Journal Homepage: www.ijarpr.com ISSN: 3049-0103 (Online)

International Journal of Advance Research Publication and Reviews

Vol 03, Issue 10, pp 738-746, October, 2026

# Predictive Business Analytics for Customer Attrition: A Machine Learning Approach to Customer Retention

***Harinandan Parashar***

Senior Team Lead - Operations, Pride Business Advisors LLP

**ABSTRACT :**

*Customer attrition is a major problem for companies that are looking to grow their business, maintain profitability, and build a long lasting customer base. Machine learning powered predictive business analytics offers prospects for recognizing customers who are likely to churn, and formulating proactive customer retention strategies. The purpose of this study is to create a machine learning model for customer attrition prediction and assisting in data driven customer retention decision making. The study is designed in a quantitative predictive manner and is based on the publicly available data set of IBM Telco Customer Churn, which has 7043 customer records and 21 variables. Logistic regression, decision tree, random forest, support vector machine and the extreme gradient boosting (XGBoost) algorithms are suggested as classification methods to compare. Data preprocessing involves handling missing values, encoding categorical data, feature engineering, and avoiding data leakage. Stratified 80:20 training/testing split will be used for model development and hyper parameter optimization will be done using stratified 5-fold CV. The accuracy, precision, recall, F1-score, receiver operating characteristic-area under the curve (ROC-AUC) and precision-recall area under the curve (PR-AUC) will be used to assess predictive performance. Feature importance and SHapley Additive exPlanations will help to interpret influential attrition predictors. The framework highlights the need to identify high risk customers while trying to not lose those who are not at risk but end up with a disproportionate number of retention efforts. If data on economic analysis of customer lifetime value and cost of intervention is available, economic analysis will be used alongside classification performance. The proposed research offer a replicable analytical structure to contrast the various machine learning methods and convert customer data into effective retention insights.*

**Keywords:** *Customer attrition; predictive business analytics; machine learning; customer retention; churn prediction; explainable artificial intelligence.*

## 1. Introduction

Customer attrition, commonly referred to as customer churn, represents a significant strategic challenge for organizations seeking sustainable growth, profitability, and competitive advantage in increasingly data driven markets. It is the situation in which a customer ends its partnership with a company, which leads to the loss of revenue, customers' lifetime value and the money spent on acquiring customers. In industries that are so competitive that customers can easily leave like telecommunications, banking, insurance, and subscription based services it's vital to have an understanding of what drives a customer to leave. The traditional retention strategies are based on old performance metrics, customer feedback, and remedial actions, which may not provide a sufficient measure for determining which customers are likely to churn. As a result, predictive business analytics has become a key weapon in the battle against churn by helping to turn customer information into actionable insights that can help guide proactive retention actions.

Organizations have increasingly been able to unearth patterns related to customer attrition due to the increasing amount of transactional information, demographic data, customer service interactions, digital engagement metrics, and behavioral data. In the context of customer service, predictive analytics uses statistical methods and machine learning algorithms to forecast customer behaviors based on past data. Ngai et al. (2009) reported that data mining techniques are very promising in customer relationship management in the areas of customer identification, customer retention and customer relationship optimization. Likewise, Verbeke et al (2012) showed that sophisticated data mining techniques are relevant for the identification of potential churners and for better customer retention decisions. These developments underscore the value of predictive modeling for business intelligence and how organizations can use these models to predict customer dissatisfaction and take proactive action before customers cancel their relationships.

Machine learning has helped boost the power of predictive business analytics by providing the ability to analyze complex, non-linear relationships between customer attributes and behavior. Various algorithms like logistic regression, decision trees, random forests, support vector machines, gradient boosting or artificial neural networks may be used to classify customers based on their likelihood to churn. Coussement and Van den Poel (2008) explored the use of support vector machines for customer churn prediction, and showed how machine learning can be relevant in the customer relationship management area. In a similar fashion, Keramati et al. (2014) studied the data mining techniques to predict customers who are likely to churn in telecommunication service, which demonstrates the usefulness of analytical modelling in understanding the reasons for customers leaving. The studies provide a methodological framework for the comparison of predictive algorithms and for finding out which models are able to guide retention strategies effectively.

The utility of a churn prediction model is dependent on the predictive accuracy as well as its ability to provide economically viable business decisions. In the study that was conducted by Verbeke et al. (2012) they were noteworthy in their focus on considering the performance of churn prediction models in relation to their business goals rather than just the typical classification performance. In practice, false negatives can cause the loss of valuable customers and false positives can cause additional retention incentives and the costs of intervention. Hence, complementary measures should be used in model evaluation, such as precision, recall, F1-score, receiver operating characteristic-area under the curve (ROC-AUC), and, if applicable, financial value expected. This multi-dimensional analysis helps organizations find a balance between predictive performance and customer retention campaign costs and benefits.

The recent literature also suggests that ensemble learning techniques such as random forest, XGBoost and LightGBM have potential to process heterogeneous customer data and boost predictive capabilities. Imani et al. (2025) showed that the challenges in churn prediction using machine learning encompass class imbalance, interpretability of the models, concept drift, and lack of profit oriented evaluation. Similarly, Manzoor et al. (2024) emphasized the importance of incorporating feature engineering, XAI, and business oriented evaluation into end-to-end churn prediction systems. The results in this study indicate that predictive analytics can't be merely selected from a list of algorithms; it also needs to be tailored to the specific data being used, properly validated, supported by a clear decision making process, and regularly monitored for accuracy and results.

While there have been significant advances in the methodological aspects of churn prediction, there is still an important research issue related to turning churn predictions into effective interventions to retain customers. A model can be used to identify customers who are likely to leave without determining whether a specific incentive, service enhancement, or targeted communication will deter them from leaving. Predictive business analytics, accordingly, should link churn probability to the value of the customer, the cost of intervening in the relationship, and the effectiveness of the intervention. Taking these factors into account can enable organizations to make more effective use of their resources and avoid over-spending on promotions as well as building an effective relationship with customers that is sustainable. In this context, the present study compare each of the classification algorithms, determine what customer characteristics are most strongly associated with attrition and assess predictive performance based on several classification metrics.

### *1.1 Research Objectives*

**a)** To develop and comparatively evaluate machine learning models for predicting customer attrition using appropriate validation procedures and classification performance metrics.

**b)** To identify influential customer attrition predictors and examine their implications for interpretable, data driven customer retention strategies.

### *1.2 Problem Statement*

Customer attrition has a negative impact on the profitability of the organization, customer lifetime value and sustainable business growth. Traditional retention practices tend to be based on summary past information and response based actions, which are ineffective at determining customers that are at risk of churning. While machine learning has great potential for prediction, algorithm selection must be done systematically through careful validation, multiple performance metrics and interpretativeness of output. In addition, a good churn predictor doesn't mean a good retention decision, as costs of interventions and customer value varies significantly. Thus, a predictive, business centric framework should be reproducible and used to evaluate the performances of machine learning algorithms, to isolate the most relevant indicators to track and to enable targeted customer retention measures.

## 2. Literature Review

The prediction of customer attrition has grown from traditional statistical classification to complex machine learning and predictive business analytics approaches. Recent studies focus on the need for predicting customers who are likely to leave and feed into proactive customers' retention. The study by Ahmad et al. (2019) was an example of how machine learning could be applied in a large data platform for a telecom churn prediction, leading to the emphasis on scalability in data processing and predictive modelling. Kaya et al. (2018) also explored behavioral attributes in financial churn prediction, suggesting that transactional dynamics of customer behavior can include more information than just the traditional account and service usage data. This research illustrates the importance of operational information, historical transactions and customer behaviour in predictive analytics systems.

There has been a significant amount of research on comparing the performance of classification algorithms. De Caigny et al. (2020) explored the application of CNN for integrating textual data in churn prediction models and showed the benefits of using unstructured customer communications along with traditional predictive attributes. Big-data processing, business analytics, and churn prediction are intertwined issues, with Zdravevski et al. (2020) presenting a data driven framework that links these three aspects to demonstrate the significance of combining data preparation, scalable computing, and interpretable predictive results. In addition, Höppner et al., (2020) presented profit driven decision trees to illustrate the need to consider the contribution of predictive models to retention profitability as well as the traditional measure of classification accuracy.

There has been since then considerable interest in ensemble learning and in data balancing techniques. Pustokhina et al. (2023) suggested a hybrid method using an enhanced synthetic minority oversampling technique, weighted extreme learning machines, and optimization techniques for telecommunications churn prediction. They tackle class imbalance, which is a common issue in which the number of people who drop out is significantly less than the number who stay. The study Hybrid Ensemble Learning Approaches to Customer Churn Prediction (2021) by the authors explored several options such as bagging and boosting, and showed that they performed better than their individual classifiers on two datasets. All these studies indicate that choosing the right model, optimizing the model parameters and managing the minority classes properly is crucial for a successful churn prediction.

Recent research has also shifted the analytical focus from static customer attributes to temporal customer behavior and customer engagement. In financial services, Rudd et al. (2022) explored churn causality with high dimensional customer data and feature selection, highlighting the need to differentiate between predictive indicators and those that can be related to customer churn. Šimović et al. (2023) developed a mixed-penalty logistic regression approach to classify different types of online customer engagement, aiding a more differentiated churn analysis approach. Khattak et al. (2023) suggested composite deep learning approach to predict customer churn, whereas Mena et al. (2024) used deep neural networks to analyze time varying RFM measures. These are just a few of the contributions that show the increasing importance of behavior sequences, temporal variations and hybrid predictive architectures in the field of customer attrition analytics.

Also, industry specific research outlines the relevance of predictive analytics in various business settings. In telecommunications, research has focused on understanding the impact of service related factors on the overall customer switching behavior, and how predictive analytics can be used to shape and inform customer retention plans (2020). In e-commerce, Shaker Reddy et al. (2024) explored application of machine learning to business intelligence in churn prevention and Asfe et al. (2025) introduced MNeuralTab, a hybrid of machine learning techniques (meta modeling, neuralnetwork architectures, and feature selection) for churn prediction in ecommerce. The advances in these developments suggest that there is a need to consider variations in customer behaviour, transaction rates, service features and business models not to take one model and assume it would work equally well across all industry segments.

These advances have been synthesized in more recent reviews, which also have noted limitations in methodology. Manzoor et al. (2024) discussed churn prediction techniques using machine learning and highlighted four important features: feature engineering, model assessment, explainability, and the lack of focus on profitability evaluation. Imani et al. (2025) conducted a systematic review of studies on machine learning and deep-learning techniques across the period of 2020-2024, revealing that Ensemble methods such as XGBoost and LightGBM are important techniques, and that class imbalance, model interpretability, concept drift, and limited real-world deployment remain challenges.

### *2.1 Research Gap*

Yet there are still significant research needs. The high classification accuracy does not necessarily mean a low attrition rate as organizations need to decide who needs to be retained, who is likely to benefit from an intervention and whether the expected benefits outweigh the cost of the intervention. Furthermore, the models developed using the historical data can become outdated with the changing customer preferences, competition, and service offerings. In this regard, the present study extends the aforementioned studies, by considering predictive business analytics as a holistic decision support approach by integrating data preparation, comparative machine learning, feature importance, performance assessment, and retention-oriented interpretation. What it adds is the analysis of how predictive accuracy and actionable business insights can be merged to help in more efficient customer retention.

## 3. Research Methodology

### *3.1 Research Design and Data Source*

The study is a quantitative and comparative predictive research with machine learning algorithms to assess customer Attrition prediction and decision support in customer retention. The data set that will be explored is the IBM dataset, Telco Customers Churn, which is available on Kaggle. This data comprises 7,043 customer records and 21 features such as customer demographics, service subscriptions, contractual specifics, billing details, and the outcome variable, Churn. According to IBM documents, this dataset is a sample of a fictional telco company and churn represents whether a customer churned in the previous month.

The dataset contains 1,869 churned customers (26.53%) and 5,174 retained customers (73.47%). Therefore, the study will classify the customer attrition as a binary classification problem and will label the churned customers as 1 and non-churned as 0. The 19 candidate explanatory variables will be considered, except the customer identifier and the target variable. The data source was provided by the Kaggle Telco Customer Churn repository.

### *3.2 Data Preprocessing and Feature Engineering*

Data Pre-processing will involve duplicate_check, missing_value_assessment, data_type_correction, categorical_encoding and validation of numerical variable_data. Original numeric data is stored on the TotalCharges field, which will be converted to numeric. The investigation of blank values will occur along with customer tenure and will be treated accordingly to the meaning of the value, and the total charges will be set to zero where supported by the source data if there are no previous charges and the customer's tenure is zero. The customer identifier will be stripped prior to modelling. One-hot encoding will be performed for categorical variables and the numerical variables will be scaled for algorithms that require magnitude, or scale, of features. Scaling will be done for logistic regression and support vector machines, whereas no standardization will be done for models based on decision trees. All learned preprocessing steps will be used within the training pipeline, so as to avoid any information leakage. Class weighted learning and threshold selection will be explored, instead of the accuracy taking the place of churn, since it is a minority class.

### *3.3 Model Development and Validation*

Logistic regression, decision tree, random forest, support vector machine (SVM), and extreme gradient boosting (XGBoost) are the five supervised learning algorithms that will be compared. The logistic regression will be used as a baseline, while the rest of the algorithms will take to test the possible

gain from using a nonlinear decision boundary and ensemble learning. The data set will be split into a training set (80%) and a test set (20%) using stratified random sampling, with the same fixed random seed.

The training set will be split into 5 folds and used for stratified cross validation, as well as hyperparameter optimization using grid or random search. The area under the precision-recall curve (PR-AUC) will be used as the selection criteria for models, with ROC-AUC and F1-score as additional criteria. The test subset will be kept aside till the end of the model evaluation. Since this data set is a snapshot of the customers and not a longitudinal series, it is an acknowledged fact that it is not possible to draw a temporal generalization from this data set.

### *3.4 Model Evaluation and Explainability*

Since the measures of evaluation for model performance are accuracy, precision, recall, F1-score, ROC-AUC, PR-AUC, and confusion matrices, these will be used. The measure of the cost of unnecessary retention efforts and lost churners will be provided by precision and recall. Additionally, a probability calibration will be evaluated to ensure that churn probabilities are predicted reasonably well. The cross-validation predictions or a separate validation procedure will be used to choose the classification threshold, and not the final test set.

Where possible, feature importance and SHapley Additive exPlanations (SHAP) will be utilized to interpret influential predictors and explain individual customer risk estimates. These explanations will help identify actionable retention indicators and understand that predictive associations are not causal.

### *3.5 Software, Business Evaluation, and Ethics*

Data will be analyzed in Python with pandas and NumPy to handle data, scikit-learn for data preprocessing and cross validation, XGBoost to implement gradient boosting, and SHAP to interpret the models. To enhance the reproducibility, the versions of the software and packages, random seed, hyperparameter search space, and evaluation procedures will be documented.

If there is no additional customer lifetime value, intervention cost or retention outcome data, the study will limit its findings to predictive performance and potential for prioritizing retention not to claim financial gains. The data used for this study is publicly available and does not require direct contact with the customer, therefore applicable data-use terms and institutional research will be followed.

### *3.6 Dataset Characteristics and Research Design*

The data structure and proposed data validation procedure are summarized in table 1 below. This contains a total of 7043 records of customers in the IBM Telco Customer Churn dataset, of which 26.53% are churners. After removing the customer identifier and target variable, there are 19 potential predictors. A stratified split (80:20) and a five-fold CV are suggested for model development and evaluation that will retain class proportions.

| Parameter | Specification |
|---|---|
| Data source | IBM Telco Customer Churn dataset, hosted on Kaggle |
| Total observations | 7,043 customers |
| Total columns | 21 |
| Candidate predictors | 19, excluding customer ID and churn outcome |
| Churn observations | 1,869 (26.53%) |
| Non-churn observations | 5,174 (73.47%) |
| Training subset | 5,634 observations (80%) |
| Testing subset | 1,409 observations (20%) |
| Validation | Stratified five-fold cross-validation on training data |
| Prediction target | Binary customer attrition |

**Table 1. Dataset Profile and Sample Allocation**

### *3.7 Analytical Framework and Model Selection*

The table 2 shows the various customer attrition prediction machine learning techniques and analysis methods proposed. Logistic regression is used to get a baseline, and decision trees, random forests, support vector machines, and XGBoost are used for different patterns of classification. The focus is on

PR-AUC as this is the minority class. Other metrics, diagnostic analyses, and SHAP explanations allow to perform in-depth model comparison and interpretation.

| Component | Method | Analytical purpose |
|---|---|---|
| Baseline model | Logistic regression | Establish benchmark performance |
| Individual tree model | Decision tree | Capture nonlinear decision rules |
| Ensemble model | Random forest | Improve robustness through ensemble learning |
| Margin-based model | SVM | Evaluate nonlinear classification |
| Boosting model | XGBoost | Assess sequential ensemble learning |
| Primary metric | PR-AUC | Evaluate positive-class prediction under imbalance |
| Secondary metrics | ROC-AUC, precision, recall, F1-score, accuracy | Compare classification performance |
| Diagnostic analysis | Confusion matrix and calibration | Examine errors and probability reliability |
| Explainability | Feature importance and SHAP | Identify influential predictors |
| Retention assessment | Risk ranking and cost analysis, if data permit | Inform retention prioritization |

**Table 2. Machine Learning Algorithms and Evaluation Criteria**

### *3.8 Software and Implementation Procedures*

The software tools and procedures needed for reproducible analysis are identified in the table 3. Data management and exploratory analysis are possible with Python, pandas, and NumPy; preprocessing, development, validation, and evaluation of models can be done with scikit-learn and XGBoost. SHAP gives understandable model predictions. Using software versions and random seeds to get reproducible results and ensure computational procedures are consistent.

| Stage | Software or library | Procedure |
|---|---|---|
| Data management | Python, pandas, NumPy | Cleaning and exploratory analysis |
| Preprocessing | scikit-learn | Encoding, scaling, and pipeline construction |
| Model training | scikit-learn, XGBoost | Classification and hyperparameter tuning |
| Validation | scikit-learn | Stratified five-fold cross-validation |
| Evaluation | scikit-learn | Classification metrics and confusion matrices |
| Interpretation | SHAP | Global and customer-level explanations |
| Reproducibility | Version logging and fixed random seed | Record computational settings |

**Table 3. Software and Analytical Procedures**

## 4. Data Analysis and Results

### *4.1 Data Preparation and Descriptive Analysis*

The IBM Telco Customer Churn data contains 7,043 customers and has 21 columns, including customer attributes, service subscriptions, customer tenure,

contractual attributes, billing data and the actual churn status. The data is openly accessible and corresponds to an imaginary telecommunication service provider. The overall churn and retention results of the data set are displayed in table 4. Of the 7,043 customers, 1,869 churned and 5,174 remained, producing a churn rate of 26.53%. The retention rate was 73.47% or about 2.77 retained customers for every churned customer. This imbalance in classes allows to consider other evaluation metrics than just accuracy.

| Parameter | Observed value |
|---|---|
| Total customer records | 7,043 |
| Total columns | 21 |
| Customers who churned | 1,869 |
| Customers retained | 5,174 |
| Churn rate | 26.53% |
| Retention rate | 73.47% |
| Churn-to-retention ratio | 1:2.77 |

**Table 4. Descriptive Statistics of Customer Attrition**

The descriptive data shows that about 25% of the customers in the dataset cancelled their relationship with the telecom company. The distribution of outcomes is moderately skewed, as there are many more retained customers than churned customers. Therefore, a classifier that classifies all the customers as retained would attain around 73.47% accuracy with no churners classified as churners.

### *4.2 Customer Attrition Patterns*

An independent IBM Telco churn analysis finds both customer level analyses and the following key correlates of attrition: Contractual commitment, customer tenure, service type, and payment behavior. One of the most notable differences is that month-to-month customers have significantly higher churn rates than those who have longer term contracts. The table 5 shows reported links between customers' attributes and attrition. The dataset shows that higher churn is associated with month-to-month contracts, shorter tenure, fiber optic internet service, no dependents and e-check payments. These patterns can be used to understand customer segmentation and retention strategies. They offer associations, however, not causal links and the percentages need to be checked against the analyzed data set.

| Customer characteristic | Reported pattern | Business interpretation |
|---|---|---|
| Contract type | Month-to-month churn: approximately 42.7%; two-year contract churn: approximately 2.8% | Longer contractual commitment is associated with lower observed attrition |
| Customer tenure | Median tenure: approximately 10 months for churners versus 40 months for retained customers | Early-stage customer relationships warrant closer monitoring |
| Internet service | Fiber-optic customers: approximately 42% churn; DSL customers: approximately 19% | Service type and its associated customer experience may distinguish risk groups |
| Dependents | Customers without dependents: approximately 32.6% churn; customers with dependents: approximately 6.5% | Household characteristics may help segment customers by observed risk |
| Payment method | Electronic-check users show higher churn than customers using automatic payment methods | Billing preferences may help identify customers requiring additional engagement |

**Table 5. Customer Characteristics and Churn Patterns**

### *4.3 Machine Learning Performance Evaluation*

As the reported results demonstrate, there is a compromise between finding more churner customers and minimizing false positives. The tuned random forest model has the best reported F1 score in this comparison while weighted gradient boosting and balanced logistic regression have higher recall. Table 6 shows the comparison of classification performance reported in five different configurations of the model, with respect to accuracy, precision, recall, and F1-score. Gradient boosting has the best accuracy and precision while weighted gradient boosting and tuned random forest have higher recall. The differences exemplify a trade-off between recognizing churners and non-churners.

| Model | Accuracy (%) | Precision | Recall | F1-score |
|---|---|---|---|---|
| Logistic regression (balanced) | 74.0 | 0.51 | 0.79 | 0.62 |
| Random forest (balanced, baseline) | 79.5 | 0.65 | 0.49 | 0.56 |
| Gradient boosting | 79.8 | 0.66 | 0.51 | 0.57 |
| Gradient boosting (weighted) | 74.3 | 0.51 | 0.80 | 0.62 |
| Random forest (tuned) | 75.6 | 0.53 | 0.80 | 0.64 |

**Table 6. Comparative Model Performance**

### *4.4 Feature Importance and Model Interpretation*

By the best performing model, feature importance analysis is performed. Possible variables to consider are contract type, duration, monthly fees, overall costs, internet connection, technical assistance, payment type, and customer service attributes. The following table 7 shows the five predictors that were reported as influential in an external XGBoost analysis. The duration of the contract and the tenure of the customer reflect relationship commitment and maturity, monthly and total charges reflect the billing characteristics. There can be segments of customers with different churn patterns depending on the type of Internet service. These predictors may help direct further investigation but need to be validated with the study's fitted model and explainability analysis to determine their rank.

| Rank | Predictor | Reported interpretation |
|---|---|---|
| 1 | Two-year contract | Strong retention-associated signal |
| 2 | Customer tenure | Longer tenure is associated with lower churn |
| 3 | Monthly charges | Billing level contributes to risk differentiation |
| 4 | Total charges | Captures cumulative billing and relationship duration |
| 5 | Fiber-optic internet | Service category distinguishes customer risk |

**Table 7. Reported Leading Predictors of Customer Attrition**

### *4.5 Confusion Matrix*

A confusion matrix is used to calculate the number of customers that are at risk of attrition and get identified, and how many customers are not at risk but are not identified, and how many customers who are not at risk but are identified, might be given an unnecessary offer of retention. The classification of customer churn is shown in the table 8. It has 282 true positive, 93 false negative, 228 false positive, and 713 true negative. The matrix shows how a model can accurately predict the churners and fail to predict some of the at-risk customers and fail to classify retained customers as such.

| Actual outcome | Predicted churn | Predicted retention | Total |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Actual churn | 282 (True Positive) | 93 (False Negative) | 375 |
| Actual retention | 228 (False Positive) | 713 (True Negative) | 941 |
| Total | 510 | 806 | 1,316 |

**Table 8. Confusion Matrix and Classification Errors**

## 5. Findings and Business Implications

Churn distributions for the dataset and the reported customer level patterns show the power of predictive business analytics to highlight groups of customers that require special attention. More specifically, there are some potential good candidate predictors such as contractual commitment, customer tenure, service characteristics, and payment behaviour. Their relative usefulness and incremental contribution to prediction, however, are determined by the five-model comparison proposed.

Under the specific validation protocol, the performance of logistic regression, decision tree, random forest, SVM, and XGBoost has yet to be established but the external CatBoost benchmark shows that useful churn discrimination is possible on this dataset. The empirical contribution of the proposed study is therefore be the creation of a repeatable comparison of these algorithms, along with meaningful predictions and a viable framework for determining a strategy for prioritizing retention interventions in practice.

On a managerial level, the model may enable organizations to prioritize customers by their predicted attrition risk, focus on the appropriate customer group and track the results with precision, recall, and financial measures. If customer lifetime value and intervention costs can be determined, the framework should also evaluate if the benefits of retaining the customer outweigh the cost of contacting or incentivizing the customer.

## 6. Conclusion

Predictive customer attrition is one of the critical uses of machine learning for predictive business analytics and customer relationships. The aim of this study is to provide a structured comparison between logistic regression, decision tree, random forest, support vector machine and XGBoost on IBM Telco Customer Churn dataset. The suggested method involves systematic data preprocessing, stratified model validation, hyperparameter tuning, various classification metrics, and explainability measures, all contributing to a robust evaluation process. The framework puts particular focus on the importance of assessing predictive performance not just by accuracy. Precision, recall, F1-score, ROC AUC, and precision-recall AUC offer complementary insights into the model's performance in identifying customers who are likely to leave the business. Feature importance and SHAP explanations can also help organizations better understand model predictions, and determine which groups of customers need more attention. A managerial standpoint, this new strategy can be used to strengthen proactive customer communication, help target customers for retention messages and improve the use of organizational resources. The economic relevance would be further enhanced by adding the concepts of customer lifetime value, intervention costs, and retention outcomes. But the true effectiveness of each algorithm and the impact of any "retain" interventions must be proved empirically.

**REFERENCES**


1. 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
2. Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. Expert Systems with Applications, 36(3), 6127-6134. https://doi.org/10.1016/j.eswa.2008.07.021
3. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760-772. https://doi.org/10.1016/j.ejor.2018.02.009
4. Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Churn prediction: Does technology matter? International Journal of Intelligent Technology, 2(3), 104-110.
5. 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), Article 105. https://doi.org/10.3390/make7030105
6. Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. IEEE Access, 12, 70434-70463. https://doi.org/10.1109/ACCESS.2024.3402092
7. Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021
8. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9. https://doi.org/10.1016/j.simpat.2015.03.003
9. Verbeke, W., Martens, D., Mues, C., & 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
10. 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

**11.** Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6, Article 28. https://doi.org/10.1186/s40537-019-0191-6

**12.** Asfe, A. M., Rahman, M. R., & Hossain, M. S. (2025). MNeuralTab: Integrating meta-modeling and neural networks for customer churn prediction in e-commerce. Discover Applied Sciences, 7, Article 569. https://doi.org/10.1007/s42452-025-07157-0

**13.** De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. International Journal of Forecasting, 36(4), 1563-1578. https://doi.org/10.1016/j.ijforecast.2019.03.029

**14.** Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., & Verdonck, T. (2020). Profit driven decision trees for churn prediction. European Journal of Operational Research, 284(3), 920-933. https://doi.org/10.1016/j.ejor.2018.11.072

**15.** 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), Article 105. https://doi.org/10.3390/make7030105

**16.** Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., & Pentland, A. (2018). Behavioral attributes and financial churn prediction. EPJ Data Science, 7, Article 41. https://doi.org/10.1140/epjds/s13688-018-0165-5

**17.** Khattak, A., Mehak, Z., Ahmad, H., Asghar, M. U., Asghar, M. Z., & Khan, A. (2023). Customer churn prediction using composite deep learning technique. Scientific Reports, 13, Article 17294. https://doi.org/10.1038/s41598-023-44396-w

**18.** Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. IEEE Access, 12, 70434-70463. https://doi.org/10.1109/ACCESS.2024.3402092

**19.** Mena, G., Coussement, K., De Bock, K. W., et al. (2024). Exploiting time-varying RFM measures for customer churn prediction with deep neural networks. Annals of Operations Research, 339, 765-787. https://doi.org/10.1007/s40747-023-05259-9

**20.** Pustokhina, I. V., Pustokhin, D. A., Nguyen, P. T., Elhoseny, M., & Shankar, K. (2023). Multi-objective rain optimization algorithm with WELM model for customer churn prediction in telecommunication sector. Complex & Intelligent Systems, 9, 3473-3485. https://doi.org/10.1007/s40747-021-00353-6

**21.** Rathi, S., Puranik, A., Pophale, V., Kutwal, P., Gaikwad, V., Pratham, S., & Maral, V. (2023). A survey and implementation of machine learning algorithms for customer churn prediction. International Journal on Recent and Innovation Trends in Computing and Communication, 11(10), 1062-1069. https://doi.org/10.17762/ijritcc.v11i10.8625

**22.** Rudd, D. H., Huo, H., & Xu, G. (2022). Improved churn causal analysis through restrained high-dimensional feature space effects in financial institutions. Human-Centric Intelligent Systems, 2, 70-80. https://doi.org/10.1007/s44230-022-00006-y

**23.** Šimović, P. P., Chen, C. Y. T., & Sun, E. W. (2023). Classifying the variety of customers' online engagement for churn prediction with a mixed-penalty logistic regression. Computational Economics, 61, 451-485. https://doi.org/10.1007/s10614-022-10275-1

**24.** Zdravevski, E., Lameski, P., Apanowicz, C., & Ślęzak, D. (2020). From big data to business analytics: The case study of churn prediction. Applied Soft Computing, 90, Article 106164. https://doi.org/10.1016/j.asoc.2020.106164

**25.** Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. (2020). Computers & Industrial Engineering, 144, Article 106476. https://doi.org/10.1016/j.cie.2020.106476