribute little to the classification task. Removing less informative variables can reduce the dimensionality of the input data and can make the modelling process more focused.

Mutual Information is one established approach to feature selection. It measures the statistical dependence between variables and can identify features that contain information about a target variable. Peng et al. (2005) examined Mutual Information based criteria for feature selection and showed how information based measures can be used to identify relevant variables. Mutual Information has also been applied specifically to customer churn prediction.

Sana et al. (2022) investigated data transformation and feature selection for telecom customer churn prediction. Their study used several publicly available telecommunications datasets and applied feature selection together with parameter optimization. The authors evaluated their models using AUC, precision, recall, and F measure and reported improvements under their experimental conditions after applying their proposed data transformation and feature selection procedures.

Edwine et al. (2022) also used Mutual Information in a telecom churn study. Their procedure first used Mutual Information to identify important features, followed by controlled ratio undersampling to address class imbalance. Random Forest, Support Vector Machine, and K Nearest Neighbour models were then optimized using Grid Search, Random Search, and Genetic Algorithms. The study evaluated the models using accuracy, recall, precision, AUC, F1 score, and Mean Absolute Error.

These studies show that feature selection is already an established part of churn prediction research. They also demonstrate that feature selection is often combined with other methodological choices. Sana et al. (2022) combined feature selection with data transformation and parameter optimization, while Edwine et al. (2022) combined Mutual Information with class balancing and several parameter optimization methods. Consequently, an observed difference in model performance may reflect several stages of the experimental procedure rather than the classifier alone.

An important distinction should therefore be made between studying the effect of feature selection and using feature selection as a common part of a model comparison. The present study follows the second approach. Mutual Information is applied consistently to the four selected models. The study does not include a separate experiment comparing models with and without Mutual Information, and therefore it does not attempt to measure the independent effect of the feature selection procedure.

https://doi.org/10.64261/hfpwc474

2.4 Class Imbalance and Evaluation Measures

Class imbalance is another recurring issue in customer churn prediction. In many telecom datasets, the number of customers who remain with a provider is substantially larger than the number who leave. A classifier can therefore obtain a relatively high overall accuracy while still failing to identify a substantial proportion of customers who actually churn. This makes the choice of evaluation measures particularly important.

Burez and Van den Poel (2009) examined class imbalance in customer churn prediction and showed that imbalance can affect the ability of predictive models to identify churners. More generally, He and Garcia (2009) describe class imbalance as a challenge for classification because standard learning procedures may be dominated by the majority class.

Recent telecom studies have addressed this problem in different ways. Edwine et al. (2022), for example, used controlled ratio undersampling before model training. Their two datasets also had different class distributions, with one containing approximately 14.5 percent churners and the other approximately 27 percent churners. Sana et al. (2022) used several publicly available datasets and evaluated models using measures beyond accuracy, including AUC, precision, recall, and F measure.

The present study uses a telecommunications dataset in which churn represents the minority class. Rather than applying an undersampling or oversampling procedure, the study retains the original class distribution and evaluates the models using several measures. Accuracy is reported together with precision, recall, F1 score, and ROC AUC. This allows the results to reflect more than the overall proportion of correct classifications.

Table 3. Evaluation measures used in previous churn studies and in the present study

Measure	Purpose	Relevance to churn prediction
Accuracy	Measures the proportion of all customer records that are classified correctly.	Provides an overall measure of classification correctness. However, accuracy alone can hide poor detection of churners when the classes are imbalanced.
Precision	Measures the proportion of customers predicted as churners who actually churned.	Indicates how reliable churn predictions are and helps assess the number of false churn predictions.
Recall	Measures the proportion of actual churners that are correctly identified.	Particularly important because missed churners may represent customers who could otherwise have been targeted for retention.
F1 score	Calculates the harmonic mean of precision and recall.	Provides a single measure that balances the ability to identify churners with the reliability of churn predictions.
ROC AUC	Measures the ability of a classifier to distinguish between churn and non churn cases across classification thresholds.	Provides a threshold independent measure of model discrimination and is useful when comparing classifiers under different probability thresholds.

https://doi.org/10.64261/hfpwc474

Confusion matrix	Summarizes true positives, true negatives, false positives, and false negatives.	Shows the types of classification errors made by each model and provides the basis for interpreting precision and recall.
Specificity	Measures the proportion of actual non churners correctly identified.	Helps assess how well a model avoids incorrectly identifying customers as likely to churn. Used in some previous churn studies but not included as a primary metric in the present study.
Mean Absolute Error (MAE)	Measures the average absolute difference between predicted and observed values.	Has been used in some previous churn studies, but it is less directly aligned with the binary classification objectives of the present experiment.

Note*. Previous telecommunications churn studies have used combinations of accuracy, precision, recall, F1 score, ROC AUC, confusion matrix based measures, and, in some cases, specificity and MAE. The present study focuses on accuracy, precision, recall, F1 score, ROC AUC, and confusion matrices because these measures collectively assess overall correctness, positive prediction reliability, churn detection, balance between precision and recall, discrimination, and classification errors.*

The evaluation procedure used in a machine learning study can affect how model performance is interpreted. A single test split provides an estimate of performance on previously unseen observations, but it does not show how consistently a model performs across different subsets of the training data. Cross validation provides an additional way to examine this variation.

Stratified cross validation is particularly relevant to binary churn classification because it maintains approximately the same class proportions across the folds. In a five fold procedure, the training data are divided into five parts. Four parts are used for training while the remaining part is used for validation. This process is repeated until each part has served as the validation set. The resulting scores can then be summarized using their mean and standard deviation (Arlot & Celisse, 2010).

ROC AUC is also widely used in binary classification because it measures discrimination across different classification thresholds. Fawcett (2006) described ROC analysis as a method for evaluating the tradeoff between the true positive rate and false positive rate across thresholds. Precision and recall are particularly relevant when the minority class is important, since they provide information about the quality of positive predictions and the proportion of actual positive cases identified by the model. Saito and Rehmsmeier (2015) further noted that precision recall analysis can be particularly informative when the classes are imbalanced.

Recent churn research uses a mixture of evaluation procedures. Edwine et al. (2022) reported accuracy, recall, precision, AUC, F1 score, and Mean Absolute Error, while Sana et al. (2022) reported AUC, precision, recall, and F measure. The 2026 systematic review also identifies differences in evaluation methods as an important consideration when interpreting findings across the literature.

2.6 Comparison of Existing Studies

The literature shows several recurring patterns. First, telecommunications customer churn remains a well-established classification problem. Second, researchers have compared a wide range of statistical and machine learning models. Third, feature selection and data transformation are frequently used to

prepare the input variables. Fourth, class imbalance is a recurring concern and is addressed through different sampling strategies. Finally, researchers use different combinations of performance measures and validation procedures.

These differences are important because they limit direct comparison of numerical results between separate studies. For example, Edwine et al. (2022) used Mutual Information, controlled ratio undersampling, and extensive parameter optimization, while Sana et al. (2022) combined feature selection with data transformation and Grid Search. Wagh et al. (2024) also examined several machine learning techniques in the telecom churn setting, further demonstrating the continued use of comparative modelling in this area.

https://doi.org/10.64261/hfpwc474

Table 4. Comparative summary of the most relevant studies

Author and year	Dataset	Feature selection	Class balancing	Models	Validation method	Main evaluation measures	Key finding
Hung, Yen, and Wang (2006)	Wireless telecommunications customer data from Taiwan	No specific feature selection method reported	Not emphasized	Decision Tree, Neural Network, and other data mining techniques	Empirical comparative evaluation	Predictive accuracy and propensity to churn	Decision Tree and Neural Network models produced useful churn prediction models using demographic, billing, contract, service, call record, and service change information. (ScienceDirect)
Verbeke et al. (2012)	Eleven real world telecommunications datasets from operators in different countries	Variable selection was examined within the comparative modelling framework	Oversampling examined	Multiple classification techniques	Extensive benchmarking across eleven datasets	Statistical performance measures and profit based measures	A large group of classifiers produced comparable performance, while profit based evaluation provided a different basis for model selection from statistical measures alone. (ScienceDirect)

https://doi.org/10.64261/hfpwc474

Edwine et al. (2022)	Two telecommunications datasets with 3,333 and 7,043 records	Mutual Information	Controlled ratio under sampling	Random Forest, Support Vector Machine, K Nearest Neighbour	Comparative experiments with Grid Search, Random Search, and Genetic Algorithm optimization	Accuracy, recall, precision, ROC AUC, F1 score, MAE	Mutual Information combined with controlled under sampling and model optimization produced strong results, with optimized Random Forest reporting the highest performance in their experimental settings. (DOI)
Sana et al. (2022)	Four publicly available telecommunications churn datasets	Univariate feature selection	No specific resampling method emphasized	K Nearest Neighbour, Naive Bayes, Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, Feed Forward Neural Network, Recurrent	10 fold cross validation	AUC, precision, recall, F measure	Data transformation, feature selection, and hyperparameter optimization affected predictive performance, with Logistic Regression, Random Forest, Feed Forward Neural Network, and Gradient Boosting among the stronger classifiers. (PLOS)

https://doi.org/10.64261/hfpwc474

				Neural Network			
Wagh et al. (2024)	Telecommunications customer churn dataset	Feature selection not central to the reported approach	SMOTE and Edited Nearest Neighbours	Multiple machine learning techniques, including Random Forest	Comparative modelling with resampling	Accuracy, precision, recall, F1 score	Resampling substantially changed the class distribution and the reported Random Forest results, illustrating the effect of class imbalance treatment on churn prediction performance. (DOI)
Mettle, Ayerko, and Appiahene (2026)	IBM Telco Customer Churn dataset; 5,634 preprocessed records	No explicit feature selection; model based feature importance was examined	No synthetic resampling; class distribution preserved	Decision Tree, Random Forest, XGBoost, Logistic Regression	80:20 stratified train test split with 5 fold cross validation on the training data; grid search for tuning	Accuracy, precision, recall, F1 score, ROC AUC	Random Forest produced the strongest reported performance, while the study used a unified comparison of four classical models and retained the original class imbalance. (Wiley Online Library)
Present study	7,032 processed telecommunications customer records	Mutual Information with SelectKBest,	No resampling	Logistic Regression, Decision Tree,	80:20 stratified holdout split; 5 fold	Accuracy, precision, recall, F1 score, ROC	Provides a controlled comparison in which all four

		top 10 features		Random Forest, Gradient Boosting	stratified cross validation on training data	AUC, confusion matrix	models use the same preprocessing, feature selection, data split, and evaluation framework.

Note. *The studies differ substantially in their datasets, feature selection procedures, class balancing strategies, model configurations, and validation protocols. Their reported performance values should therefore not be treated as directly comparable numerical benchmarks. The main purpose of this comparison is to show how the experimental design of the present study relates to previous research*.

https://doi.org/10.64261/hfpwc474

The literature therefore supports the need for careful experimental control when comparing classifiers. A comparison is more informative when the models are exposed to the same observations, processed using the same procedures, and evaluated using the same criteria. Such an approach does not replace broader studies involving different datasets or advanced optimization methods. Instead, it provides a controlled basis for examining how selected classifiers behave under identical experimental conditions.

2.7 Research Gap

The existing literature provides substantial evidence that machine learning can be used for telecommunications customer churn prediction. It also provides extensive comparisons of classification algorithms and several established approaches to feature selection, class imbalance, and model optimization. The recent systematic review by Jain, Khunteta, and Sharma further demonstrates the maturity and breadth of the research area, covering 199 studies published between 2010 and 2024.

The main gap addressed by this study is therefore not the introduction of a new classification algorithm or a new feature selection method. Instead, the study focuses on experimental consistency. Existing studies often combine different datasets, preprocessing procedures, feature selection methods, sampling strategies, optimization techniques, and evaluation protocols. These differences make it difficult to determine how much of an observed performance difference is associated with the classifier itself.

This study addresses that issue through a controlled comparison of Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. All four models use the same dataset, the same training and test split, the same preprocessing framework, the same Mutual Information based feature selection procedure, and the same evaluation measures. Five fold stratified cross validation is applied to the training data, while the holdout test set remains separate for final evaluation. The resulting comparison provides evidence about the relative behaviour of the four models within one consistent experimental setting.

The literature reviewed in this section therefore provides the basis for the methodology used in the present study. The next section describes the dataset, data preparation, preprocessing pipeline, feature selection procedure, model configurations, training process, validation strategy, and evaluation measures used in the experiment.

3. Methodology

3.1 Research Design

This study uses a quantitative experimental design to compare four supervised machine learning models for telecommunications customer churn prediction. The models evaluated are Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Each model was trained and evaluated using the same dataset preparation procedure, preprocessing pipeline, feature selection method, and evaluation criteria.

The experimental design was intended to provide a controlled comparison of the selected classifiers. The

same training and test partitions were used for all four models, and the same preprocessing and feature selection procedures were applied within the modelling pipeline. Model performance was examined in two ways. First, the trained models were evaluated on a separate holdout test set that was not used during model training. Second, five fold stratified cross validation was performed on the training data to examine the consistency of model performance.

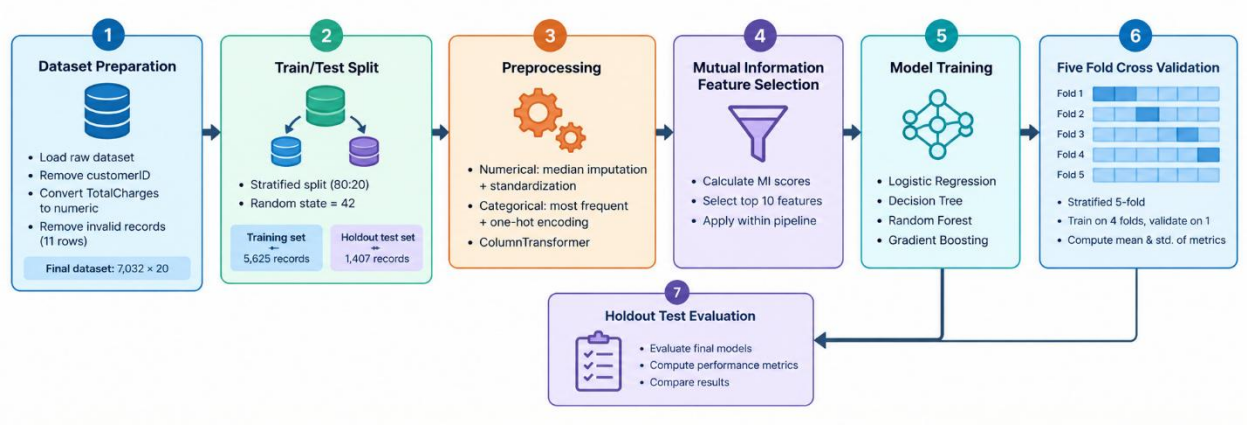


Figure 2: Overall experimental workflow showing dataset preparation → train/test split → preprocessing → Mutual Information feature selection → model training → five fold cross validation → holdout test evaluation

3.2 Dataset

The study uses a telecommunications customer churn dataset containing 7,043 customer records and 21 variables. The dataset contains information about customer demographics, services, account characteristics, contract arrangements, and billing. The target variable, Churn, indicates whether a customer left the service.

The original dataset includes the customer identifier customerID. Because this identifier does not represent a meaningful behavioural or service characteristic, it was removed before model development. The TotalCharges variable required additional processing because some records contained blank or invalid values. The variable was converted to a numeric representation, and 11 records with invalid or missing values were removed.

After these preparation steps, the resulting dataset contained 7,032 observations and 20 variables, including the target variable.

The target distribution consisted of 5,163 customers who did not churn and 1,869 customers who churned. Churn therefore represented approximately 26.6 percent of the processed dataset, while approximately 73.4 percent of customers remained with the provider. The distribution shows a moderate class imbalance, with non churn observations forming the majority class.

Table 5. Dataset characteristics

Item	Value
Original rows	7,043

https://doi.org/10.64261/hfpwc474

Original columns	21
Removed identifier	customerID
Invalid TotalCharges records removed	11
Final rows	7,032
Final columns	20
Non churn records	5,163
Churn records	1,869

Note. The dataset contained 7,043 customer records and 21 columns before preparation. The customerID identifier was removed, and 11 records with invalid or blank TotalCharges values were excluded. The resulting dataset contained 7,032 records and 20 modelling columns, with 5,163 non churn records and 1,869 churn records.

3.3 Data Preparation

Data preparation was performed before model training. The customer identifier was removed because it does not provide predictive information about customer behaviour. TotalCharges was converted to a numeric variable so that it could be processed as a numerical feature. Records that could not be converted into valid numerical values were excluded.

The target variable was separated from the predictor variables before the train and test datasets were created. The remaining variables consisted of numerical and categorical attributes. Numerical variables included measures such as tenure and charges, while categorical variables represented customer characteristics, subscribed services, contract arrangements, payment methods, and related account information.

The processed data were divided into training and holdout test sets using an 80:20 stratified split with a random state of 42. Stratification was used to preserve the proportion of churn and non churn observations in both partitions. The resulting training set contained 5,625 observations, while the holdout test set contained 1,407 observations.

The test set was kept separate from the model development process. It was not used to fit preprocessing transformations, select features, train the models, or perform cross validation. This separation provides an independent basis for evaluating the final trained models.

https://doi.org/10.64261/hfpwc474

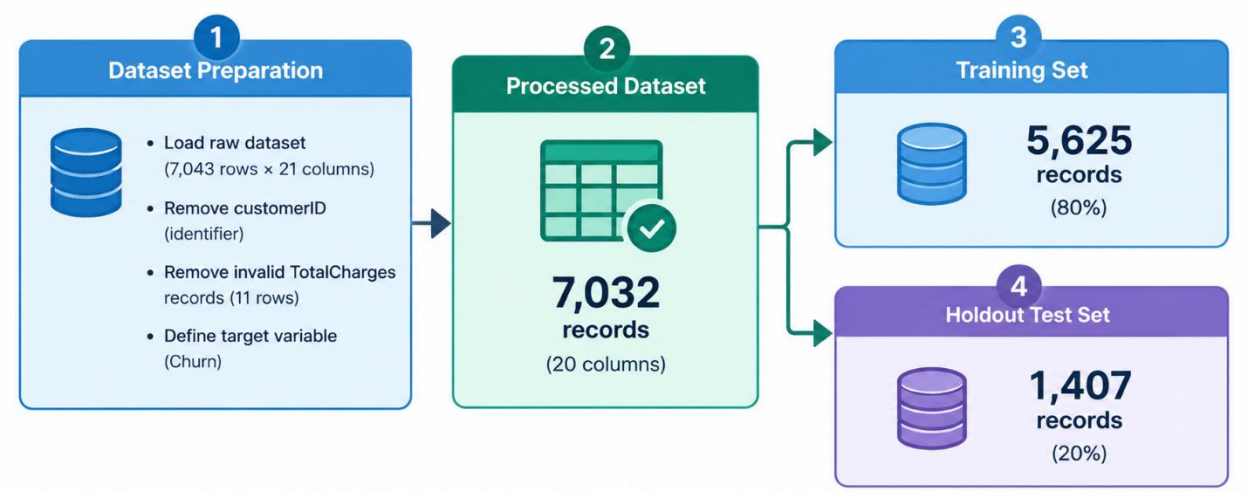


Figure 3. Dataset preparation and partitioning diagram showing 7,032 processed records divided into 5,625 training records and 1,407 holdout test records.

Figure 3 here: Dataset preparation and partitioning diagram showing 7,032 processed records divided into 5,625 training records and 1,407 holdout test records

3.4 Data Preprocessing

The dataset contains both numerical and categorical variables, so separate preprocessing procedures were used for the two variable types. The preprocessing operations were implemented using a scikit learn pipeline so that the transformations were applied consistently during model training and evaluation.

For numerical variables, missing values were handled using median imputation. The resulting values were then standardized using StandardScaler. For categorical variables, missing values were handled using the most frequent category. The categorical variables were subsequently converted into numerical representations using OneHotEncoder with handle_unknown="ignore".

The preprocessing operations were combined using a ColumnTransformer. This allowed the numerical and categorical variables to follow their respective preprocessing procedures while remaining part of a single modelling workflow.

Importantly, the preprocessing transformations were fitted using the training data rather than the complete dataset. During evaluation, the transformations learned from the training data were applied to validation and test observations. This prevents information from the holdout test set from influencing the fitted preprocessing parameters.

3.5 Feature Selection

After preprocessing, Mutual Information was used to select the most informative features for the classification task. Mutual Information measures the dependence between an input variable and the target variable and can be used to identify features that contain useful information about the outcome.

The study used SelectKBest with Mutual Information as the scoring method and selected the 10 highest scoring features. The same feature selection procedure was used for all four models.

Feature selection was incorporated into the modelling pipeline rather than being performed separately on the complete dataset before the train and test split. This design ensures that the feature selection step is fitted using the appropriate training data during model development. It also prevents information from the holdout test observations from influencing the selected feature set.

The purpose of feature selection in this study was to provide a common reduced feature space for the four classifiers. The experiment does not include a separate comparison between models trained with and without feature selection. Therefore, the study does not attempt to estimate the independent effect of Mutual Information on predictive performance.

Table 6. Selected features after Mutual Information ranking

Rank	Feature	Mutual Information score
1	Contract: Month to month	0.092029
2	tenure	0.074148
3	Contract: Two year	0.060715
4	PaymentMethod: Electronic check	0.058354
5	OnlineSecurity: No	0.057884
6	TechSupport: No	0.052659
7	TotalCharges	0.044276
8	MonthlyCharges	0.043477
9	InternetService: Fiber optic	0.036158
10	InternetService: No	0.036089

Note. *Mutual Information scores were calculated on the training portion of the dataset after the preprocessing transformations. Categorical variables were one hot encoded before feature scoring. The ten highest scoring transformed features were retained using SelectKBest with k = 10. The scores represent the amount of information each transformed feature provides about the churn target under the applied estimation procedure. They should not be interpreted as causal effects or as evidence that Mutual Information itself improved model performance.*

Short interpretation for Section 4.1

The Mutual Information analysis identified Contract: Month to month as the highest ranked feature, with a score of 0.096518. Tenure was ranked second at 0.074887. Other highly ranked transformed features included TechSupport: No, Contract: Two year, OnlineSecurity: No, and PaymentMethod: Electronic check. MonthlyCharges, TotalCharges, InternetService: Fiber optic, and OnlineBackup: No also appeared among the ten selected features. The results indicate that contract characteristics, customer tenure, service related variables, payment method, and charges contained measurable information about the churn target within the training data.

3.6 Machine Learning Models

Four supervised classification algorithms were selected for the comparative experiment.

3.6.1 Logistic Regression

Logistic Regression was included as a statistical classification baseline. It estimates the probability of a binary outcome using a logistic function. Its relatively simple structure makes it useful for

comparison with more complex ensemble methods.

3.6.2 Decision Tree

The Decision Tree classifier represents classification decisions through a sequence of branching rules. It can model nonlinear relationships and interactions between variables without requiring the same linear assumptions as Logistic Regression.

3.6.3 Random Forest

Random Forest is an ensemble method that combines predictions from multiple decision trees. Breiman (2001) introduced Random Forest as a method that uses multiple randomized trees to improve classification and regression performance. Its inclusion allows the study to examine whether combining multiple trees provides an advantage over a single Decision Tree.

3.6.4 Gradient Boosting

Gradient Boosting was included as a sequential ensemble method. The approach builds a series of weak prediction models, with later models focusing on reducing the errors made by earlier models. Friedman (2001) established Gradient Boosting as a general statistical learning approach based on this principle.

The four algorithms were trained using the same processed training data and selected features. This common setup allows their predictive performance to be compared without changing the data preparation procedure between models.

3.7 Model Configuration

The final Gradient Boosting model used in the experiment was configured with 100 estimators, a learning rate of 0.1, a maximum tree depth of 3, and a random state of 42. These settings define the final Gradient Boosting configuration used for the reported results.

The other classifiers were evaluated within the same experimental framework. Model selection was based on the observed performance of the evaluated classifiers rather than on an assumption that a particular algorithm would perform best before testing.

Table 7. Model configurations

Model	Main parameters	Random state	Number of estimators
Logistic Regression	max_iter = 1000; pipeline preprocessing	42	Not applicable
Decision Tree	Default configuration; pipeline preprocessing	42	Not applicable
Random Forest	n_estimators = 200; pipeline preprocessing	42	200
Gradient Boosting	n_estimators = 100; learning_rate = 0.1; max_depth = 3; pipeline preprocessing	42	100

Note. *All four models were trained using the same preprocessing and Mutual Information feature selection pipeline.*

https://doi.org/10.64261/hfpwc474

The random state was fixed at 42 where applicable to support reproducibility. The reported Gradient Boosting parameters correspond to the final configuration used in the experiment.

3.8 Training and Validation Procedure

The training procedure was designed to keep model development separate from final evaluation. The training partition contained 5,625 observations and was used for preprocessing, feature selection, model fitting, and cross validation. The holdout test partition contained 1,407 observations and remained separate until final evaluation.

Five fold stratified cross validation was used to examine model performance on the training data. The training set was divided into five folds while preserving the class distribution as closely as possible within each fold. Each model was trained on four folds and evaluated on the remaining fold. The process was repeated until every fold had been used as the validation fold.

The cross validation procedure used shuffling with a random state of 42. Performance was summarized using the mean and standard deviation across the five folds. The measures recorded included accuracy, weighted precision, weighted recall, weighted F1 score, and ROC AUC.

Because preprocessing and feature selection were part of the modelling pipeline, the transformations and feature selection procedure were fitted within the appropriate training portion during model development. This prevents validation information from being used to fit the corresponding transformations.

Cross validation was performed only on the training partition. The holdout test set was not included in the five folds.

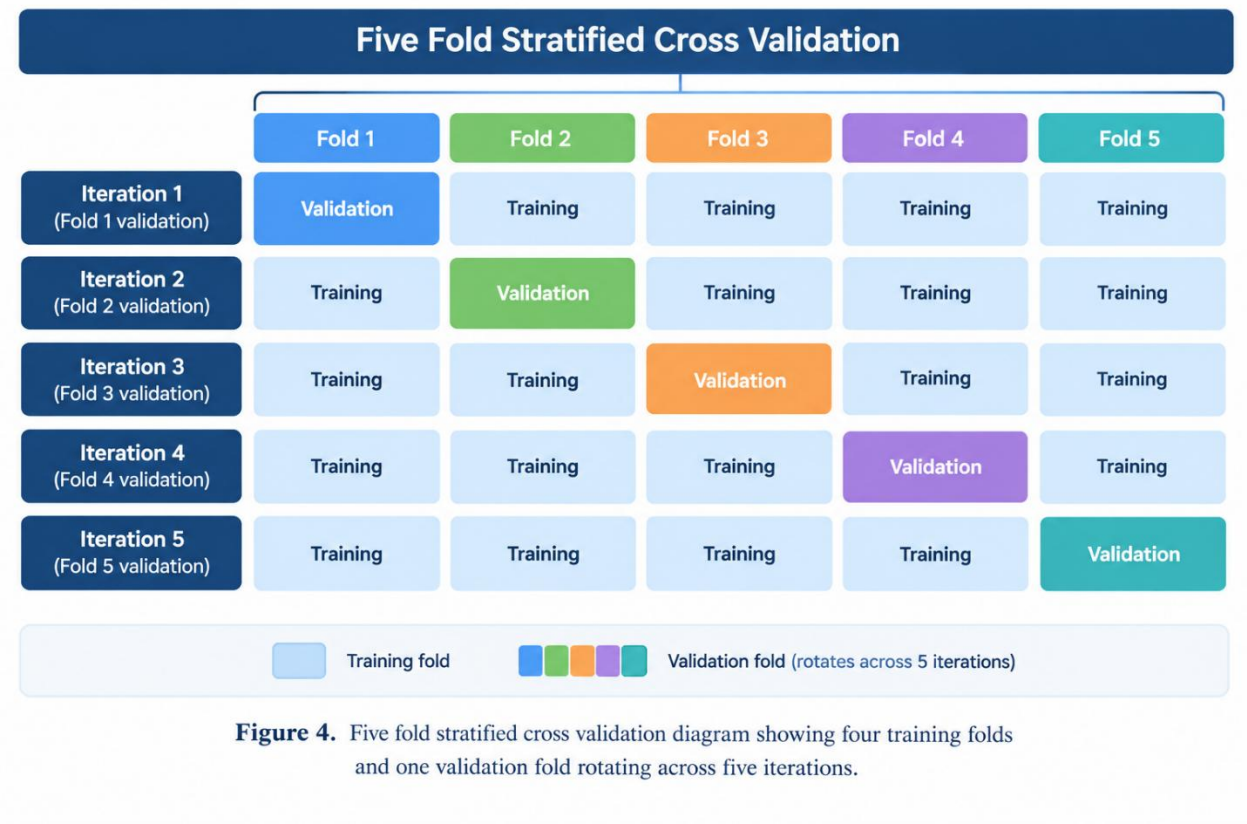


Figure 4. Five fold stratified cross validation diagram showing four training folds and one validation fold rotating across five iterations.

https://doi.org/10.64261/hfpwc474

Figure 4 here: Five fold stratified cross validation diagram showing four training folds and one validation fold rotating across five iterations

3.9 Holdout Test Evaluation

After model development, the classifiers were evaluated using the separate holdout test set. This dataset contained 1,407 observations that were not used to fit the models or determine the preprocessing transformations and selected features.

The holdout evaluation provides the main basis for comparing the final models under unseen data. Accuracy, precision, recall, F1 score, and ROC AUC were used to describe predictive performance. Confusion matrices were also examined to show the numbers of correctly and incorrectly classified churn and non churn observations.

The use of several evaluation measures is important because accuracy alone may not fully describe performance when the target classes are not evenly distributed. Precision indicates the proportion of predicted churn cases that were actually churn cases, while recall indicates the proportion of actual churn cases that were correctly identified. F1 score combines precision and recall into a single measure. ROC AUC measures the ability of the model to distinguish between the two classes across classification thresholds.

3.10 Performance Measures

The following measures were used throughout the evaluation.

Accuracy represents the proportion of all test observations that were correctly classified:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TPTP represents true positives, TNTN represents true negatives, FPFP represents false positives, and FNFN represents false negatives.

Precision measures the proportion of observations predicted as churn that were actually churn cases:

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

Recall measures the proportion of actual churn cases that were correctly identified:

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

F1 score provides a combined measure of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

ROC AUC measures the model's ability to distinguish between churn and non churn observations across different classification thresholds.

Confusion matrices were also used to examine the distribution of correct and incorrect predictions for each class.

3.11 Reproducibility and Implementation

The experiment was implemented in Python using commonly used machine learning and data analysis libraries.

Pandas was used for data manipulation, NumPy for numerical operations, and scikit learn 1.9.1 for preprocessing, feature selection, model training, cross validation, and evaluation.

A fixed random state of 42 was used for the main data split, Mutual Information feature selection, and stochastic modelling procedures. The final Gradient Boosting model was saved as models/model.joblib for subsequent use by the project application.

The implementation included automated tests covering the data processing, machine learning, and application components. The final repository test suite reached 130 passing tests. The test workflow completed successfully after the reproducibility changes.

3.12 Methodological Summary

The complete experimental procedure can be summarized as follows:

1. Load the original telecommunications churn dataset.
2. Remove the customer identifier.
3. Convert TotalCharges to a numeric variable.
4. Remove the 11 records containing invalid TotalCharges values.
5. Separate the target variable from the predictor variables.
6. Divide the processed data into stratified training and holdout test sets using an 80:20 ratio and random state 42.
7. Apply numerical and categorical preprocessing through a ColumnTransformer.
8. Apply Mutual Information based SelectKBest feature selection with random state 42 and retain the 10 highest scoring features.
9. Train Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting classifiers using the common pipeline.
10. Evaluate each model using five fold stratified cross validation on the training data.
11. Evaluate the final models on the separate holdout test set.
12. Compare accuracy, precision, recall, F1 score, ROC AUC, and confusion matrices.
13. Identify the strongest performing model within the experimental conditions.

https://doi.org/10.64261/hfpwc474

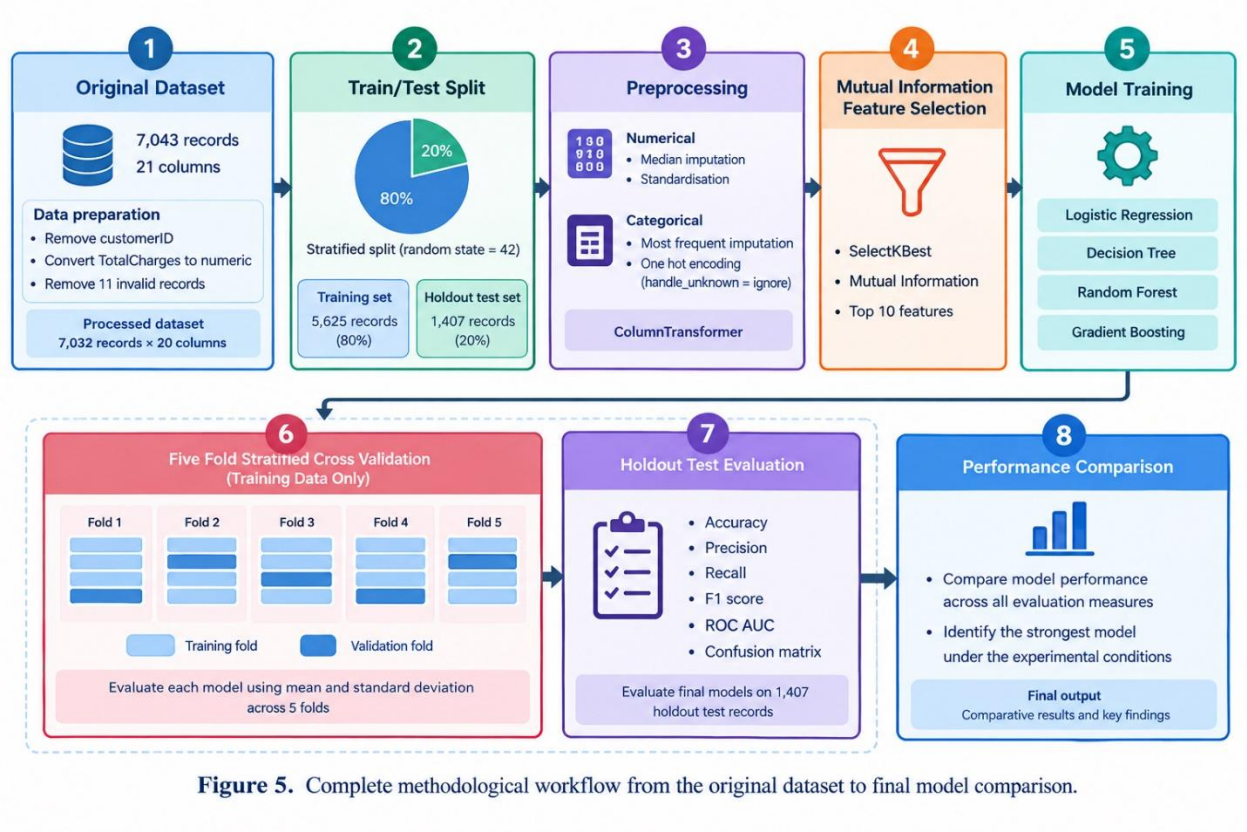


Figure 5. Complete methodological workflow from the original dataset to final model comparison.

Figure 5: Complete methodological workflow from the original dataset to final model comparison

The methodology was designed to reduce differences between the experimental conditions of the four classifiers. The same dataset, data partition, preprocessing approach, feature selection method, validation strategy, and evaluation measures were used throughout the comparison. This provides the basis for the results presented in the next section.

4. Results

4.1 Dataset and Feature Selection Results

The raw dataset corresponds to the IBM Telco Customer Churn sample dataset, which contains 7,043 records and 21 columns before preparation. IBM publishes the sample in its customer churn code pattern repository (IBM, n.d.). The version used in this study was stored locally as Churn_Dataset.csv and processed to remove the customer identifier and the 11 records with invalid TotalCharges values. The IBM source provides the provenance for the underlying dataset used in the experiment.

After the initial data preparation, the dataset contained 7,032 customer records and 20 variables, including the churn target. The dataset contained 5,163 customers who did not churn and 1,869 customers who churned. Thus, non churn customers represented approximately 73.4 percent of the observations, while churn customers represented approximately 26.6 percent.

The processed dataset was divided into 5,625 training observations and 1,407 holdout test observations using stratified sampling. The class distribution was preserved during the split so that the training and test sets maintained similar proportions of churn and non churn observations.

Mutual Information based feature selection was then applied within the modelling pipeline using a fixed random state of 42. The SelectKBest procedure retained the 10 highest scoring features for model development. The selected features were used consistently across all four classifiers.

Table 8: Dataset preparation and partition results

Dataset stage	Observations	Variables
Original dataset	7,043	21
After data preparation	7,032	20
Training set	5,625	20
Holdout test set	1,407	20

Table 9: Class distribution in the processed dataset.

Churn status	Number of customers	Percentage
No churn	5,163	73.4%
Churn	1,869	26.6%
Total	7,032	100%

4.2 Holdout Test Performance

The four classifiers were evaluated using the same holdout test set containing 1,407 observations. The results showed differences in predictive performance across the models.

Logistic Regression achieved an accuracy of 78.82 percent and a ROC AUC of 83.39 percent. Decision Tree produced an accuracy of 72.00 percent and a ROC AUC of 65.01 percent. Random Forest achieved 77.40 percent accuracy and a ROC AUC of 79.30 percent. Gradient Boosting achieved 78.61 percent accuracy and the highest ROC AUC among the four models at 83.60 percent.

The difference between Gradient Boosting and Logistic Regression was small. Gradient Boosting had a slightly lower accuracy than Logistic Regression, while its ROC AUC was slightly higher. The ROC AUC difference was 0.21 percentage points.

Table 10: Holdout test performance of the four classifiers

Model	Accuracy	Precision	Recall	F1 score	ROC AUC
Logistic Regression	78.82%	78.06%	78.82%	78.33%	83.39%
Decision Tree	72.00%	72.29%	72.00%	72.14%	65.01%
Random Forest	77.40%	76.50%	77.40%	76.83%	79.30%
Gradient Boosting	78.61%	77.66%	78.61%	77.96%	83.60%

Among the four models, Gradient Boosting produced the highest numerical ROC AUC under the experimental conditions used in this study, while Logistic Regression produced the highest overall accuracy. The differences between the two models were small. Decision Tree recorded the lowest accuracy and ROC AUC.

https://doi.org/10.64261/hfpwc474

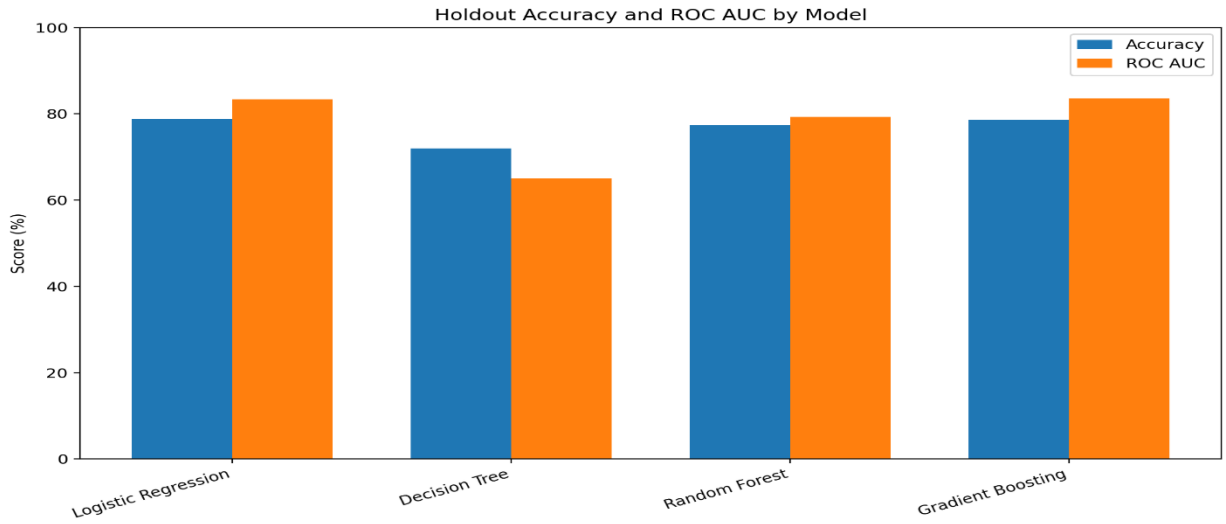


Figure 6. Comparison of accuracy and ROC AUC across Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting on the holdout test set.

4.3 Detailed Gradient Boosting Performance

Because Gradient Boosting produced the highest ROC AUC in the holdout comparison, its class level results were examined in greater detail.

The model achieved an overall accuracy of 78.61 percent, weighted precision of 77.66 percent, weighted recall of 78.61 percent, and weighted F1 score of 77.96 percent. Its ROC AUC was 83.60 percent.

The confusion matrix contained 913 correctly classified non churn customers and 193 correctly classified churn customers. The model incorrectly classified 120 non churn customers as churn and 181 churn customers as non churn.

Table 11: Detailed Gradient Boosting classification performance

Measure	Value
Accuracy	78.61%
Precision (weighted)	77.66%
Recall (weighted)	78.61%
F1 score (weighted)	77.96%
ROC AUC	83.60%

https://doi.org/10.64261/hfpwc474

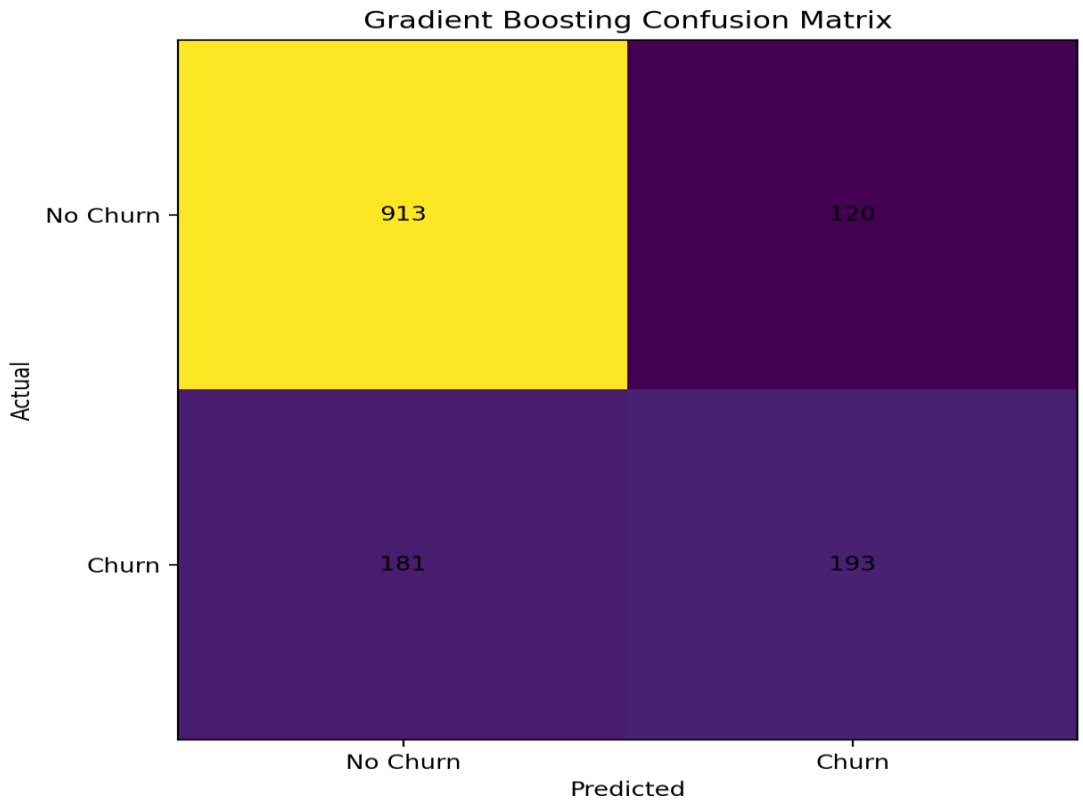


Figure 7: Confusion matrix for the Gradient Boosting model.

The class level results provide additional information about the distribution of these predictions. For the non churn class, precision was 83.46 percent, recall was 88.38 percent, and F1 score was 85.85 percent. For the churn class, precision was 61.66 percent, recall was 51.60 percent, and F1 score was 56.19 percent.

Table 12: Class level classification report for Gradient Boosting

Class	Precision	Recall	F1 score	Support
No churn	83.46%	88.38%	85.85%	1,033
Churn	61.66%	51.60%	56.19%	374

4.4 Cross Validation Results

Five fold stratified cross validation was performed using the training data. The holdout test set was not included in the cross validation procedure.

For Gradient Boosting, the mean ROC AUC across the five folds was approximately 84.48 percent, with a standard deviation of approximately 0.65 percentage points. This indicates that the model produced similar ROC AUC values across the five validation folds.

Table 13. Five fold cross validation results for all four models

https://doi.org/10.64261/hfpwc474

Model	Accuracy	Precision	Recall	F1 score	ROC AUC
Logistic Regression	79.86% ± 0.67%	79.10% ± 0.76%	79.86% ± 0.67%	79.33% ± 0.74%	84.29% ± 0.52%
Decision Tree	73.48% ± 1.24%	73.65% ± 1.20%	73.48% ± 1.24%	73.56% ± 1.21%	66.83% ± 1.75%
Random Forest	77.51% ± 0.48%	76.51% ± 0.44%	77.51% ± 0.48%	76.83% ± 0.44%	80.99% ± 0.85%
Gradient Boosting	79.63% ± 0.34%	78.62% ± 0.31%	79.63% ± 0.34%	78.79% ± 0.22%	84.48% ± 0.65%

Note. Values are the mean ± standard deviation across five stratified folds using the training portion of the dataset. The holdout test set was not used during cross validation. Gradient Boosting and Logistic Regression had similar cross validation ROC AUC values of 84.48% ± 0.65% and 84.29% ± 0.52%, respectively.

The standard deviation of approximately 0.65 percentage points for Gradient Boosting ROC AUC indicates limited variation in this measure across the five validation folds within this dataset.

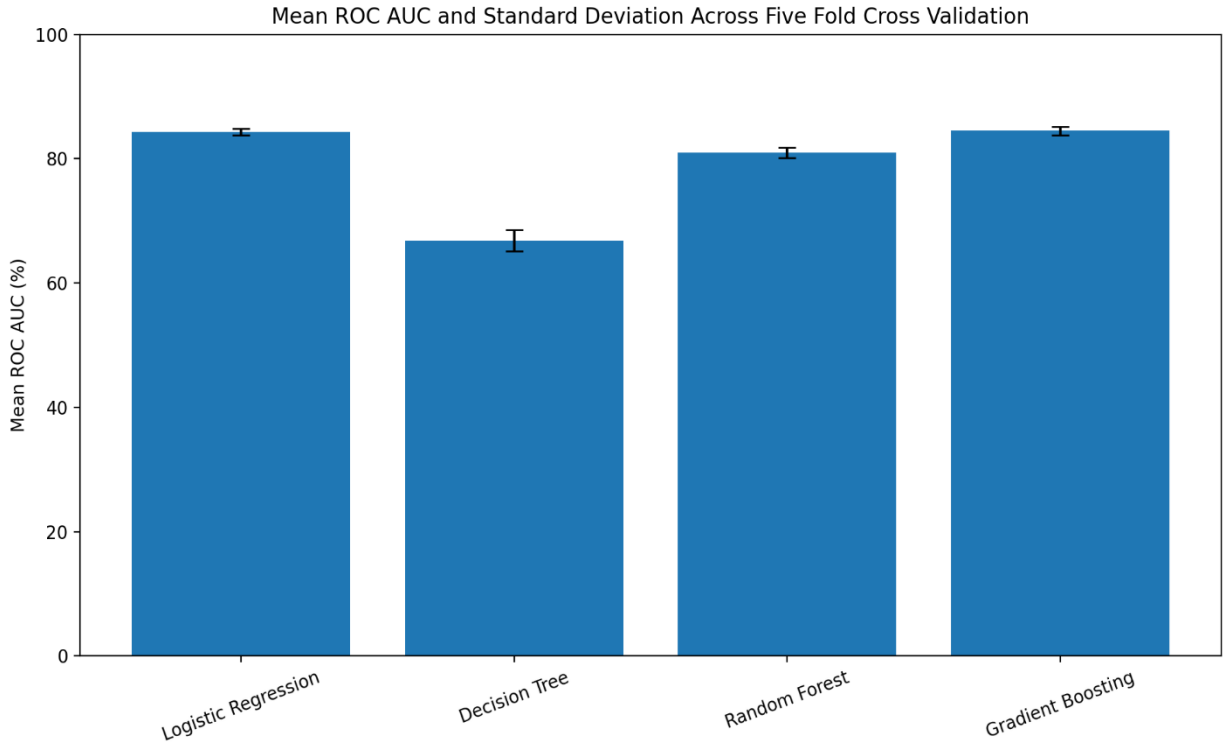


Figure 8. Mean ROC AUC with standard deviation across five fold cross validation.

4.5 ROC Analysis

ROC AUC was used to examine the ability of the four classifiers to distinguish between churn and non churn customers across classification thresholds. The final holdout results showed ROC AUC values of 83.39 percent for Logistic Regression, 65.01 percent for Decision Tree, 79.30 percent for Random Forest, and 83.60 percent for Gradient Boosting.

https://doi.org/10.64261/hfpwc474

Gradient Boosting therefore produced the highest numerical ROC AUC in the holdout evaluation, although its value was only 0.21 percentage points higher than that of Logistic Regression.

The ROC analysis also shows a substantial difference between the two tree based models and the individual Decision Tree. Random Forest produced a higher ROC AUC than Decision Tree, while Gradient Boosting produced the highest numerical ROC AUC among the four evaluated classifiers.

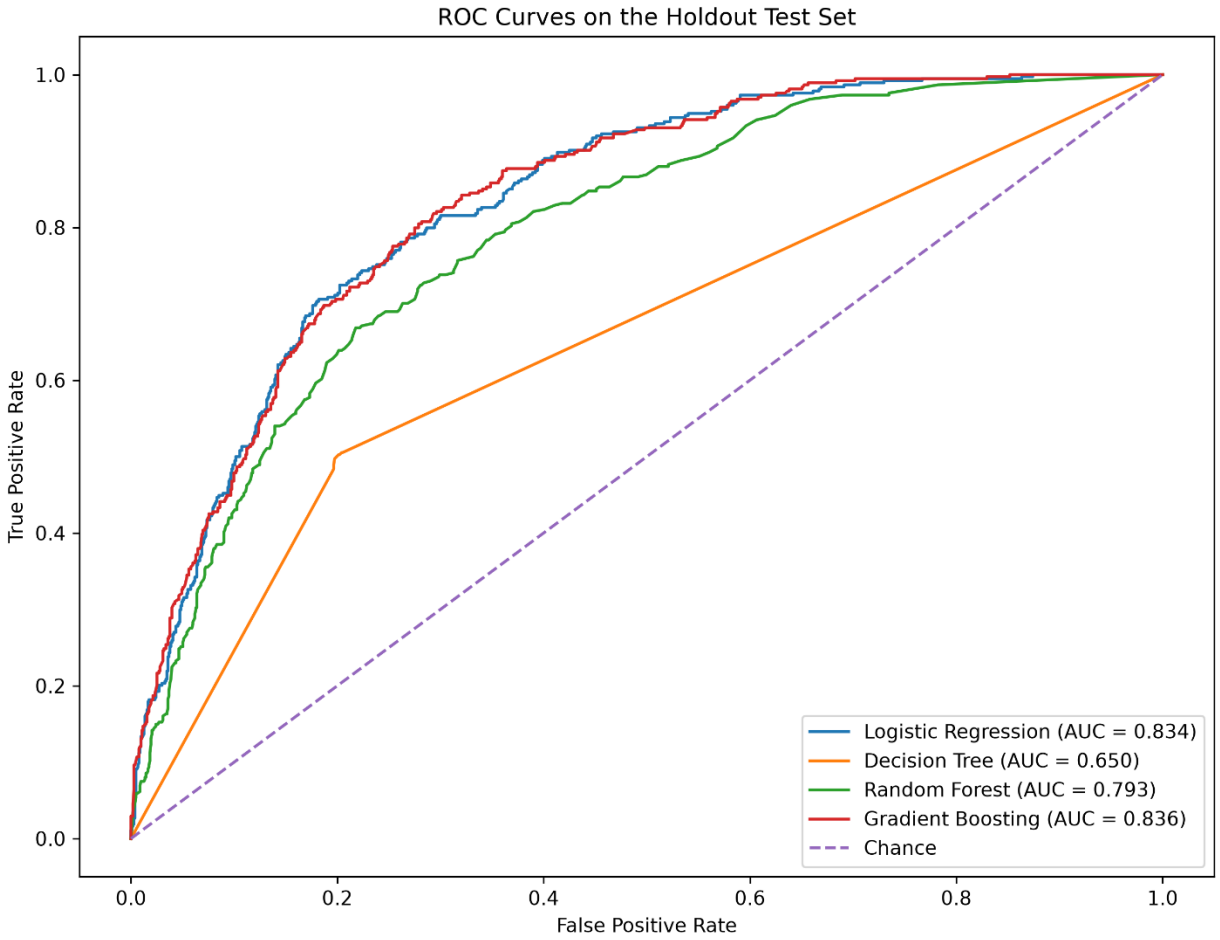


Figure 9. ROC curves for Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting on the holdout test set.

4.6 Summary of Results

The experimental results show that the four classifiers did not perform equally under the same preprocessing and feature selection conditions. Gradient Boosting achieved the highest numerical holdout ROC AUC at 83.60 percent, while Logistic Regression achieved the highest holdout accuracy at 78.82 percent. The difference between these two models was small for both measures.

Random Forest produced an accuracy of 77.40 percent and a ROC AUC of 79.30 percent. Decision Tree produced the lowest performance among the four classifiers, with an accuracy of 72.00 percent and a ROC AUC of 65.01 percent.

The detailed Gradient Boosting results also show a difference between performance on the majority and minority

classes. The model correctly identified 51.60 percent of actual churn cases, compared with 88.38 percent of non churn cases. Its churn class F1 score was 56.19 percent, while the corresponding non churn F1 score was 85.85 percent.

The five fold cross validation results provide an additional assessment of model stability. Gradient Boosting achieved a mean ROC AUC of approximately 84.48 percent with a standard deviation of approximately 0.65 percentage points across the five folds.

These results provide the empirical basis for the discussion in the following section. The next section considers the findings in relation to previous research, examines the differences between the four models, and discusses the implications and limitations of the experimental results.

5. Discussion

5.1 Interpretation of the Model Comparison

The results show that the four classifiers produced different levels of predictive performance despite being trained and evaluated under the same experimental conditions. Gradient Boosting recorded the highest numerical ROC AUC on the holdout test set at 83.60 percent, while Logistic Regression recorded the highest accuracy at 78.82 percent. The difference between the two models was small. Gradient Boosting achieved an accuracy of 78.61 percent and Logistic Regression achieved 78.82 percent, a difference of 0.21 percentage points.

This result shows that the model with the highest numerical value depends partly on the evaluation measure being considered. If ROC AUC is used as the measure of discrimination, Gradient Boosting produced the highest numerical value. If overall accuracy is considered, Logistic Regression produced a higher value. The observed ROC AUC difference is small, so it should not be interpreted as evidence of statistically significant superiority.

The similar performance of the two models also suggests that the relationship between the available customer characteristics and churn can be captured reasonably well by a relatively simple classification method. Logistic Regression provides a useful baseline because its structure is less complex than the ensemble methods examined in this study. Its performance being close to that of Gradient Boosting indicates that additional model complexity did not produce a large improvement in the measured holdout performance under the conditions of this experiment.

5.2 Performance of the Tree Based Models

Decision Tree produced the lowest performance among the four classifiers, with an accuracy of 72.00 percent and a ROC AUC of 65.01 percent. Random Forest produced higher values, with an accuracy of 77.40 percent and a ROC AUC of 79.30 percent. The observed difference is specific to the dataset, feature selection procedure, model configurations, and evaluation protocol used here.

The difference between these two models is consistent with the general principle behind ensemble tree methods. A single Decision Tree makes predictions through one sequence of learned decision rules. Random Forest combines predictions from multiple trees and can therefore reduce some of the instability associated with relying on one tree. Breiman (2001) developed Random Forest as an ensemble approach designed to improve predictive performance through the combination of multiple randomized trees.

https://doi.org/10.64261/hfpwc474

The results from this study should not, however, be interpreted as evidence that Random Forest will always outperform a single Decision Tree for customer churn prediction. The observed difference is specific to the dataset, feature selection procedure, model configurations, and evaluation protocol used here.

5.3 Gradient Boosting and Logistic Regression

The close results between Gradient Boosting and Logistic Regression are one of the main findings of the experiment. Gradient Boosting achieved a ROC AUC of 83.60 percent compared with 83.39 percent for Logistic Regression. The difference was 0.21 percentage points. Accuracy followed a similar pattern, with Logistic Regression at 78.82 percent and Gradient Boosting at 78.61 percent.

The results suggest that increasing model complexity did not produce a substantial improvement in this experiment. Gradient Boosting can represent nonlinear relationships through an ensemble of sequential decision trees, while Logistic Regression models the relationship between the predictors and the outcome using a linear decision function. The similar results indicate that the additional flexibility of Gradient Boosting provided only a small improvement in discrimination on the holdout data.

This finding is relevant when considering model selection in practical settings. A more complex model is not automatically preferable simply because it can represent more complicated relationships. Where predictive performance is similar, factors such as interpretability, computational requirements, maintenance, and the intended use of the model may also become relevant. The present study does not evaluate those operational factors, so no conclusion is made about which model would be preferable in a particular business environment.

5.4 Churn Class Performance

Although the overall results were reasonably consistent across several measures, the class level results reveal an important limitation. The Gradient Boosting model correctly identified 51.60 percent of customers who actually churned. Its precision for the churn class was 61.66 percent and its F1 score was 56.19 percent.

The corresponding results for the non churn class were considerably higher. Non churn recall was 88.38 percent and the F1 score was 85.85 percent. The confusion matrix contained 181 churn customers that were incorrectly classified as non churn. In contrast, 120 non churn customers were incorrectly classified as churn.

This difference is important because overall accuracy does not show how the model behaves for each class. The dataset contains substantially more non churn customers than churn customers. A model can therefore achieve relatively high overall accuracy while still missing a considerable proportion of customers who actually leave.

The results indicate that the model is better at identifying customers who remain with the provider than customers who churn. This does not invalidate the overall model comparison, but it places an important qualification on the results. If the primary objective of a future application were to identify as many potential churners as possible, recall for the churn class would require particular attention.

5.5 Relationship to Previous Research

The findings are consistent with the broader literature in showing that several machine learning methods can be used for telecom customer churn prediction, but their performance varies according to the experimental conditions. Previous studies have compared a wide range of classifiers and have

used different combinations of preprocessing, feature selection, class balancing, and parameter optimization.

Sana et al. (2022) examined telecom churn prediction using data transformation and feature selection and evaluated several models using AUC, precision, recall, and F measure. Their study demonstrates the importance of evaluating churn models using several performance measures rather than relying on accuracy alone. The present study follows a similar principle by reporting accuracy, precision, recall, F1 score, and ROC AUC. However, the experimental procedures differ, particularly in the treatment of class imbalance and model optimization.

Edwine et al. (2022) also used Mutual Information for feature selection in a comparative telecom churn study. Their methodology included controlled ratio undersampling and parameter optimization before evaluating several classifiers. The present study uses Mutual Information as a common feature selection procedure but does not use undersampling or a separate experiment comparing models with and without feature selection. Consequently, the findings should not be interpreted as evidence that Mutual Information itself improved predictive performance. The purpose of using it here was to maintain a consistent feature selection procedure across the four models.

The results also fit within the broader development of churn prediction research. Earlier studies established the use of data mining and statistical learning methods for telecom churn, while later work has expanded the range of models, feature selection methods, sampling procedures, and optimization techniques. Recent systematic research shows that the field now contains a large body of studies with considerable variation in datasets and experimental protocols. This variation reinforces the value of controlled comparisons when the purpose is to examine differences between selected algorithms.

5.6 Cross Validation and Model Stability

The five fold stratified cross validation results provide additional evidence about model performance. Gradient Boosting achieved a mean ROC AUC of approximately 84.48 percent with a standard deviation of approximately 0.65 percentage points across the five folds.

The relatively small standard deviation indicates that Gradient Boosting produced similar ROC AUC values across the validation folds in this experiment. This suggests that its measured discrimination was not strongly dependent on one particular training and validation partition.

The cross validation results should nevertheless be interpreted alongside the holdout test results. Cross validation was performed only on the training data, while the holdout test set was reserved for final evaluation. The two procedures therefore answer slightly different questions. Cross validation examines performance consistency during model development, while the holdout evaluation provides the final comparison using observations that were kept outside the model development process.

The difference between the Gradient Boosting mean cross validation ROC AUC of approximately 84.48 percent and its holdout ROC AUC of 83.60 percent is not, by itself, evidence of a major change in model behaviour. The values come from different evaluation procedures and should not be treated as directly interchangeable estimates.

5.7 Implications for Customer Churn Prediction

The findings show that a controlled model comparison can produce useful information even when the differences

between some models are small. In this study, Logistic Regression and Gradient Boosting produced almost identical holdout accuracy and ROC AUC values, while Random Forest produced lower values and Decision Tree produced the lowest values among the four classifiers.

The results also demonstrate why churn prediction should not be assessed using a single metric. The Gradient Boosting model had a ROC AUC of 83.60 percent and an overall accuracy of 78.61 percent, but it identified only 51.60 percent of the actual churn cases. A business application that focuses heavily on retaining customers at risk of leaving may therefore require a different evaluation emphasis from an application concerned mainly with overall classification accuracy.

Threshold selection could be one way to address this issue in future work. Changing the probability threshold can alter the balance between identifying more churn customers and producing more false positive predictions. However, threshold selection was not part of the experimental comparison reported in this study, and the current results should therefore be interpreted using the default classification procedure used during evaluation.

5.8 Methodological Contribution

The main contribution of the experiment is methodological consistency rather than the introduction of a new machine learning algorithm. Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting are established classification methods, and Mutual Information is an established feature selection technique.

The study instead evaluates these models within the same experimental framework. The same dataset was used for every model. The same 80:20 stratified partition was applied, the same preprocessing procedure was used, and the same Mutual Information based feature selection approach was applied. The models were evaluated using the same holdout test set and five fold stratified cross validation on the training data.

This design reduces one source of variation that occurs when results from separate studies are compared. Differences in reported performance can arise from the dataset, preprocessing, sampling strategy, feature selection, parameter optimization, or evaluation protocol. By keeping these factors consistent within the experiment, the comparison provides clearer evidence about the relative behaviour of the four selected classifiers under the same conditions.

5.9 Limitations

Several limitations should be considered when interpreting the findings.

First, the study uses a single telecommunications customer churn dataset. The results may therefore not represent other telecommunications markets, providers, customer populations, or datasets with different characteristics.

Second, the dataset contains a moderate class imbalance, with churn customers representing approximately 26.6 percent of the processed observations. No oversampling or undersampling procedure was applied. The results therefore reflect model performance under the original class distribution used in the experiment.

Third, the study uses Mutual Information feature selection as a common component of the modelling pipeline but does not include an ablation experiment comparing performance with and without feature selection. It is therefore not possible to determine how much of the observed performance is attributable specifically to the feature selection procedure.

https://doi.org/10.64261/hfpwc474

Fourth, the model comparison does not provide evidence about business or financial outcomes. Predicting churn accurately does not by itself demonstrate that a retention intervention would increase customer retention or produce financial benefits.

Fifth, the Gradient Boosting model achieved a churn recall of only 52 percent on the holdout test set. This means that a substantial proportion of actual churn cases were not identified. The result limits the extent to which the current model should be interpreted as a complete solution for identifying customers at risk of leaving.

Finally, the study evaluates four selected classifiers rather than the full range of approaches used in current churn research. More extensive hyperparameter optimization, alternative sampling strategies, additional algorithms, external validation datasets, and threshold optimization could produce different results.

5.10 Discussion Summary

The experiment demonstrates that the four selected classifiers behaved differently under a common preprocessing and feature selection framework. Gradient Boosting produced the highest numerical holdout ROC AUC at 83.60 percent, while Logistic Regression produced the highest accuracy at 78.82 percent. The difference between the two models was small, indicating that the additional complexity of Gradient Boosting did not produce a large improvement under the conditions of this experiment.

Random Forest performed below these two models but substantially above the single Decision Tree. The class level results for Gradient Boosting also showed a clear difference between non churn and churn prediction. The model identified 52 percent of actual churn cases, compared with 88 percent of non churn cases.

The findings support the value of controlled model comparison while also showing the limits of relying on overall accuracy or ROC AUC alone. The results should be understood as evidence from one dataset and one experimental configuration rather than as a general ranking of machine learning algorithms for telecommunications churn prediction.

The next section presents the conclusion of the study, summarizes the main findings, states the practical and methodological implications, and identifies directions for further investigation.

6. Conclusion

This study examined the performance of four supervised machine learning models for telecommunications customer churn prediction. Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting were evaluated using the same dataset, preprocessing procedure, Mutual Information based feature selection method, training and test partition, and evaluation framework. The purpose was to provide a controlled comparison of commonly used classifiers rather than to propose a new prediction algorithm.

The results showed clear differences between the four models. Gradient Boosting achieved the highest numerical ROC AUC on the holdout test set at 83.60 percent, while Logistic Regression achieved the highest accuracy at 78.82 percent. The difference between these two models was small. Random Forest achieved 77.40 percent accuracy and a ROC AUC of 79.30 percent, while Decision Tree produced 72.00 percent accuracy and a ROC AUC of 65.01 percent. These results show that model performance varied under the same experimental conditions, but they do not establish that one algorithm is universally superior for customer churn prediction.

The detailed results also revealed an important limitation. Although the Gradient Boosting model achieved an

overall accuracy of 78.61 percent, its recall for the churn class was only 51.60 percent. The model therefore failed to identify a substantial proportion of customers who actually churned. This finding demonstrates why overall accuracy should not be considered in isolation when evaluating churn prediction models. Precision, recall, F1 score, ROC AUC, and confusion matrices provide additional information about how the model behaves for each class.

The five fold stratified cross validation results provided additional evidence about model consistency. Gradient Boosting achieved a mean ROC AUC of approximately 84.48 percent with a standard deviation of approximately 0.65 percentage points across the five folds. The similarity between the cross validation and holdout results provides supporting evidence that the model maintained comparable discrimination across the evaluated data partitions within this experimental dataset.

The study contributes an empirical comparison of four established classifiers under a consistent experimental framework. Its main value lies in controlling the conditions under which the models were compared. The same preprocessing and Mutual Information based feature selection procedure was used for each classifier, while the holdout test set remained separate from model development. This design provides a clearer comparison than directly comparing performance values reported by studies that use different datasets, sampling procedures, feature selection methods, and evaluation protocols.

Several limitations restrict the interpretation of the findings. The experiment used a single telecommunications dataset and did not include external validation. The study also retained the original class distribution rather than applying oversampling or undersampling. In addition, the experiment did not include an ablation study to determine the independent effect of Mutual Information feature selection. The results also do not measure the financial effect of using a churn prediction model in an operational retention programme.

Future research can extend this work by evaluating the models on additional telecommunications datasets and by testing the effect of different class balancing strategies. Further experiments could compare model performance with and without feature selection, investigate broader hyperparameter optimization, and examine probability threshold selection when the primary objective is to identify a larger proportion of potential churners. External validation and longitudinal evaluation would also provide stronger evidence about how well the models transfer to new customer populations and changing business conditions.

Overall, the findings show that the four evaluated classifiers produced different results when tested under the same experimental conditions. Gradient Boosting recorded the highest numerical ROC AUC, while Logistic Regression achieved higher accuracy. The small difference between these models, together with the relatively low recall for the churn class, shows that model selection for customer churn prediction requires consideration of several performance measures and the intended use of the prediction system. The results provide a controlled empirical basis for further work on improving churn identification and evaluating predictive models across broader telecommunications datasets.

References

Arlot, S., & Celisse, A. (2010). A survey of cross validation procedures for model selection. *Statistics Surveys, 4*, 40–79. https://doi.org/10.1214/09-SS054

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

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine, 162*(1), 55–63. https://doi.org/10.7326/M14-0697

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter selection techniques. *Expert Systems with Applications, 34*(1), 313–327. https://doi.org/10.1016/j.eswa.2006.09.038

De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. *Journal of Discrete Mathematical Sciences and Cryptography, 25*(7), 1965–1985. https://doi.org/10.1080/09720529.2022.2133238

Edwine, N., Wang, W., Song, W., & Ssebuggwawo, D. (2022). Detecting the risk of customer churn in telecom sector: A comparative study. *Mathematical Problems in Engineering, 2022*, 8534739. https://doi.org/10.1155/2022/8534739

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

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

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(3), 515–524. https://doi.org/10.1016/j.eswa.2005.09.080

IBM. (n.d.). Telco customer churn dataset. IBM customer churn prediction sample. https://github.com/IBM/telco-customer-churn-on-icp4d

Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. *Machine Learning and Knowledge Extraction, 7*(3), 105. https://doi.org/10.3390/make7030105

Jain, H., Khunteta, A., & Sharma, N. (2026). Telecom customer churn prediction: A detailed systematic review of data, methodology, and evaluation methods. Telecommunication Systems, 89(3), 1–48. https://doi.org/10.1007/s11235-026-01491-3

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E, 69*(6), 066138. https://doi.org/10.1103/PhysRevE.69.066138

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*(2), 472–484. https://doi.org/10.1016/j.eswa.2005.04.043

Lima, E. O., Mues, C., & Baesens, B. (2009). Domain knowledge integration in data mining using decision tables:

Case studies in churn prediction. *Journal of the Operational Research Society, 60*(8), 1096–1106. https://doi.org/10.1057/jors.2008.161

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14–23. https://doi.org/10.1002/widm.8

Mettle, H. N. A., Ayerko, E. H., & Appiahene, P. (2026). Machine learning based customer churn prediction in telecommunication industry. *Applied Computational Intelligence and Soft Computing, 2026*, 8496186. https://doi.org/10.1155/acis/8496186

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Nguyen, V. T., Chan, J., & Bailey, J. (2014). Reconsidering mutual information based feature selection: A statistical significance view. In *Proceedings of the AAAI Conference on Artificial Intelligence, 28*(1). https://doi.org/10.1609/aaai.v28i1.8953

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max dependency, max relevance, and min redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Saito, T., & Rehmsmeier, M. (2015). The precision recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Rahman, M. S. (2022). A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. *PLOS ONE, 17*(12), e0278095. https://doi.org/10.1371/journal.pone.0278095

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217. https://doi.org/10.1016/S0377-2217(03)00069-9

Verbraken, T., Verbeke, W., & Baesens, B. (2014). Profit optimizing customer churn prediction with Bayesian network classifiers. *Intelligent Data Analysis, 18*(1), 3–24. https://doi.org/10.3233/IDA-130625

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.031

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing, 14*, 431–446. https://doi.org/10.1016/j.asoc.2013.09.017

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization, 14*, 100342. https://doi.org/10.1016/j.rico.2023.100342

Wei, C. P., & Chiu, I. T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications, 23*(2), 103–112. https://doi.org/10.1016/S0957-4174(02)00030-1