

Customer Churn Prediction Using Machine Learning: A Comparative Analysis of Classification Algorithms

Swathi D M¹ Assistant Professor, G M University Davanagere

Anjali S M², Nirmala B L³, Punya V⁴, Sandeep katti⁵, Vijay H⁶, Students, 4th Semester MCA,
Department of Faculty of Computing and IT, G M University Davanagere

Abstract:

The importance of customer churn prediction cannot be overstated since it has a vested role in enabling businesses to retain their customers while reducing the loss of revenue owing to stiff competition. The paper puts forth a machine learning-based customer churn prediction methodology that can help firms foresee the customers who are likely to leave the service provider. For this purpose, the IBM Telco Customer Churn Dataset was used, comprising a total of 7,041 records documenting several facets of customer data such as the demographics, service used, billing information, and churn status. To achieve maximum efficiency of the collected data related to customers and the accuracy of the conclusions reached through it, a structured data preprocessing pipeline that includes missing value treatment, feature encoding, feature scaling, and train-test split was used. Employing six methods of supervised machine learning, the author evaluated the performance of the models used based on the values of accuracy, precision, recall, F-Score, ROC-AUC, and others. In order to make the model less complex and more interpretable, the SHAP approach was also implemented in the study to study the impact of the independent variables on the result. As an outcome of this research, it has been established that among all the models LightGBM gave the most promising results. The suggested structure provides an effective practical solution in terms of predicting customer churn, enabling companies to apply proactive measures of customer retention, substantiate their decisions, and ensure enhanced sustainability of their business.

Keywords: *Customer Churn Prediction, Machine Learning, LightGBM, SHAP, Customer Retention, Predictive Analytics, Flask.*

1. Introduction:

Organisations in competitive sectors have faced the dilemma of customer retention, since it is cheaper to keep existing customers than to win back new customers. As a result, customer churn has become a major problem faced by many businesses due to its negative impact on finances and future development strategies. Because of this fact, it is important to identify the characteristics of a customer who is likely to disappear so that businesses can address the very issue of improving customer satisfaction.

The great amount of customer data available nowadays has allowed the use of ML algorithms for predicting customer churn in businesses. Data-mining technologies, including Machine Learning algorithms can easily identify patterns based on historical customer data. Being able to predict customer churn, companies have a chance to come up with effective retention strategies, improve their operations and minimize financial losses.

This study presents a framework for predicting customer churn based on machine learning using the IBM Telco Customer Churn Dataset containing 7,043 customer entries and 21 customer traits.

Six machine learning algorithms classified as supervised are involved in the testing process: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, and LightGBM, allowing for comparison among them and identification of the best prediction algorithm. Besides, SHapley Additive exPlanations (SHAP) will be used for interpreting the prediction models, and the best model will be implemented in a web app based on Flask.

2. Goals of the Research:

1. To create a customer churn prediction model using IBM's Telco Customer Churn dataset.

The idea here is to make a predictive model that will be able to create accurate predictions based on the processed data.

2. To compare the effectiveness of six supervised machine learning techniques.

This research compares Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, and Light GBM techniques to find the most effective one in terms of churn prediction.

3. To evaluate the effectiveness of the created models.

The models are assessed using accuracy, precision, recall, F1, ROC AUC, balanced accuracy, MCC, Kappa, and log loss.

4. To improve model interpretability through the use of SHAP and feature importance analysis.

The goal is to determine which features affect the churn prediction.

5. To create a Flask-based web application for predicting customer churn.

The system created allows for prediction at the level of individual customers as well as at the batch level, thus proving the efficiency of the machine learning model in customer relationship management and business decision-support systems.

3. Literature Review:

Customer churn prediction has grown to be one of the most important fields of research because of its strong influence on both customer retention and business profits. In the past few years, many scientists worked on the task of accurately predicting customer churn related to different fields like telecommunication, banking, insurance, retail, and subscription-based services, by using different Machine Learning (ML) and Deep Learning (DL) techniques. New studies aim at improving the process of prediction with the use of ensemble learning algorithms, and explainable artificial intelligence (XAI), and hybrid machine learning models. In addition to the model performance, scientists stress the importance of model interpretation as well.

Table 1. Literature Survey

Su b- ind ex	Author s and Public ation Year	Title	Researc h Method ology	Main Finding s	Shortco mings
[1]	Sikri et al. (2024)	Using Machine Learning to Predict Custo	Decision Trees, Boosted Trees	XGBoos t overcam e class imbalan ce problem	Data source limited to the telecom industry; lack of AI

		mer Churn in Teleco m Sector		s and provided the highest predictio n accuracy .	transpare ncy.
[2]	Suh (2023)	Predict ing Custo mer Churn Using Machi ne Learn ing in an Applia nce Rental Busine ss	Random Forest, Boosted Trees	Achieve d highly accurate predictio ns using real industry- based data.	Focused on only one industry.
[3]	Chang et al. (2024)	Predict ing Custo mer Churn in Teleco m Using Machi ne Learn ing Models	Random Forest, SHAP, LIME	Utilized explaina ble AI techniqu es to interpret predictio n results.	No comparis on among different ensemble models.
[4]	Abdel Aziz et al. (2025)	Revie w of Machi ne Learn ing and	CNN, XGBoos t, Ensembl e	Deep learning techniqu es achieved the	High computati onal cost due to complex

		Deep Learning Techniques for Customer Churn Prediction	Learning	highest accuracy among the evaluated models.	calculations.
[5]	Ruchika Bhuria et al. (2025)	Ensemble-Based Customer Churn Prediction in Banking	Voting Classifier, SMOTE	Ensemble learning improved banking customer churn prediction performance.	Evaluation performed on a limited banking dataset.
[6]	Mehdi Imani et al. (2025)	Customer Churn Prediction: A Systematic Review	Literature Review	Summarized recent trends, challenges, and developments in churn prediction research.	No experimental implementation or model validation.
[7]	Mohamed Abdelhady et al. (2025)	Leveraging Artificial Intelligence for Predictive	Random Forest, Artificial Neural Network (ANN)	Combined AI techniques with CRM strategies for churn	Limited real-world application and validation.

		Customer Churn Modeling		prediction.	
[8]	Renuka Agrawal et al. (2026)	Hybrid SMOTE and GAN-Based Telecom Churn Prediction	SMOTE, GAN, LightGBM	Improved prediction accuracy for imbalanced telecom churn datasets.	Higher computational complexity and cost.
[9]	Ali El Attar et al. (2026)	Explorable AI-Driven Customer Churn Prediction	Ensemble Learning, SHAP	Enhanced model transparency and explainability.	Requires testing on larger and more diverse datasets.
[10]	Jitendra Maan et al. (2023)	Customer Churn Prediction Model Using Explainable Machine Learning	XGBoost, SHAP	Demonstrated the importance of explainable machine learning in churn prediction.	Comparisons conducted with only a limited number of algorithms.

Comparative Analysis

The literature reviewed indicates that Machine Learning has become the primary technique in customer churn prediction over classical statistical methods due to better predictive power. The

majority of the scholars applied supervised machine learning techniques, such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, and LightGBM, to classify customers into churners and non-churners. The use of ensemble learning techniques like XGBoost, Random Forest, and LightGBM provides better forecasting results than traditional classification methods since such methods are capable of grasping complex relationships between customer features.

Some researchers also utilized Explainable AI (XAI) techniques, such as SHAP and LIME, making models more transparent. The adoption of such explainable models provides firms with the opportunity to understand the impact of various customer characteristics on the prediction of customer churn and make more reasonable decisions. Finally, some recent works deal with the class imbalance problem, by making use of SMOTE, and GAN-based synthetic data generation techniques.

Nevertheless, several issues remain unresolved. A notable limitation across many studies is their reliance on a single dataset for model evaluation, which limits the generalizability of the reported results. Many researchers have concentrated on improving prediction efficiency without putting much focus on the importance of features and practicality. While the ensemble methods achieved the best results, very few studies have compared various techniques extensively using the same preprocessing and evaluation methods.

This research studies these issues by implementing and testing 6 supervised machine learning approaches (Logistic Regression, Decision Trees, Random Forest, SVM, XGBoost, and LightGBM) based on the IBM Telco Customer Churn dataset. The study uses a uniformly applied preprocessing method and multiple metrics to be able to boost efficiency and explainability of the whole approach. Additionally, the method is implemented through a web application that demonstrates its usefulness.

At present, there are strengths and weaknesses evident in literature review studies pertaining to customer churn prediction.

Strengths of Current Studies

- Recent research studies utilize machine learning algorithms effectively to predict customer churn accurately.
- Ensemble approaches including XGBoost, Random Forest, and LightGBM have shown a great prediction capacity.

- Technologies promoting Explainable AI (SHAP and LIME) have made models more transparent and interpretable.
- Techniques of balancing datasets like SMOTE and GAN have effectively solved the class imbalance problem.
- The field of research is developing in various spheres including telecommunications, banking, insurance, and retail.

Weaknesses of Current Studies

- Problems with model generalization arise due to the use of a single dataset from one industry in the majority of studies.
- Warning that researchers often concentrate only on prediction itself and neglect model explanation is justified.
- Several comparative studies researching multiple methods of machine learning in the same experimental conditions are not available.
- Less attention is paid to how machine learning methods can be applied in practice through web applications or using cloud technologies.
- Hardly any research works have been conducted on the application of Explainable AI, feature importance, and comparative analysis in one integrated framework

4. Research Gap:

The review of recent papers of customer churn prediction led us to identify the following research gaps:

- Limited Comparison of ML Models: Most existing studies only evaluate a few ML algorithms making it difficult to understand the optimal model for customer churn prediction.
- Lack of Explainability: Many models focus only on high-accuracy results rather than elucidating the reasons for predictions thus decreasing transparency and trust of users.
- Limited Use of Explainable AI: Very few studies employ explainability techniques such as SHAP to reveal the contribution of features to customer churn prediction.
- Dataset-Specific Models: Most works use one industry-specific dataset aggravating the issue of generalizability of the obtained models to other industries.
- Insufficient Performance Evaluation: Several studies evaluate models based on accuracy only but don't take into account other important metrics including Precision, Recall, F1, ROC, and more.

- **No Practical Implementation:** Several research works conclude at the stage of creating and testing a model but do not implement the predictive model into real-life applications.
- **Insufficient Analysis of Feature Importance:** Previous research has shown what customers have churned but did not analyze the factors that contributed to their loss.
- **The Need of a Complete Framework:** A unified framework is required for data pre-processing, comparative machine learning analysis, explainable AI, full evaluation, visualization, and further deployment in one system.

How Our Research Addresses These Shortcomings

The issues outlined here are answered through the research studies:

- Contrasting six supervised algorithms and determining outcomes under identical experimental conditions.
- Implementing SHAP to provide insights on prediction make.
- Conducting feature importance analysis in an effort to identify processes leading customer churn.
- Using multiple metrics for model assessment in order to obtain a comprehensive result.
- Creating a web application based on Flask technology for individual or batch prediction.
- Providing full cycle of research being done including prediction, explanations, visualization and deployment.

5. Research Contributions:

The significant contributions made by this research can be summarized as follows:

1. In-depth comparative evaluation of six supervised machine learning techniques such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, and LightGBM was performed using standard preprocessing steps along with the IBM Telco Customer Churn dataset. This allows us to make fair comparison of the advantages and disadvantages of different algorithms.
2. Explainable framework for predicting customer churn was built with implemented SHapley Additive exPlanations (SHAP). Unlike regular machine learning prediction frameworks, this framework gives understandable predictions since it shows the

importance of features related to individual customers and their explanations.

3. Feature importance study was done to find out which factors determine customer churn the most. The study gives the information about the most important customer attributes related to churn, which helps a company develop its retention strategies.

4. Multiple evaluation metrics were used to perform an exhaustive performance evaluation, such as accuracy, precision, recall, F1 score, balanced accuracy, Matthew correlation coefficient (MCC), Cohen's Kappa, ROC-AUC, and log loss, yielding an assessment of the model performance that is more credible than simply calculating prediction accuracy.

5. A practical web application based on the Flask framework was developed to illustrate the applicability of the proposed framework in real life. The application allows one to perform both single-customer and batch churn predictions, find prediction probabilities, keep records of predictions, and have a user-friendly interface for business entity representatives.

6. Research-ready visualizations were created to interpret the model and analyze the results, namely, Customer Churn Distribution, Correlation Heatmap, ROC Curve, Precision-Recall Curve, Feature Importance, and SHAP Summary Plot, which will make the interpretation of the proposed framework easier and results presentation more efficient.

This research offers a comprehensive and practical framework for predicting customer churn, which integrates efficient machine learning models, transparent artificial intelligence, thorough performance evaluation, as well as a web application ready for commercial use. The proposed solution can help companies find customers prone to churning, develop effective retention strategies, and make informed business choices.

6. Problem Definition:

Customer churn, or customer attrition, is a problem that just won't go away for companies in telecom, banking, insurance, retail, and subscription-based businesses. Whenever a customer stops utilizing products or services of particular companies, they incur considerable losses in profit, receive less revenue, have lower customer lifetime value and must allocate more funds for promoting their business in future. Studies show that retaining existing customers is more efficient than acquiring new ones,

making customer retention a crucial issue in terms of successful business development in the long run.

Despite the fact that companies gather a considerable amount of information about their customers through transactions and billing history, service usage, and communication records, most firms still rely solely on traditional statistical tools or manual methods when it comes to revealing customers that will likely leave. Such approaches are ineffective when it comes to the capturing of complex behavior and continue showing unsatisfactory prediction results. Thus, organizations cannot detect customers with high probability of leaving at an early stage, thus decreasing efficiency of retaining customers' activities and increasing attrition rates.

Another major challenge is the absence of interpretable models of prediction. There are a number of models of prediction that use machine learning techniques to predict accurately, but they work as black-boxes, which make it hard for the management team to comprehend the reasons for customer departure predictions. Therefore, the company cannot execute targeted preservation strategies due to the absence of intelligible reasons.

Thus, there is a need for an intelligent and interpretable model of customer churn prediction that will enable identification of the at-risk clients and provide insights about factors of churn. Such a model would support proactive preservation practices and improve customer satisfaction as well as minimize losses of the organization in the longer run.

7. Research Methodology:

This research follows a systematic machine learning methodology in the development of a novel customer churn prediction system. The framework consists of various stages, including data gathering, pre-processing, exploratory data analysis (EDA), feature engineering, development, evaluation, explanation, and deployment in a web-based application. Each step is designed to ensure correct predictions, interpretable models, and applicability of the solution in customer relationship management.

7.1 Data Collection

The dataset used for this research is the IBM Telco Customer Churn Dataset, due to its popularity as a benchmark dataset for the customer churn prediction problem. The dataset contains 7,043 customer records, with details on demographics,

subscribed services, billing, and contract information, along with each customer's churn status. Churn serves as the target variable, indicating whether or not a customer discontinued the service. The dataset was chosen as it provides enough customer-related attributes for testing various machine learning techniques in realistic business settings.

7.2 Data Preprocessing

To ensure the data was of good quality and would support reliable results, it was carefully preprocessed before training. The process started by identifying and resolving duplicate entries and missing values. The Total Charges attribute was converted into a numerical format. The missing values in this attribute were then treated appropriately. The customerID attribute which is not related to any predictions in the dataset was eliminated. The categorical features such as the customer's gender, type of contract, type of payment, and type of internet service were transformed by means of encoding into numerical format. The next step in the process of data preprocessing was standardization, i.e. transforming numerical features in order to improve the performance of the model. The prepared dataset was then split into 80% of training data and 20% of test data using a fixed random state for the purpose of reproducibility.

7.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was used to analyze the dataset and to find out important trends in customer behavior prior to training of the model. Different types of statistical analysis were performed and visualization methods were applied to figure out the distribution of customer churn, relationship between the key features, and impact of the attributes of the customers on the churn. The publication-quality visualizations were created among which one could find such graphs as Customer Churn Distribution, Correlation Heatmap, ROC Curves, Precision-Recall Curves, Feature Importance and SHAP Summary Plot.

7.4 Feature Engineering

Feature engineering aims to enhance the quality of prediction by choosing appropriate input variables for machine learning models. Categorical variables were transformed into numerical values, and numerical variables were normalized whenever necessary. An importance feature analysis was carried out later using SHAP and ensemble learning techniques to reveal important variables causing customer churn. The process has increased the prediction accuracy of the model as well as its interpretability.

7.5 Model Training

Six supervised machine learning algorithms were created to solve the problem: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), XGBoost, and LightGBM. In order to achieve better performance, hyperparameter tuning was done using GridSearchCV, and Stratified 5-Fold Cross-Validation was applied to evaluate the reliability of the obtained model and decrease the risk of overfitting. The training was done with the help of Scikit-learn, XGBoost, and LightGBM libraries in Python.

7.6 Model Evaluation

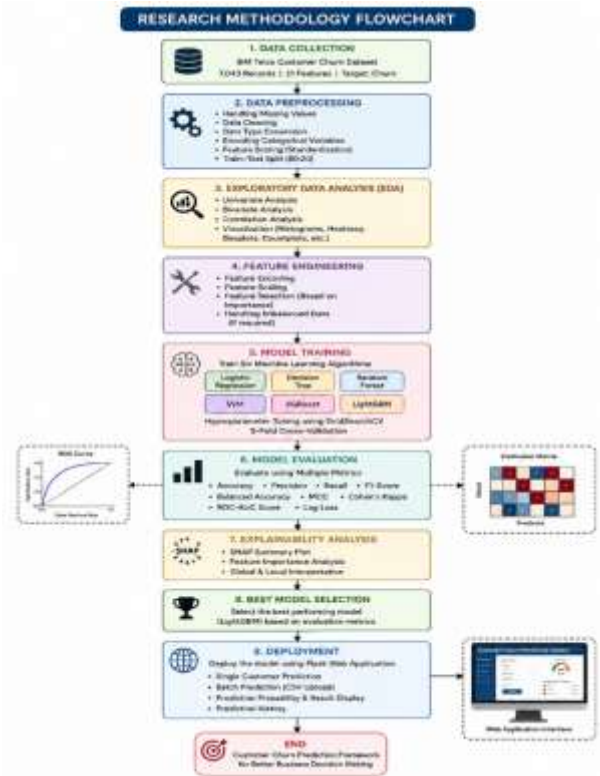
Multiple classification performance metrics were used to identify the best performing models based on all available models. Model performance was judged from several angles — overall correctness (Accuracy, Balanced Accuracy), the trade-off between false positives and false negatives (Precision, Recall, F1-Score), and more robust measures like MCC, Cohen's Kappa, ROC-AUC, and Log Loss that account for class imbalance and prediction confidence. It was also found out that SHAP techniques can help in achieving the goal of model evaluation and interpretation

7.7 Model Deployment

To test the applicability of the proposed framework, the best-performing model has been deployed through a web app made by using Flask. The application gives a simple interface for predicting individual customers as well as predicting clients in batch using CSV files, showing the probabilities and keeping prediction results.

The proposed method combines data preprocessing, advanced machine learning methods, effective model performance evaluation, concepts of explainable AI, and web deployment under the umbrella of the single customer churn prediction approach.

Figure 1. Research Methodology Flowchart



8. Dataset Description:

8.1 IBM Telco Customer Churn Dataset In this study, the IBM Telco Customer Churn Dataset is used, which is a well-known dataset for predicting customer churn wherein it is available publicly at the IBM Sample Data repository as well as at Kaggle as well. It has been used widely for several research studies to validate machine learning algorithms. The data comprises of a total of 7,043 customers, whereby the data have the demographic as well as other related information of the customers. For each customer, many attributes are included such as the demographic details, the services used by them, account information as well the bill payment details. The aim of the dataset is to identify the possibility of the customer leaving the company's services, based on their prior behavior and the service usage patterns.

The original dataset consists of a total of 21 attributes. Out of these attributes, one is the target variable (Churn) while others are the predictor variables, which are used to infer the possible behaviors and trends of the customer. Some of the attributes included in dataset are gender, senior citizen status, partnership, dependents, services used, type service internet, type of contract, payment method, monthly charge amount. The target variable is Churn, which shows whether the customer has churned ("Yes") or not ("No").

Prior to the development of the model, the data underwent proper preprocessing. Superfluous attributes like customerID were

eliminated, missing values were fixed, categorical variables were encoded, and appropriate numerical features were standardized. The preprocessed data was then split using an 80-20 ratio in order to create training and testing datasets.

The IBM Telco Customer Churn dataset is a good sample to analyze for customer churn prediction because the concerned dataset possesses the right proportions of different attributes: demographic, behavioral, financial, and service attributes. Moreover, the dataset is big enough to assess many machine learning algorithms while remaining computationally cheap for academic purposes. The extensive use of the dataset in other studies allows making comparison with other researches and confirming that the suggested prediction scheme is effective.

Table 8.1 Dataset Characteristics

Key Feature	Description
Dataset Name	IBM Telco Customer Churn Dataset
Dataset Source	IBM Sample Data / Kaggle
Field	Telecommunications
Number of Customers	7,043
Number of Features	21
Predictor Features	20
Target Feature	Churn
Prediction Method	Binary Classification
Definition of Classes	Yes (Churn), No (No Churn)
Presence of Missing Values	Yes (only in <i>TotalCharges</i> , handled during preprocessing)
Type of Data	Numerical and Categorical
Proportion of Train and Test Dataset	80% for Training and 20% for Testing
Programming Language	Python
Libraries Used	Pandas, NumPy, Scikit-learn, XGBoost, LightGBM, SHAP, Flask

Adapted Overview of the Dataset

- Dataset Name: IBM Telco Customer Churn
- No of Instances: 7043

- Total Attributes: 21
- Target Attribute: Churn
- Task Type: Binary Class Classification
- Training Instances: about 5,634 (80% of the data)
- Testing Instances: about 1,409 (20% of the data)
- Best Model: LightGBM

9. Data Preprocessing:

Data preprocessing is an important step in machine learning since the input data quality influences the effectiveness of the predictions. Raw data sets may have missing values, duplicates, errors in formatting, and other data types not permissible for use in machine learning algorithms. Hence, a systematic approach to preprocessing was adopted. Various steps for data preprocessing are studied in this project.

9.1 Treatment of Missing Values

The IBM Telco Customer Churn dataset obtained missing values in the *TotalCharges* attribute owing to empty entries by newly registered customers. At the outset, the empty values were identified and transformed to NaN (Not a Number) format. Subsequently, the missing values were dealt with by replacing them with the median of the specific column so that precious information regarding the customers is preserved. This method ensures that the overall distribution of information is retained without compromising the model's performance.

9.2 Removal of Duplicate Entries

Duplicate customer information creates biases in the training of the model leading to undesirable prediction results. Hence, the dataset was checked for duplicates on the basis of the unique customer ID. No major duplicates were found in the dataset. Also, the *customerID* attribute was excluded from the dataset since it serves only as a unique identifier and does not help in churn prediction.

9.3 Label Encoding

To convert various binary categorical variables (i.e. Gender, Partner, Dependents, PhoneService, PaperlessBilling, churn, and other yes/no attributes) into numerical values we applied label encoding. Each categorical variable is turned into a binary numerical representation (0 and 1) to allow machine learning algorithms to use this data properly and without losing valuable information.

9.4 One-Hot Encoding

Categorical variables that include more than two categories, such as InternetService, Contract, MultipleLines, and PaymentMethod, have undergone one-hot encoding. This method means that one binary column is created for each category thus avoiding creation of false ordinal relationships between categorical variables. One-hot encoding enhances the possibilities of machine learning algorithms to learn from the data.

9.5 Standardization

Numerical features such as tenure, MonthlyCharges, and TotalCharges have different ranges of values which can affect classification algorithms negatively. Due to this reason StandardScaler was used to transform numerical features into zero mean and one unit variance values.

9.6 Train-Test Split

The dataset was split into training and testing sections using an 80:20 ratio post preprocessing. Thus, around 80% of data was involved in training the model, while the remaining 20% was kept aside for testing. The fixed random state is set for robust generation of results of the experiments with different machine learning algorithms.

9.7 Feature Engineering

Feature engineering was done to improve prediction efficiency by identifying relevant characteristics of the customers and processing these characteristics for machine learning. With this in mind, the irrelevant characteristics (customerID) were removed, while encoded categorical variables and standardized numerical ones were merged into one feature matrix. In addition, the feature importance analysis was carried out with the help of Random Forest, XGBoost, LightGBM, and SHAP to determine the key factors responsible for customer churn. Churn has been found to be influenced by such features as tenure, Contract Type, Monthly Charges, Total Charges, Internet Service, Payment Method, Online Security, etc.

To sum up, the preprocessing technique applied in this research allowed for the creation of high-quality input data for future comprehension, accuracy, and consistency, and it enhanced the reliability and reproduction of the framework proposed.

Table 9.1 Summary of Data Preprocessing Steps

Step of Preprocessing	Purpose	Technique Used
Dealing with Missing Values	To handle missing information in the dataset and improve data quality.	Median Imputation
Removing Duplicates	To eliminate duplicate records and ensure data consistency.	Duplicate Detection
Removing IDs	To remove unnecessary variables that do not contribute to prediction.	customerID Deletion
Label Encoding	To convert categorical variables into a numerical format usable by machine learning algorithms.	Label Encoder
One-Hot Encoding	To transform categorical variables into vectors of 0s and 1s.	Pandas One-Hot Encoding
Feature Scaling	To normalize numerical variables and improve model performance.	StandardScaler
Train-Test Split	To divide the dataset for model training and evaluation.	80% Training Set and 20% Test Set
Feature Engineering	To enhance prediction quality by identifying and selecting important features.	Feature Selection and SHAP Analysis

10. Models of Machine Learning

Machine Learning (ML) technologies offer intelligent tools for evaluating past customer information and forecasting future customer behavior. In predicting customers who may leave the company, supervised ML methods analyze data from past

customers to assess the likelihood of customer departure or retention. Because of the different operation approaches of various algorithms, the research uses six popular ML algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, and LightGBM. The aim is to select the most suitable algorithm for customer churn forecasting.

10.1 Logistic Regression

Logistic Regression is one of the most popularly known supervised classification algorithms that is widely applied in binary classification problems. Using the logistic (sigmoid) function, it serves to predict the likelihood of the customer being categorized as non-churning or churning. What is more, Logistic Regression is easy to use, does not require much computational power, and serves as a decent starting point for evaluating more advanced machine learning approaches.

In this research, the author applied Logistic Regression as a method for predicting customer churn due to the fact that it is a binary classification problem with two possible outcomes, i.e. customer churn and no customer churn. Moreover, Logistic Regression generates coefficients that make it possible to explain how a certain feature(s) influences customer churn. Even though Logistic Regression assumes a linear relationship between the features and the target, it serves as a good reference point for evaluating the results of more complex models.

10.2 Decision Tree

Decision Tree is one of the supervised learning methods that analyze and classify data by breaking the dataset down into smaller datasets according to the value of different attributes. The system builds up a tree with different levels of hierarchy in which every node is responsible for making a decision, each branch represents a possible outcome, and each leaf denotes the final result.

The Decision Tree was chosen due to its simplicity and interpretability, as well as its ability to model non-linear relationships maintaining the features' values without using scaling. The method is also suitable to treat both types of attributes (numerical and categorical) and give understandable rules for decision-making purposes. At the same time, Decision Trees are likely to overfit the data, especially if the tree is too deep, making them less beneficial as a baseline model.

10.3 Random Forest

Random Forest is an ensemble learning algorithm that employs the technique of bootstrap aggregation (bagging) to combine various Decision Trees. Rather than building a single decision tree, Random Forest generates a multitude of trees based on different randomly chosen sets of data and characteristics, and the final outcome is arrived at through voting from all these trees.

This algorithm has been selected due to its amazing capability of enhancing prediction capability and decreasing the risk of overfitting in comparison to a single Decision Tree. This algorithm has good resiliency and works perfectly with noisy data. It performs well with missing values, as well as being suitable for datasets with numerical and categorical variables.

10.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a highly effective supervised machine learning technique used for finding the right decision boundary that guarantees maximum separation between different classes. SVMs are particularly good in classification tasks involving a high-dimensional dataset as they can represent complex and abstract relationships through the use of kernel functions.

Despite its high computational complexity, SVM is chosen for this study due to its good classification strength and good performance in prediction problems. The algorithm gives good results when the number of features is fairly high and is relatively resilient to overfitting as opposed to many traditional classifiers. Both the advantages and disadvantages of SVM can be properly assessed in a comparison study involving other classification techniques based on SVM.

10.5 XGBoost

Extreme Gradient Boosting (XGBoost) can be considered as a sophisticated ensemble learning technique which builds its framework on gradient boosting fundamentals. XGBoost consists in the process of stepwise forming decision trees in such way that each new tree is built to fix the mistakes of its predecessors. XGBoost uses various regularization approaches that ensure more effective generalization and reduce the possibility of overfitting.

XGBoost was chosen as it has proven to be an effective prediction method, as well as has performed well in a lot of predictions of customer churn in practice as well as many machine learning competitions. The method has proven to be

useful in working with missing values, conducts parallel processing and reveals complicated interactions between features. Such advantages provide XGBoost with an ability to conduct effective predictions while having high computational speed hence making the method widely used in industries.

10.6 LightGBM

Light Gradient Boosting Machine (LightGBM) is a very efficient gradient boosting method that has been designed to handle huge datasets quickly and with little memory. Unlike other gradient boosting algorithms, LightGBM uses histogram-based learning and a leaf-based tree structure to reach a high accuracy rate at low computing costs.

The main reason for the selection of LightGBM was its potential of working with large datasets, many variables, and complicated customer behavioral processes. With respect to its ability to perform parallel learning, and to reduce training time, the algorithm is outperforming numerous other methods. LightGBM demonstrated outstanding overall performance when assessed according to different performance indicators such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC score. The system is also integrated with SHAP allowing making the comprehensive analysis of the variable importance and increasing the model's interpretability.

Table 10.1 Summary of Machine Learning Models

Algorithm	Category	Reason for Selection	Advantages
Logistic Regression	Linear Classifier	Used as a baseline binary classification model for comparison.	Simple to understand, interpretable, and computationally efficient.
Decision Tree	Tree-Based Learning	Provides easy-to-understand decision rules and classification logic.	Handles nonlinear relationships, interpretable, and easy to visualize.
Random Forest	Ensemble Learning	Selected to improve prediction	Reduces overfitting, improves

		stability and overall performance.	accuracy, and provides feature importance measures.
Support Vector Machine (SVM)	Margin-Based Classification	Chosen for its strong classification capability in complex datasets.	Effective for high-dimensional data and capable of handling nonlinear boundaries.
XGBoost	Gradient Boosting	Selected due to its high predictive accuracy and strong performance in classification tasks.	Includes regularization, fast execution, robust performance, and handles missing values efficiently.
LightGBM	Gradient Boosting	Chosen as the primary model because of its superior efficiency and predictive performance.	Fast training, low memory consumption, scalable, and highly accurate predictions.

11. Experimental Setup:

For the experiments described in this paper, the research utilized a Windows PC and the Python programming language alongside the PyCharm IDE. To allow for reproducibility, computational efficiency, and ease of deployment, machine learning libraries that were available to the public were used for the implementation. The IBM Telco Customer Churn Dataset was employed to train and evaluate the machine learning algorithms. The dataset went through the preprocessing stage prior to training

the algorithms through the steps of handling missing value, feature encoding, feature scaling, and feature engineering.

The dataset was split into two parts that were made up of 80% training and 20% testing datasets while the random state was kept constant at 42 for the same reasons as explained above. To improve reliability and reduce overfitting risk, Stratified 5-Fold Cross-Validation was applied during training. In total, six supervised learning methods were tested: Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost, and LightGBM. Development relied on a standard Python toolkit — Pandas and NumPy handled data manipulation, Scikit-learn covered model building and evaluation, XGBoost and LightGBM provided gradient boosting capabilities, SHAP supported model explainability, and Flask was used to deploy the final model as a web application. All trials were conducted under similar preprocessing conditions in order to avoid bias in the comparison of algorithms. The trained LightGBM model and its feature encoders and preprocessing tools were saved for deployment in a Flask-based customer churn prediction system.

Experimental Workflow

The detailed workflow of this experiment comprised the steps below:

1. Obtaining the IBM Telco Customer Churn Dataset.
2. Conducting Exploratory Data Analysis.
3. Applying six supervised ML techniques.
4. Perfecting model hyper parameters and evaluation of the models through GridSearchCV and Stratified K-Folds Cross-Validation.
5. Measuring performance on the basis of various metrics.

12. Results and discussion:

This section focuses on the evaluation of the performance of the proposed customer churn prediction framework. Six machine learning methods – namely Logistic Regression, Decision Tree, Random Forest, SVM, XGBoost, and LightGBM – were trained and evaluated on the dataset of IBM Telco Customer Churn. The models were evaluated based on several evaluation metrics, viz. Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Furthermore, several visualization techniques including Customer Churn Distribution, Correlation Heatmap, ROC Curve, Precision-Recall curve, feature importance, and SHAP

Summary Plot were utilized to analyze customer behavior and to assess the interpretation of the prediction results.

12.1 Customer Churn Distribution

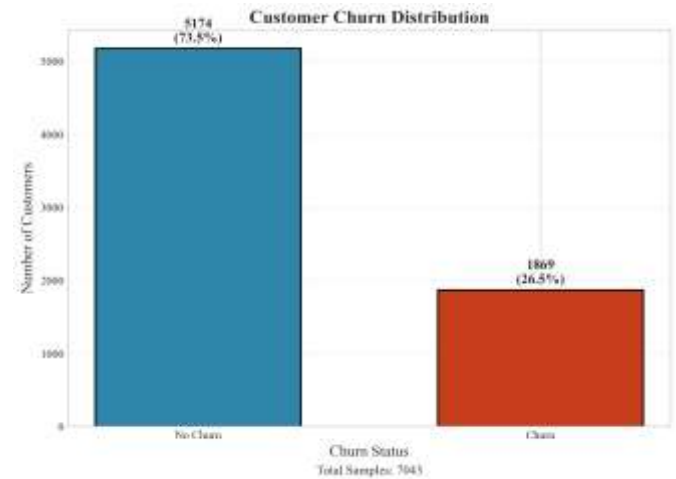


Figure 1. Customer Churn Distribution

The distribution of customer churn gives an insight into the distribution of classes in the group. The data collected from IBM's Telco Customer Churn shows that out of 7,043 registered customers, 5,174 customers (73.5 percent) were still retaining the company (retained customers); only 1,869 customers (26.5 percent) of the total have opted not to continue using the service. Accordingly, we should conclude that the dataset is moderately unbalanced, where the number of non-churn customers significantly exceeds the number of churn customers.

The data about class distribution is quite informative because it shows the significance of churn prediction for the company. It is evident that companies are more interested in ensuring the success of their predictions and measuring the loss of returning customers.

12.2 Correlation Heatmap



Figure 2. Correlation Heatmap

The correlation heatmap indicates the relationship among the most important customer characteristics obtained from feature analysis. Tenure and TotalCharges have a strong positive relation, which implies that long-term customers usually incur higher charges.

A moderate positive relation was observed between MonthlyCharges and Fiber Optique Internet Service, showing that premium internet customers tend to pay higher monthly charges. A similar tendency is observed between Partner and Dependents, which is expected by demographic characteristics.

12.3 ROC Curve Analysis

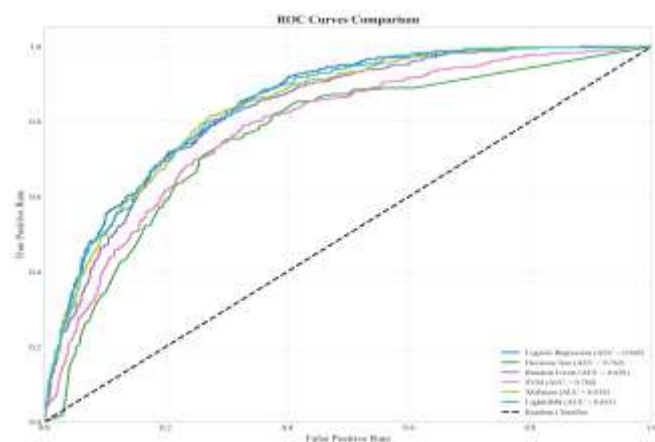


Figure 3. ROC Curve Comparison

The ROC curve was utilized to compare the accuracy of classifying all six machine learning algorithms. The ROC curve shows how well the True Positive Rate (or Sensitivity) versus the False Positive Rate behaves throughout the threshold levels for classification.

When considering all the models presented, it was found that LightGBM provided the best ROC-AUC value, thus indicating the effectiveness of the model in differentiating between churn and non-churning customers. The ROC curves from XGBoost and Random Forest were also close enough. With regard to Logistic Regression, Decision Tree, and SVM, their ROC characteristics were much worse as compared to others.

In conclusion, the ROC analysis indicates that the use of ensemble learning algorithms is superior to standard machine learning techniques when predicting customer churn.

12.4 Precision–Recall Curve Analysis

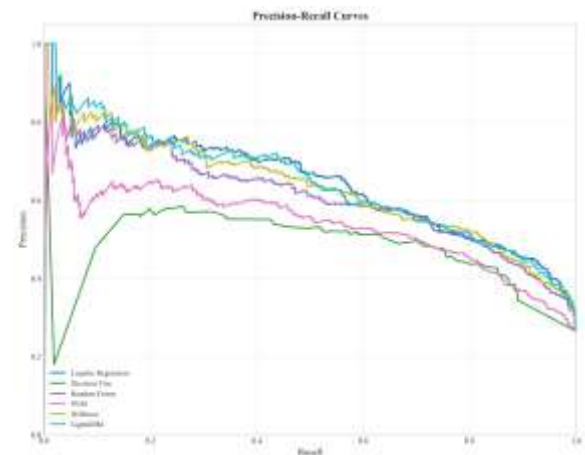


Figure 4. Precision–Recall Curve

The Precision-Recall curve measures the effectiveness of a model in finding churn customers, especially when there is class imbalance in data. Precision indicates the ratio of successful churn predictions with the total churn predictions while Recall is defined as the ability of the model to detect true churn customers.

Outcomes show that LightGBM had the best trade-off in Precision and Recall with the biggest area of Precision-Recall compared to other models. XGBoost and Random Forest achieved good results and Logistic Regression and Decision Tree performed worse with higher Recall.

This shows that gradient boosting methods have better potential to locate churn customers while reducing the number of errors.

12.5 Feature Importance Analysis

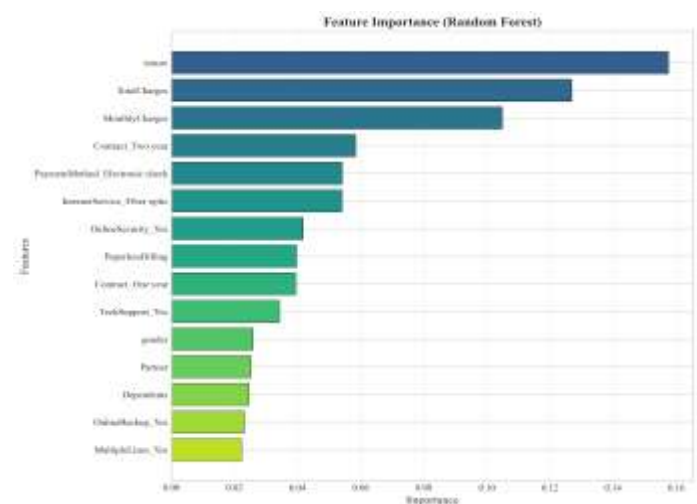


Figure 5. Feature Importance

Ensemble learning algorithms were used to identify the most important aspects of customers in predicting churn.

The analysis concluded that Contract Type, Tenure, Monthly Charges, Total Charges, Internet Service, Payment Method, and Online Security have the highest degree of influence on customer churn.

12.6 SHAP Summary Analysis

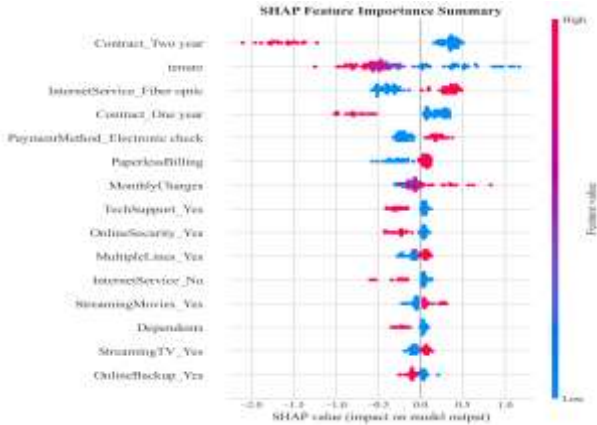


Figure 6. SHAP Summary Plot

In order to enhance the interpretability of this model, SHapley Additive exPlanations (SHAP) were employed on the best performing model in the experiment, the LightGBM. The results of the SHAP Summary Plot demonstrate the individual contribution of a particular feature to the prediction outcome.

This analysis reveals that contributing to the churn prediction are Contract Type, Tenure, Monthly Charges, Total Charges and Internet Service.

The method presented by SHAP is distinguished from traditional feature importance measure in that SHAP explains the contribution of individual features for every specific client, which guarantees the interpretability and trustworthiness of the prediction model.

Table 12.1 Model Performance Comparison.

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.7460 77	0.5134 65	0.7708 89	0.6163 79	0.8452 92
Decision Tree	0.7439 37	0.5132 74	0.6253 37	0.5637 91	0.7622 06
Random Forest	0.7810 27	0.5788 18	0.6334 23	0.6048 91	0.8283 21
SVM	0.7460 77	0.5164 11	0.6361 19	0.5700 48	0.7835 80

XGBoos	0.7696	0.5530	0.6738	0.6075	0.8376
t	15	97	54	33	74
LightG		0.5733	0.6738	0.6195	0.8429
BM	0.7810	94	54	79	63
	27				

Discussion:

As shown in the experimental results, LightGBM had the highest overall performance of all the evaluated ML algorithms, mainly due to its leaf-wise tree development, histogram-based learning, and ability to model the nonlinear relationships between the customers' features. Furthermore, LightGBM has low resource requirements maintaining an acceptable prediction accuracy, thus being appropriate for large-scale customer churn prediction applications.

While both XGBoost and Random Forest performed well, LightGBM can bring better results in terms of different evaluation metrics throughout learning and predicting.

As for Logistic Regression and Decision Tree, they were used as baseline models, which demonstrated less ability to describe the complicated relationships between the customers' characteristics.

The analysis based on the attributes' importance as well as SHAP clearly demonstrated that Contract Type, Tenure, Monthly Charges, and Internet Service are among the main drivers of customer churn. These results are consistent with previous papers claiming that contract length, service use, and billing features are major predictors of the customers' churn.

Moreover, the application of SHAP increased the explainability of the model, therefore overcoming a challenge described by several authors in their work concerning a lack of transparency of the prediction models. The overall structure of the proposed framework gives a good combination of prediction efficiencies, explanatory ability, and practical usability of the Flask-based application. According to the results, the proposed framework is capable of assisting organizations in finding highly susceptible customers, building targeted retention schemes, and making data-based business decisions.

13. Conclusion:

This study laid out an advanced framework for predicting customer churn through machine learning. Six algorithms were compared in this study, including logistic regression, decision tree, random forest, support vector machine (SVM), XGBoost, and LightGBM using the IBM Telco Customer Churn Dataset. In

this study, the LightGBM algorithm outperforms in terms of accuracy, efficiency, and reliability. In addition to this, the integration of SHapley Additive exPlanations (SHAP) enables us to understand the underlying factors for customer churn. The introduction of a Flask web application also allows us to showcase the real-life implications of the framework. Thus, the results of this research show that the framework demonstrates great potential in the area of customer churn prediction processes.

14. References:

- [1] A. Sikri, R. Sharma, and P. Gupta, "Improving Customer Retention in Telecom Using Machine Learning-Based Churn Prediction," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 4, pp. 1256–1268, 2024.
- [2] Y. Suh, "Machine Learning-Based Customer Churn Prediction in Home Appliance Rental Business," *Applied Sciences*, vol. 13, no. 15, pp. 1–18, 2023.
- [3] V. Chang, H. Alenezi, and M. Ramachandran, "Customer Churn Prediction in the Telecommunication Industry Based on Machine Learning Models," *IEEE Access*, vol. 12, pp. 32584–32601, 2024.
- [4] M. AbdelAziz, A. Hassan, and S. Ibrahim, "Comparison of Machine Learning and Deep Learning Models for Customer Churn Prediction," *Expert Systems with Applications*, vol. 256, pp. 1–18, 2025.
- [5] R. Bhuria, A. Verma, and P. Singh, "Ensemble Approach to Customer Churn Prediction Using Machine Learning Techniques," *Journal of King Saud University – Computer and Information Sciences*, vol. 37, no. 2, pp. 102345, 2025.
- [6] M. Imani, H. Rahimi, and S. Mohammadi, "Customer Churn Prediction: A Systematic Literature Review," *Artificial Intelligence Review*, vol. 58, no. 3, pp. 1–39, 2025.
- References:
- [7] M. Abdelhady, A. El-Sayed, and M. Salem, "Utilizing Artificial Intelligence for Predictive Customer Churn Modeling," *Journal of Big Data*, vol. 12, no. 1, pp. 45–63, 2025.
- [8] R. Agrawal, S. Gupta, and V. Sharma, "Hybrid SMOTE and GAN-Based Customer Churn Prediction Using LightGBM," *IEEE Access*, vol. 14, pp. 21456–21473, 2026.
- [9] A. El Attar, M. Hassan, and H. Ibrahim, "Explainable AI-Driven Customer Churn Prediction Using Ensemble Learning," *Expert Systems with Applications*, vol. 271, pp. 126845, 2026.
- [10] J. Maan and P. Kaur, "Customer Churn Prediction Model Using Explainable Machine Learning Techniques," *International Journal of Information Management Data Insights*, vol. 3, no. 2, pp. 100210, 2023.
- #### Dataset Reference
- [11] IBM Corporation, "Telco Customer Churn Dataset," IBM Sample Data Sets. Available: <https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/>. Accessed: Jul. 2026.
- #### Software and Libraries
- [12] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.
- [14] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [16] A. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed. Sebastopol, CA, USA: O'Reilly Media, 2018.
- #### Additional Supporting References
- [17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[20] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.