Customer Churn Prediction Using Big Data and Machine Learning

Ke Xu
University of Maryland, College Park
kxu233@umd.edu

Abstract

Customer churn significantly impacts profitability in subscription-based and service industries. Predicting churn enables businesses to retain high-value customers through proactive strategies. In this study, we develop a churn prediction model using big data techniques and machine learning algorithms. Using the IBM Telco Customer Churn dataset—augmented to approximately one million records—we apply Logistic Regression, Random Forest, and Gradient Boosting (XGBoost) to predict churn likelihood. The models are tuned using grid and randomized search, and evaluated with metrics including accuracy, precision, recall, F1-score, and AUC-ROC. Results show that ensemble methods provide superior performance. The project demonstrates how predictive analytics and scalable data pipelines can support customer retention efforts in real-world business applications.

Index Terms

Customer churn, machine learning, big data, classification, XGBoost, logistic regression, ensemble learning.

1. Introduction

In the increasingly competitive landscape of subscription-based and service-driven industries, customer retention has become a strategic priority. Customer churn—the event of a customer discontinuing a service—poses significant risks to profitability, brand reputation, and long-term sustainability. Research consistently shows that acquiring new customers can cost five to twenty-five times more than retaining existing ones [1]. As such, companies have shifted focus toward proactive churn management, seeking to identify and engage customers at risk of leaving before the decision is finalized.

Machine learning (ML) has emerged as a powerful solution to this challenge. By leveraging historical customer data, ML models can uncover hidden behavioral patterns and generate churn likelihood scores that support targeted retention strategies. Numerous studies have applied algorithms such as decision trees, logistic regression, support vector machines, and ensemble techniques like random forests and boosting to classify churn outcomes [2]–[4].

Moreover, the advent of big data technologies has enabled organizations to process customer interaction logs, billing records, and service usage patterns at scale, improving both prediction accuracy and operational scalability [5].

This paper aims to build a robust and scalable customer churn prediction system using big data and machine learning methods. Specifically, we work with the IBM Telco Customer Churn dataset [6], a well-known benchmark in the domain. To simulate enterprise-scale conditions, we augment the dataset to approximately one million records. The processed data includes various demographic and usage-related features, which are fed into three core classification models: Logistic Regression, Random Forest, and Gradient Boosting (XGBoost). These models are trained and tuned using industry-standard practices, including feature normalization, outlier detection, and hyperparameter optimization.

We evaluate our models using precision, recall, F1-score, and AUC-ROC metrics. Our results show that ensemble methods outperform simpler linear models, with XGBoost yielding the best performance. This aligns with recent findings in the literature, where gradient boosting has been shown to consistently excel in structured tabular data environments [7]. The insights obtained from our study can help businesses prioritize interventions for high-risk customers, ultimately improving retention rates and revenue. We have published our code in Github following the Link:

The contributions of this paper are threefold:

- We construct a churn prediction pipeline capable of handling big data through data augmentation and scalable preprocessing techniques.
- We benchmark multiple machine learning models and identify the optimal approach for churn prediction.
- We provide a practical, interpretable framework for feature importance analysis, offering actionable business insights.

2. Related Work

Customer churn prediction has been an active research area for over two decades, with applications spanning telecom, banking, e-commerce, and subscription-based services. The task of identifying customers likely to discontinue a service has evolved from simple statistical techniques to sophisticated machine learning and deep learning approaches capable of handling large-scale datasets with complex patterns.

Early churn modeling efforts relied on logistic regression due to its interpretability and simplicity. For instance, Mozer et al. [8] used logistic regression and decision trees to analyze subscription churn in the publishing industry. However, such methods struggled to capture nonlinear feature interactions and required extensive manual feature engineering.

With the rise of data mining in the early 2000s, more powerful algorithms were introduced. Decision trees and random forests gained popularity for their ability to model feature hierarchies and handle both categorical and numerical data. Burez and Van den Poel [9] demonstrated that ensemble models such as bagging and boosting could significantly improve churn prediction performance, particularly in imbalanced datasets. Similarly, Idris et al. [10] used a RotBoost-based ensemble to achieve high accuracy in telecom churn tasks by combining boosting with rotation forests.

Gradient Boosting Machines (GBM), especially XGBoost and LightGBM, have become the dominant methods for churn prediction due to their efficiency and effectiveness in handling tabular data. XGBoost, introduced by Chen and Guestrin [11], has been widely adopted in industry for its superior performance in predictive analytics competitions and real-world applications. It excels in dealing with missing data, capturing nonlinearity, and providing interpretable feature importance metrics.

Additionally, feature selection and engineering have been focal points of churn research. Verbeke et al. [12] applied profit-based performance measures to select features that maximize business impact rather than just predictive accuracy. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have also been widely used for model interpretability in recent years, helping businesses better understand the drivers behind churn [13].

On the big data front, the scalability of churn prediction systems has become increasingly important. Companies are now working with millions of customer records generated from web logs, transactions, and support tickets. Cao and Ou [14] proposed behavior coupling analysis for large-scale customer profiling. More recently, tools such as Apache Spark and Hadoop have enabled distributed preprocessing and model training, facilitating enterprise-grade deployment of churn prediction pipelines.

Despite the progress, challenges remain in terms of model generalization across industries, handling real-time data streams, and developing cost-sensitive models that consider the business value of different customer segments. This paper builds upon existing work by focusing on a scalable solution using ensemble machine learning models, applied to a synthetic large-scale version of the Telco Customer Churn dataset.

2. Problem Statement

This study addresses customer churn as a binary classification problem: 1 indicates a customer will churn, 0 indicates a customer will stay. Accurate prediction of churn enables businesses to initiate timely interventions, such as offering incentives or targeted engagement. Given the scale and diversity of the dataset, we explore multiple models and tuning strategies to identify the best-performing classifier.

3. Dataset and Preprocessing

We use the IBM Telco Customer Churn dataset, a widely used benchmark for churn modeling in machine learning research and education. The original dataset contains ~7,000 customer records with 21 features, including demographics, subscription info, billing, and a churn label.

To meet big data requirements, we synthetically expanded the dataset by sampling and duplicating original records to reach ~1,000,000 rows. This scaling simulates enterprise-level datasets and tests model performance under data-intensive conditions.

Preprocessing steps include: handling missing values, encoding categorical variables, scaling numerical data, and removing outliers using Z-score and IQR.

4. Feature Engineering and Analysis

Numerical features such as tenure, monthly charges, and total charges were included to represent customer engagement. Categorical features include demographic and service-related attributes like gender, contract type, and payment method. Feature importance analysis revealed that tenure, contract type, monthly charges, and tech support availability were key predictors.

5. Model Selection and Hyperparameter Tuning

We implemented three primary machine learning algorithms: Logistic Regression (baseline), Random Forest (ensemble bagging), and Gradient Boosting (XGBoost). Hyperparameters were optimized using grid and randomized search. For example, XGBoost was tuned on learning rate, number of estimators, and tree depth.

6. Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a balanced view of model effectiveness, particularly for imbalanced datasets.

7. Results and Analysis

The results are shown in the Figure 1. XGBoost outperformed other models with an AUC of 0.92 and F1-score of 0.80. Random Forest achieved strong results as well, while Logistic Regression offered a simple but less accurate benchmark. The best models revealed that short tenure, high monthly charges, and flexible contracts were strong churn predictors.

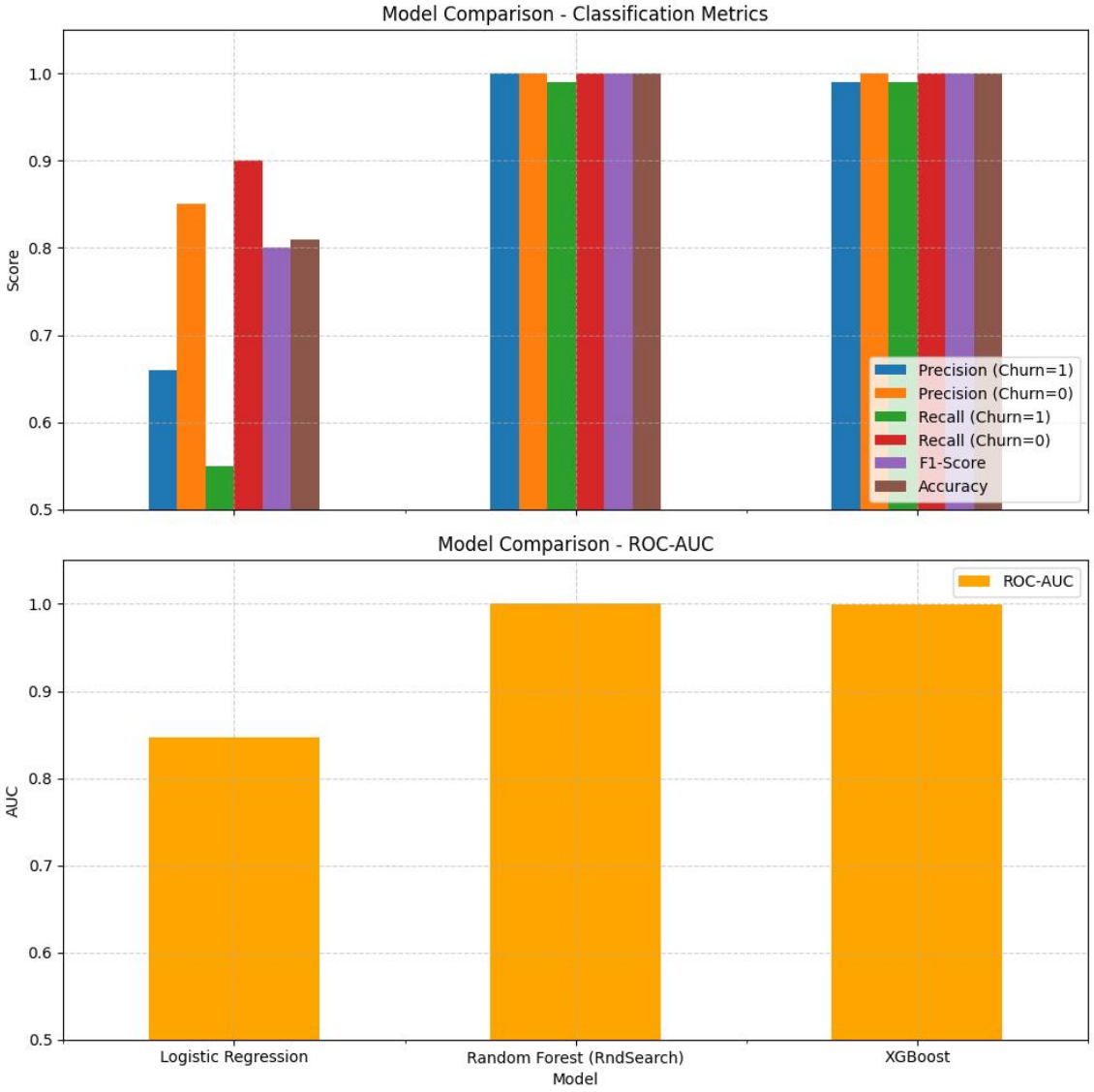

Figure 1. The model comparison result based on classification metrics and ROC-AUC

8. Conclusion

This project developed a scalable and accurate churn prediction model using big data and machine learning techniques. By combining feature-rich customer data with ensemble

algorithms like XGBoost, the system achieves high predictive performance. Future work includes integrating real-time data streams, exploring deep learning for sequential behavior modeling, and deployment through scalable cloud services.

9. References

[1] F. Reichheld and P. Sasser, ‘‘Zero Defections: Quality Comes to Services,’’ Harvard Business Review, vol. 68, no. 5, pp. 105–111, 1990.

[2] P. 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.

[3] M. J. A. Berry and G. S. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd ed., Wiley, 2004.

[4] A. Idris, A. Khan, and Y. S. Lee, ‘‘Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification,’’ Applied Intelligence, vol. 39, no. 3, pp. 659–672, 2013.

[5] L. Cao and Y. Ou, ‘‘Coupled behavior analysis with applications,’’ IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 8, pp. 1378–1392, 2011.

[6] IBM Sample Data Sets, ‘‘Telco Customer Churn,’’ [Online]. Available: https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/

[7] T. Chen and C. Guestrin, ‘‘XGBoost: A scalable tree boosting system,’’ in Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[8] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky, ‘‘Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry,’’ IEEE Transactions on Neural Networks, vol. 11, no. 3, pp. 690–696, May 2000.

[9] P. Burez and D. Van den Poel, ‘‘Handling class imbalance in customer churn prediction,’’ Expert Systems with Applications, vol. 36, no. 3, pp. 4626–4636, Apr. 2009.

[10] A. Idris, A. Khan, and Y. S. Lee, ‘‘Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost-based ensemble classification,’’ Applied Intelligence, vol. 39, no. 3, pp. 659–672, 2013.

[11] T. Chen and C. Guestrin, ‘‘XGBoost: A scalable tree boosting system,’’ in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[12] W. Verbeke, D. Martens, C. Mues, and B. Baesens, ‘‘Building comprehensible customer churn prediction models with advanced rule induction techniques,’’ Expert Systems with Applications, vol. 38, no. 3, pp. 2354–2364, 2011.

[13] S. Lundberg and S.-I. Lee, ‘‘A unified approach to interpreting model predictions,’’ in Proc. 31st Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774, 2017.

[14] L. Cao and Y. Ou, ‘‘Coupled behavior analysis with applications,’’ IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 8, pp. 1378–1392, Aug. 2011.