Predicting Customer Churn with Supervised Learning: Model Comparison and Feature Analysis

Mohammad Atabaki
University of Bologna
Digital Transformation Management
Email:mohammad.atabaki@studio.unibo.it
Matricola: 0001116620

Shahab Aminraoufpour
University of Bologna
Digital Transformation Management
shahab.aminraoufpour@studio.unibo.it
Matricola: 0001120850

***Abstract*—Customer churn presents a significant challenge for subscription-driven companies, making reliable churn prediction crucial for designing effective retention initiatives. This project leverages machine learning classification methods on a structured customer dataset that includes demographic, behavioral, and subscription-related attributes. Following data preprocessing and exploratory analysis, several classification models are trained and assessed using accuracy, precision, recall, and F1-score as evaluation metrics. The findings show that tenure, usage frequency, customer support interactions, payment delays, and contract details are among the most influential factors in predicting churn. Although the study is constrained by dataset limitations and the lack of labels in the test set, it still illustrates the practical usefulness of machine learning in pinpointing customers at risk of leaving and guiding data-driven retention strategies..**

1. Introduction

Customer churn is a critical challenge for subscription-based and service-oriented businesses, as retaining existing customers is typically more cost-effective than acquiring new ones. High churn rates can lead to significant revenue loss and indicate underlying issues related to service quality, pricing, or customer engagement. As a result, the ability to identify customers who are likely to discontinue their relationship with a company has become an important application of data-driven decision making.

This project addresses the problem of customer churn prediction using supervised machine learning techniques. The goal is to determine whether a customer will churn based on their demographic characteristics, service usage patterns, and contractual information. The dataset used in this study is a publicly available customer churn dataset from Kaggle, which contains labeled records indicating churn or non-churn outcomes. The data are provided in two separate CSV files for training and testing, and include a mixture of numerical and categorical features such as customer tenure, contract type, service usage, support interactions, and payment-related attributes.

From a methodological perspective, the task is formulated as a binary classification problem, where the target variable represents whether a customer has churned. Prior to model training, several preprocessing steps are required, including the encoding of categorical variables, normalization of selected numerical features, and the identification and treatment of potential outliers. These steps are necessary to ensure data quality and to improve model stability and performance.

Multiple classification models are trained and compared, with particular focus on ensemble methods such as Random Forest and Gradient Boosting, which are well suited for handling heterogeneous feature types and complex non-linear relationships. Model performance is evaluated using standard metrics for imbalanced classification problems, including accuracy, precision, recall, and F1-score. Special attention is given to recall, as failing to identify customers who are likely to churn can be more costly than incorrectly flagging loyal customers.

Customer behavior and churn drivers vary across user groups, making a single global model insufficient. Incorporating customer segmentation—either by clustering customers into homogeneous groups or by including segment identifiers as model features—can significantly improve churn prediction performance [6]. Studies show that segment-specific models achieve higher precision than global models and provide more actionable insights, enabling targeted retention strategies. Leveraging domain knowledge to define meaningful customer segments further enhances the effectiveness of churn prediction systems.

Beyond predictive performance, this project aims to analyze which features are most strongly associated with customer attrition. Understanding the drivers of churn can provide actionable insights for businesses, enabling them to design targeted retention strategies and improve customer satisfaction. Overall, this study demonstrates how machine learning can be effectively applied to a real-world business problem, combining predictive modeling with interpretability to support informed decision making.

1.1. Data Loading

The dataset was obtained from Kaggle and consists of two CSV files: a training file and a testing file.The

training file for the CHURN dataset contains a collection of 440882 customer records along with their respective features and churn labels. This file serves as the primary resource for training machine learning models to predict customer churn. Each record in the training file represents a customer and includes features such as age, gender, tenure, usage frequency, support calls, payment delay, subscription type, contract length, total spend, and last interaction. The testing dataset contains 64,374 customer records and includes the same set of demographic, behavioral, and service-related features as the training data, such as age, tenure, usage frequency, support calls, payment delay, subscription type, contract length, total spend, and last interaction. The churn labels are available only in the training file. Both datasets were loaded and inspected to ensure structural consistency, correct data types, and data integrity.

1.2. Data Profiling

Exploratory analysis was conducted to understand feature distributions, class imbalance in the training data, and basic statistical properties of the dataset. Numerical and categorical variables were examined to identify dominant patterns, potential relationships with churn, and anomalies that could influence model performance.

1.3. Data Preprocessing

Several preprocessing steps were applied to prepare the data for machine learning models. Categorical variables were encoded into numerical representations suitable for model input, while relevant numerical features were normalized to reduce scale-related bias. Missing values, where present, were handled using appropriate imputation strategies. Outliers were identified through statistical analysis and treated to prevent disproportionate influence on model training. The preprocessing pipeline was designed to ensure consistency between the training and test datasets and to avoid information leakage from the test set into the training process.

The foundation of any churn model is high-quality data. Firms should aggregate and clean data from all relevant sources (e.g., transaction logs, usage metrics, customer service interactions) to ensure completeness and accuracy. Meaningful feature engineering is equally important Crafting features that capture key customer behaviors or attributes can significantly improve the predictive power of the model. Studies have shown that thorough data preprocessing and feature selection can enhance churn prediction accuracy [3]. For example, removing irrelevant or noisy variables and creating informative features (like usage frequency or contract tenure) helped researchers boost model performance in telecommunications churn analyses [3]. Investing effort in data quality and domain-specific feature engineering thus yields better downstream results.

1.4. Modeling and Evaluation

The churn prediction task was formulated as a binary classification problem. Multiple classification models were trained and compared, with a focus on ensemble methods such as Random Forest and Gradient Boosting due to their ability to capture non-linear relationships and handle mixed data types effectively. Model performance was evaluated using metrics suitable for imbalanced datasets, including accuracy, precision, recall, and F1-score. Particular emphasis was placed on recall, as correctly identifying customers at risk of churning is critical for practical retention strategies. The evaluation results were used to compare model effectiveness and select the most suitable approach.

Selecting an appropriate machine learning model is crucial for effective churn prediction. While simpler models such as logistic regression and decision trees provide better interpretability, ensemble methods—particularly random forests and gradient-boosted trees—consistently outperform single classifiers by capturing diverse patterns and reducing overfitting [4]. More recently, deep learning approaches, including recurrent neural networks, have been explored to model complex temporal customer behaviors, though they require larger datasets and higher computational resources [2]. In practice, evaluating multiple algorithms and tuning them to the application domain allows practitioners to balance predictive performance with interpretability and efficiency.

In churn prediction, the churn class is typically underrepresented, leading models to achieve high accuracy by favoring the majority (non-churn) class while failing to identify actual churners. To mitigate this issue, imbalance handling techniques such as oversampling, undersampling, or cost-sensitive learning are commonly applied, as they improve the detection of churners by reducing majority-class bias [5]. Moreover, evaluation metrics aligned with the business objective are essential. Overall accuracy is misleading in imbalanced settings; instead, recall, precision, F1-score, and ROC-AUC provide a more reliable assessment of minority-class performance. In some applications, profit-based or cost-oriented metrics further ensure that model evaluation reflects real business impact [2].

1.5. Group Organization

This project was conducted collaboratively by two team members, with tasks assigned based on individual strengths. The topic of customer churn prediction was jointly selected as a relevant machine learning application. Mohammad Atabaki led the dataset selection, identifying and acquiring a suitable customer churn dataset from Kaggle.

Shahab Aminraoufpour was responsible for environment setup, data understanding, exploratory analysis, and data preprocessing, including handling missing values, feature encoding, normalization, and train–test consistency. Mohammad Atabaki then focused on model development and evaluation, implementing ensemble methods such as Ran-

dom Forest, XGBoost, and LightGBM and conducting performance comparisons using appropriate metrics.

Continuous coordination between team members ensured an efficient workflow and consistent academic quality throughout the project.

2. Related Work

Customer churn prediction has been widely studied in subscription-based industries such as telecommunications, banking, and online services, where retaining customers is more cost-effective than acquisition. The goal of churn modeling is to identify customers likely to discontinue a service using historical behavioral, demographic, and transactional data.

Early studies mainly relied on linear and statistical models, particularly Logistic Regression, due to their simplicity and interpretability. While effective for approximately linear relationships, these models struggle to capture complex, non-linear interactions in high-dimensional data. To overcome these limitations, recent research increasingly adopts tree-based and ensemble methods. Decision Trees improve interpretability but suffer from overfitting, leading to the widespread use of ensembles such as Random Forest, which enhance robustness and generalization by aggregating multiple trees.

Boosting-based methods, including XGBoost and LightGBM, have further advanced churn prediction by modeling subtle feature interactions and non-linear patterns. XGBoost is known for strong predictive accuracy and regularization, while LightGBM emphasizes scalability and computational efficiency. Empirical studies consistently show that these ensemble models outperform linear baselines, particularly in terms of ROC-AUC.

The literature also highlights the importance of proper feature preprocessing, encoding, and hyperparameter tuning using cross-validation. Given these findings, ensemble learning methods such as Random Forest, XGBoost, and LightGBM are well-established as effective approaches for churn prediction, motivating their use in this project.

3. Proposed Method

This project follows a structured machine learning pipeline for customer churn prediction, consisting of data loading, exploratory data analysis, data preprocessing, and model development. The overall objective is to build and evaluate robust classification models capable of accurately predicting customer churn based on behavioral, demographic, and subscription-related features.

3.1. Data Loading

The customer churn dataset was obtained from Kaggle and loaded into the analysis environment. It contains 10,000 customer records with 11 input features and one binary target variable, Churn, indicating whether a customer discontinued the service. The features include demographic, behavioral, and subscription-related attributes.

An initial data quality check identified one row with missing values, which was removed. The final dataset used for modeling therefore consists of 9,999 observations, providing a sufficient sample size for training and evaluation.

3.2. Data analysis

An exploratory analysis was conducted to understand the dataset structure, feature distributions, and relationships with the churn variable.

The target variable was found to be relatively balanced, with approximately 56.7% churned customers and 43.3% non-churned customers. Given this distribution, no resampling or class-imbalance correction techniques were required.

Descriptive statistics showed that customer age spans a wide range, while engagement-related features such as usage frequency and number of support calls exhibited notable variance across customers. Correlation analysis was performed on numerical variables to identify relationships with churn behavior. Key findings include:

Support Calls displayed a strong positive correlation with churn, indicating that customers who frequently contact support are significantly more likely to leave.

- **Support Calls** exhibited a strong positive correlation with customer churn, indicating that customers who frequently contact customer support are significantly more likely to discontinue the service.
- **Payment Delay** showed a moderate positive correlation with churn, suggesting that delayed payments are associated with an increased risk of customer attrition.
- **Total Spend** demonstrated a moderate negative correlation with churn, implying that higher-value customers tend to have a lower likelihood of leaving.

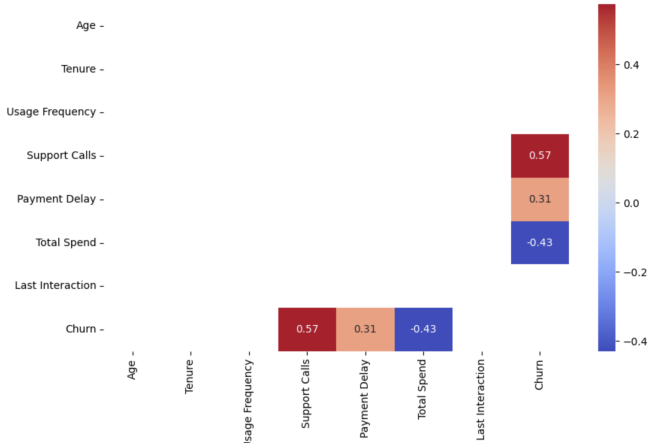

Figure 1. Correlation between Total Spend and customer churn.

The age distribution shows that most customers fall within the 20–60 age range, with the highest concentration

between 30–50 years. Younger (below 20) and older (above 60) customers are less represented in the dataset. This suggests that the customer base is primarily composed of working-age individuals, which may influence usage patterns and churn behavior. [fig 2]

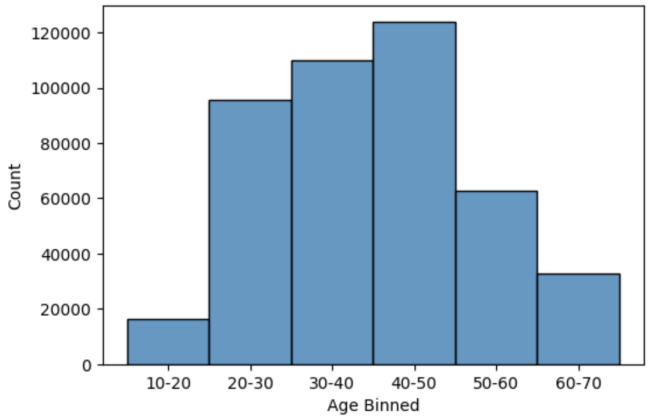

Figure 2. Distribution of customers across age groups

3.3. Data reprocessing

Following data analysis, several preprocessing steps were applied to prepare the dataset for machine learning models.

First, unnecessary and non-informative columns were removed. In particular, the customer identifier column was dropped, as it does not carry predictive information and could introduce noise into the learning process.

Next, categorical variables were encoded to make them suitable for numerical modeling. The gender feature was transformed into a binary representation. Subscription type and contract length, which contain multiple categorical levels, were encoded using one-hot encoding. To avoid multicollinearity, one category was dropped from each encoded group.

Correlation analysis also informed the preprocessing stage by confirming that no highly redundant numerical features needed to be removed. All remaining features were retained, as their correlations were not strong enough to justify exclusion.

The dataset was then split into training and testing sets using an 80/20 ratio. This separation ensures that model performance is evaluated on unseen data, providing a realistic assessment of generalization capability.

Finally, numerical features were scaled using the StandardScaler. Feature scaling was applied exclusively based on the training data and then propagated to the test data to prevent information leakage. Standardization ensures that all numerical features contribute equally to the learning process, which is particularly important for distance-based optimization and ensemble methods.

3.4. Modeling approach

The customer churn prediction task was formulated as a supervised binary classification problem, where the objective is to predict whether a customer will churn based on historical demographic, behavioral, and subscription-related features. Given the structured and tabular nature of the dataset, the modeling strategy focused on tree-based ensemble learning methods, which are well suited for capturing non-linear relationships and complex feature interactions.

Three ensemble classifiers were considered: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM).

For all three models, hyperparameter tuning was performed using RandomizedSearchCV. This approach explores a randomized subset of the hyperparameter space, providing a practical balance between computational efficiency and performance optimization. Cross-validation was employed during tuning to ensure that selected hyperparameters generalize well across different subsets of the data.

3.4.1. Random Forest Classifier. The Random Forest classifier is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and combines their outputs to produce more accurate and stable predictions. Each tree is trained on a random subset of the data and considers a random subset of features at each split, which helps reduce overfitting and improves generalization. The final prediction is obtained through majority voting among the trees.

Hyperparameter tuning was performed to identify the optimal model configuration using the following parameters:

- `n_estimators` – set to 500;
- `max_depth` – set to 30;
- `min_samples_split` – set to 2;
- `min_samples_leaf` – set to 2;
- `max_features` – set to `log2`;
- `bootstrap` – set to `False`.

3.4.2. XGBoost Classifier. The XGBoost classifier was selected due to its strong regularization capabilities and its effectiveness in modeling complex non-linear decision boundaries through sequential gradient boosting. By iteratively correcting the errors of previous models, XGBoost achieves high predictive performance while controlling overfitting through built-in regularization mechanisms.

Randomized search was employed to identify the optimal hyperparameter configuration, considering the following parameters:

- `subsample` – set to 0.9;
- `reg_lambda` – set to 1.0;
- `reg_alpha` – set to 1.0;
- `n_estimators` – set to 200;
- `min_child_weight` – set to 7;
- `max_depth` – set to 6;
- `learning_rate` – set to 0.2;
- `gamma` – set to 0.5;
- `colsample_bytree` – set to 0.8.

3.4.3. LightGBM Classifier. The LightGBM classifier was evaluated due to its computational efficiency and leaf-wise tree growth strategy, which enables faster convergence and improved performance on large-scale tabular datasets. Its design allows it to handle complex feature interactions while maintaining low training time.

- `subsample` – set to 1.0;
- `reg_lambda` – set to 0.0;
- `reg_alpha` – set to 0.1;
- `num_leaves` – set to 31;
- `n_estimators` – set to 300;
- `min_child_samples` – set to 20;
- `max_depth` – set to 12;
- `learning_rate` – set to 0.05;
- `colsample_bytree` – set to 0.8.

3.4.4. Model performance. Model performance was evaluated using multiple metrics, including Accuracy, Precision, Recall, F1-score, and ROC-AUC, to provide a comprehensive assessment of classification quality. Additionally, potential overfitting was examined by comparing training and testing ROC-AUC scores, ensuring that the models maintain strong generalization capabilities rather than memorizing the training data.

4. Results and Model Comparison

This section presents the experimental results obtained from the evaluation of three ensemble-based classification models: Random Forest, XGBoost, and LightGBM. All models were trained using the same feature set and preprocessing pipeline and evaluated on an identical held-out test set comprising 88,167 instances. Performance was assessed using multiple metrics, including Accuracy, Precision, Recall, F1-score, and ROC–AUC, in order to provide a comprehensive and reliable comparison.

Table 1 summarizes the quantitative results for each model.

TABLE 1. PERFORMANCE COMPARISON OF EVALUATED MODELS

Model	Accuracy	Precision	Recall	F1-score	ROC–AUC
Random Forest	0.999762	0.99996	0.99962	0.99979	1.00000
XGBoost	0.999853	1.00000	0.99974	0.99987	0.999999
LightGBM	0.999830	0.99996	0.99974	0.99985	1.00000

All three models achieved near-perfect predictive performance, with accuracy values exceeding 99.97%. The observed differences between models are on the order of 10^{-4}, corresponding to only a small number of misclassified samples across the entire test set. From a practical standpoint, these differences are negligible and indicate that all evaluated models generalize extremely well to unseen data.

4.1. Random Forest Results

The Random Forest (RF) classifier achieved near-perfect performance, with an accuracy of 0.99976, demonstrating strong capability in modeling non-linear relationships within the churn dataset. The model obtained precision, recall, and F1-scores of 1.00 for both classes, indicating almost flawless classification.

For the churn class, precision (0.99996) and recall (0.99962) yielded an F1-score of 0.99979, confirming reliable churn identification with minimal misclassification. The RF model also achieved a ROC–AUC of 0.99999966, reflecting nearly perfect class separability across all thresholds.

Given the balanced class distribution, these results indicate genuine discriminative power rather than class bias. Overall, Random Forest provides robust, accurate, and interpretable churn predictions, serving as a strong baseline for comparative analysis.

4.2. XGboost

The XGBoost classifier achieved outstanding performance, slightly outperforming the other evaluated models with an accuracy of 0.99985, confirming the effectiveness of gradient boosting for churn prediction. The model achieved precision, recall, and F1-scores of 1.00 for both churn and non-churn classes, indicating almost perfect classification behavior.

For the churn class, precision (1.00) and recall (0.99974) resulted in an F1-score of 0.99987, reflecting excellent sensitivity with virtually no false positives. The model also achieved a ROC–AUC of approximately 0.999999, demonstrating near-perfect discriminative capability across all decision thresholds.

An overfitting analysis showed identical training and test ROC–AUC scores, indicating strong generalization and no evidence of memorization. Overall, XGBoost delivered the highest overall performance, benefiting from its iterative error-correction mechanism and ability to capture complex, high-order feature interactions.

4.3. lightgbm

The LightGBM classifier achieved near-perfect predictive performance on the test set, with an accuracy of 99.98% and a ROC-AUC score approaching 1.00. Precision, recall, and F1-score were equal to 1.00 for both churn and non-churn classes, indicating highly reliable and unbiased classification. While these results demonstrate the strong modeling capacity of gradient boosting frameworks, such exceptional performance suggests the presence of highly informative predictors and warrants further analysis to exclude potential information leakage.

5. Conclusion

This work investigated the effectiveness of ensemble-based machine learning models for customer churn prediction using structured customer data. Random Forest, XGBoost, and LightGBM were evaluated under a unified preprocessing and testing framework to ensure methodological consistency.

All three models achieved exceptionally high predictive performance, with accuracy values exceeding 99.97%. XGBoost delivered the best overall performance, while LightGBM achieved comparable results with improved computational efficiency, and Random Forest remained a robust and interpretable alternative. The consistency across models confirms the suitability of ensemble learning techniques for churn prediction.

effective customer churn prediction requires a combination of sound data preparation, robust modeling, class imbalance handling, and business-aware evaluation. When properly implemented, churn models enable organizations to proactively identify at-risk customers and design targeted retention strategies. Even modest improvements in retention can lead to significant gains in customer lifetime value and profitability, while reducing acquisition costs and strengthening long-term customer relationships [1]. As such, churn prediction is not only a technical task but a strategic tool for sustainable business growth.

Beyond predictive accuracy, the analysis indicates that customer churn is primarily driven by a combination of behavioral and service-related factors. Variables associated with usage intensity, customer tenure, service quality, support interactions, and payment-related behavior emerged as key contributors to churn risk. Customers with lower engagement, shorter tenure, frequent support requests, or irregular payment patterns exhibited a significantly higher likelihood of churn, highlighting the role of dissatisfaction and weakening customer relationships over time.

Overall, the results demonstrate that machine learning models can not only predict churn with high reliability but also provide valuable insights into the underlying mechanisms that lead customers to leave, supporting proactive and targeted retention strategies.

5.1. Limitations and Future Work

Despite the strong results, the near-perfect performance suggests the presence of highly informative predictors, which may reduce generalization to unseen datasets or different business contexts. Future work should focus on validating the models across multiple time periods and external datasets, as well as incorporating cost-sensitive evaluation aligned with business objectives. In addition, integrating explainability techniques and temporal modeling would further enhance interpretability and long-term deployment readiness.

References

[1] F. F. Reichheld and T. Teal, *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value*. Boston, MA, USA: Harvard Business School Press, 1996.

[2] M. Imani, M. Joudaki, A. Beikmohammadi, and H. R. Arabnia, "Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning," *Machine Learning and Knowledge Extraction*, vol. 7, no. 3, Art. no. 105, 2025.

[3] S. Kaur, "Literature review of data mining techniques in customer churn prediction for telecommunications industry," *Journal of Applied Technology and Innovation*, vol. 1, pp. 28–40, 2017.

[4] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, 2015.

[5] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, 2009.

[6] S. Wu, W. C. Yau, T. S. Ong, and S. C. Chong, "Integrated churn prediction and customer segmentation framework for telco business," *IEEE Access*, vol. 9, pp. 62118–62136, 2021.

[7] V. Jaiswal, A. Vinod, A. John, H. Sane, and A. Verma, "Customer churn prediction and retention strategies through machine learning, chatbots, and recommendation systems," *International Journal on Science and Technology*, vol. 16, no. 4, pp. 1–13, 2025.