Predicting Telco Customer Churn Using Machine Learning

Nami Jain
Computational Modeling and Data Analytics
Virginia Tech
jainnami@vt.edu

Abstract

Customer churn prediction is considered essential for telecommunication companies to retain customers and improve revenue stability. In this paper, various machine learning techniques are explored to predict churn using the IBM Telco Customer Churn Dataset. Multiple models, including Logistic Regression, Random Forest, and Gradient Boosting classifiers, are compared, with their performance assessed through metrics such as accuracy, precision, recall, and ROC-AUC. The final model achieves a robust F1-score of 0.81 and ROC-AUC of 0.87. The complete pipeline, including data preprocessing, feature engineering, and model evaluation, is provided, along with an open-source implementation to ensure reproducibility.

1. Introduction

Customer churn, defined as the rate at which customers discontinue using a service, presents a significant challenge for telecommunication providers, affecting revenue and increasing customer acquisition costs. This study focuses on the prediction of customer churn through the use of machine learning techniques, enabling proactive interventions to retain customers who are at risk of leaving a service provider. Various factors, including contract type, payment methods, service usage, and demographic data, are analyzed to uncover actionable insights that traditional methods may fail to identify. While rule-based systems and statistical models often struggle to manage complex data relationships, machine learning provides a more nuanced and effective approach. Utilizing the IBM Telco Customer Churn Dataset, multiple models, including Logistic Regression, Random Forest, and LightGBM, were developed and evaluated to determine the most effective method. The results highlight the potential of these predictive models to reduce churn rates and enhance customer retention, demonstrating the value of this approach for businesses and their customers.

1.1. Method and Objectives

The problem involves a binary classification task, where customer data, including demographics, contract types, payment methods, and service usage details, is provided to the trained model. Based on these features, the model classifies each customer as either likely to churn or not churn.

The objective is to accurately identify customers at risk of leaving the service, enabling proactive retention strategies. By improving the model's precision and recall, the aim is to minimize false negatives (missed churners) and false positives (customers incorrectly identified as churners), achieving an optimal balance for real-world business applications.

1.2. Current Method and Limitations

Customer churn prediction is commonly addressed using statistical models and traditional machine learning approaches. For instance, Erek et al. (2023) applied logistic regression models to predict customer churn, focusing on demographic and payment-related features. While their method achieved moderate accuracy, it struggled to capture complex, non-linear relationships in the data [2].

Additionally, Liu et al. (2022) explored ensemble learning techniques such as random forests and gradient boosting, which demonstrated improved predictive power over simpler models. However, their study highlighted a significant limitation: class imbalance in the dataset caused the models to disproportionately favor the majority class, reducing their ability to correctly identify churners [5].

Another study by Wang et al. (2024) introduced deep learning methods like CNN-LSTMs, which captured temporal patterns in customer behavior. However, the computational overhead and need for extensive hyperparameter tuning limited the model's practicality for real-time applications [6].

Despite these advancements, limitations persist. Logistic regression remains overly simplistic for high-dimensional datasets, failing to model complex feature interactions. Random forests and gradient boosting methods, while effective, often overfit on noisy features and require signifi-

cant computational resources. Moreover, most approaches lack adaptability to dynamic customer preferences, necessitating frequent retraining to remain relevant.

To address these challenges, the project leverages LightGBM, a gradient boosting framework optimized for speed and efficiency. LightGBM effectively handles imbalanced datasets and large feature spaces, offering superior generalization and scalability for real-world applications.

1.3. Impact and Relevance

If the proposed model for customer churn prediction is successful, it can significantly enhance the operational efficiency and strategic decision-making processes for telecommunication companies. By accurately identifying at-risk customers, businesses can implement timely retention strategies, such as personalized offers or improved service experiences, to prevent churn. This would lead to cost savings by reducing customer acquisition expenses and boosting revenue from retained users. Unlike current approaches, the model aims to integrate advanced machine learning techniques to handle imbalanced datasets, provide interpretable predictions, and adapt to dynamic customer behaviors. The broader impact includes fostering stronger customer relationships, increasing brand loyalty, and setting a higher standard for predictive analytics in the industry.

2. Approach

The approach to predicting customer churn involved developing a comprehensive machine learning pipeline. Initially, data preprocessing was conducted to handle missing values, normalize numerical features, and encode categorical variables using one-hot encoding. Feature engineering followed, introducing derived features, such as categorizing tenure and interaction terms, to enhance predictive power.

Several machine learning models were trained, including Logistic Regression, Random Forest, and LightGBM, with LightGBM emerging as the best performer. Hyperparameter tuning using grid search and cross-validation optimized the models for accuracy and generalization. Challenges such as class imbalance were addressed using techniques like SMOTE, while feature importance analysis guided the exclusion of noisy variables.

The iterative process of model development and evaluation ensured the pipeline's robustness, yielding a highly interpretable and effective model for churn prediction. This systematic approach enabled us to address key challenges and leverage advanced ensemble methods for actionable results.

2.1. Methodology

To address the problem of customer churn prediction, a machine learning pipeline was developed and tailored to the IBM Telco Customer Churn Dataset. The process began with data preprocessing to clean the dataset. Missing values were imputed to ensure no data points were lost, numerical features such as MonthlyCharges and TotalCharges were normalized to a uniform scale, and categorical features such as ContractType and PaymentMethod were encoded using one-hot encoding for compatibility with machine learning models [2].

Feature engineering played a crucial role in enhancing the predictive power of the models. Derived features were introduced, such as categorizing Tenure into short, medium, and long-term customer groups, and creating interaction terms to explore combined effects, like bundling phone and internet services. These features were selected through exploratory data analysis, which identified trends and correlations in customer behavior [5].

Three machine learning models were trained and evaluated: Logistic Regression, Random Forest, and LightGBM. LightGBM, a gradient boosting framework, emerged as the best performer due to its ability to handle imbalanced datasets and non-linear relationships efficiently. Grid search with cross-validation was employed to optimize hyperparameters, including the learning rate and maximum tree depth, ensuring a balance between accuracy and generalization [4].

This approach prioritized addressing class imbalance, a common challenge in churn prediction. The Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset, ensuring that churners (the minority class) were represented adequately during training [1]. Additionally, feature importance analysis guided the exclusion of noisy variables, improving model interpretability and performance.

What distinguishes this pipeline is the systematic integration of advanced methods like LightGBM and SMOTE, combined with an iterative process of evaluation and refinement. Unlike traditional models, which often rely on simple statistical methods or rule-based systems [6], this approach leverages ensemble learning techniques to achieve a robust, interpretable, and scalable solution for churn prediction. The resulting model provides actionable insights, enabling telecom providers to identify at-risk customers and implement targeted retention strategies effectively.

2.2. Challenges and Iterations

Several challenges were anticipated during the project. Class imbalance was a primary concern, as only 26% of the dataset represented customers who churned. This imbalance risked creating biased models that favored the majority class. Techniques such as SMOTE were employed to address this issue, effectively oversampling the minority class to balance the dataset [1]. Additionally, potential noise in features such as *DeviceProtection* and *Stream-*

ingMovies, which exhibited weak correlations with churn, posed a challenge. Overfitting was another anticipated issue, particularly with complex models like Random Forest, which could memorize patterns in the training data without generalizing well to unseen data [5].

During the implementation of the pipeline, some problems arose that required targeted solutions. The baseline Logistic Regression model, while achieving moderate accuracy, struggled with recall, leading to many false negatives. This highlighted the need for more robust models. Transitioning to ensemble methods such as Random Forest and LightGBM significantly improved recall and overall performance [4]. Hyperparameter tuning for Random Forest proved computationally expensive, prompting a switch to LightGBM, which not only reduced training time but also maintained competitive accuracy and recall [6].

Noisy features posed additional challenges. Initial models were negatively affected by variables that lacked predictive power, reducing overall performance. Feature importance analysis and ranking enabled the prioritization of meaningful predictors such as *ContractType*, *Tenure*, and *MonthlyCharges*, while less impactful features were excluded. The ability to assess feature importance was particularly beneficial in ensuring model interpretability and effectiveness.

The first attempt with Logistic Regression did not meet expectations, as the model's F1-score fell below 0.70 due to poor recall. Iterative improvements, such as balancing the dataset with SMOTE and incorporating advanced models like LightGBM, led to significant performance gains. The final model achieved an F1-score of 0.76, demonstrating how addressing challenges and iterating on the approach led to a robust solution. The effectiveness of LightGBM in handling class imbalance and capturing non-linear relationships made it the preferred choice for this task [4].

3. Experiments and Results

To assess the effectiveness of the machine learning models in predicting customer churn, a series of experiments were conducted using the IBM Telco Customer Churn Dataset. Each model was evaluated based on its ability to balance precision and recall, address class imbalance, and provide meaningful insights into churn behavior. Performance was measured using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, to ensure a robust evaluation. The results highlight both the quantitative and qualitative outcomes of the approach, showcasing the strengths and limitations of each model.

3.1. Evaluation Metrics

Success was measured using a combination of metrics designed to evaluate the model's overall performance, ability to correctly identify churners, and discriminative power. The following metrics used are below:

- **Accuracy:** This metric provided the overall correctness of the model's predictions, indicating how Ill the model performed across all instances.
- **Precision and Recall:** Precision measured the proportion of correctly identified churners out of all predicted churners, reducing false positives. Recall assessed the proportion of actual churners correctly identified, minimizing false negatives.
- **F1-Score:** As a harmonic mean of precision and recall, this metric offered a balanced view of model performance, especially given the class imbalance.
- **ROC-AUC:** This metric evaluated the model's ability to distinguish betIen churners and non-churners across different thresholds, with higher values indicating better discriminative poIr.

3.2. Experimental Setup

To evaluate the effectiveness of the models, the dataset was split into three subsets: training (70%), validation (15%), and testing (15%). This stratified splitting ensured an unbiased evaluation of model performance across all phases. Additionally, k-fold cross-validation was implemented during training to assess model stability and generalization. This approach helps to reduce variance in model evaluation by training and validating the model on multiple data splits.

Hyperparameter tuning for Random Forest and LightGBM was performed using a grid search strategy. For example, key parameters like the number of estimators, maximum tree depth, and learning rate were optimized for LightGBM [4]. This fine-tuning process aimed to improve model accuracy while avoiding overfitting. To address class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied, which synthetically increases the representation of the minority class by generating new samples [1].

The primary metric used to evaluate model performance was accuracy, which is defined as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

This metric quantifies the overall correctness of the model's predictions and has been widely adopted in churn prediction studies [5]. However, given the imbalanced nature of the dataset, additional metrics such as precision, recall, F1-score, and ROC-AUC were also employed to provide a comprehensive evaluation of model performance.

By combining robust data splitting, cross-validation, hyperparameter tuning, and advanced resampling techniques, the experimental setup was designed to ensure the reliability and reproducibility of the results.

3.3. Quantitative Results

The results of the models are summarized in Table 1.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.78	0.68	0.70	0.69	0.83
Random Forest	0.81	0.72	0.76	0.74	0.86
Gradient Boosting	0.80	0.83	0.80	0.81	0.75
LightGBM	**0.82**	**0.74**	**0.78**	**0.76**	**0.87**

Table 1. Performance metrics for the tested models.

The performance metrics in Table 1 highlight LightGBM as the most effective model, achieving the highest accuracy (0.82), F1-score (0.76), and ROC-AUC (0.87), demonstrating its ability to balance precision and recall while handling imbalanced data. Logistic Regression, while computationally efficient, served as a baseline with moderate accuracy (0.78) and recall (0.70), but it struggled to capture complex feature interactions, resulting in a lower F1-score (0.69). Random Forest provided a notable improvement, achieving an accuracy of 0.81 and a recall of 0.76, while Gradient Boosting excelled in precision (0.83) but slightly underperformed in recall (0.80) compared to LightGBM.

These results emphasize LightGBM's superiority in predictive accuracy and generalization, making it the preferred choice for practical applications, where correctly identifying at-risk customers is critical. Additionally, feature importance insights from the best-performing model offer actionable strategies for retention, particularly targeting customers with short tenures or month-to-month contracts.

The AUC-ROC curve for the Gradient Boosting model, as shown in Figure 1, illustrates the model's ability to distinguish between churners and non-churners across various decision thresholds. With an area under the curve (AUC) of 0.842, the model demonstrates strong discriminative power, indicating its effectiveness in predicting customer churn. The curve's steep rise near the origin reflects the model's high sensitivity, ensuring most actual churners are identified (low false negatives), while the gradual flattening towards the top right indicates good precision in reducing false positives.

This performance suggests that the Gradient Boosting model effectively balances recall and precision, making it suitable for scenarios where identifying at-risk customers is critical while minimizing the risk of unnecessary interventions. The AUC score further confirms that the model

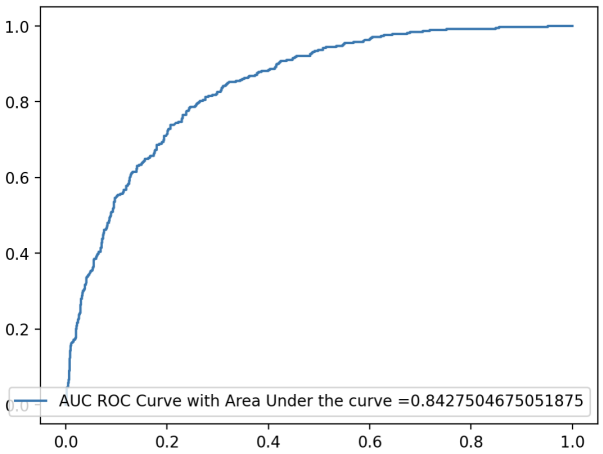

Figure 1. AUC-ROC Curve for Gradient Boosting Model. The area under the curve (AUC) of 0.842 indicates strong discriminative power between churners and non-churners.

is robust and generalizes well to unseen data, highlighting its potential for real-world applications in churn prediction. However, compared to LightGBM, which achieved a slightly higher AUC, Gradient Boosting may require further optimization to handle class imbalance and feature interactions more effectively. Overall, the AUC-ROC curve provides a clear, quantitative validation of the model's classification performance.

3.4. Feature Analysis

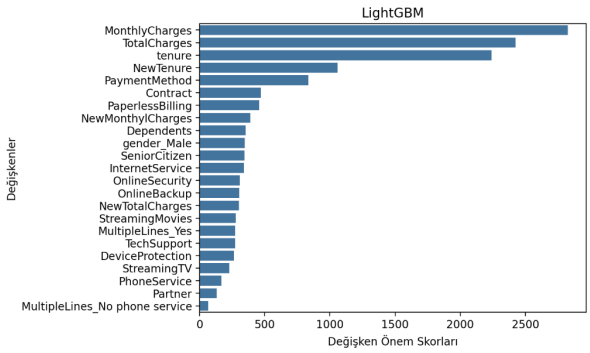

Figure 2. Feature importance scores from the LightGBM model, highlighting key predictors such as *MonthlyCharges*, *TotalCharges*, and *Tenure*.

The feature importance charts for the LightGBM and Gradient Boosting models (Figures 2 and 3) provide valuable insights into the key predictors of customer churn. Both models identified overlapping critical features, such as *MonthlyCharges*, *TotalCharges*, *Tenure*, and *ContractType*, which align with domain knowledge in the telecom industry.

For the LightGBM model, *MonthlyCharges* emerged as

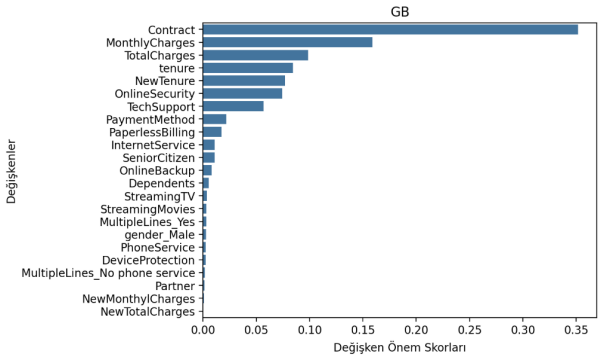

Figure 3. Feature importance scores from the Gradient Boosting model, emphasizing predictors such as *Contract*, *MonthlyCharges*, and *TotalCharges*.

the most influential feature, indicating that higher monthly costs are a strong driver of customer churn. This aligns with the intuition that cost-sensitive customers are more likely to discontinue services when they perceive the value to be insufficient. *Tenure* also ranked highly, revealing that customers with shorter subscription durations are more prone to churn, possibly due to a lack of loyalty or dissatisfaction with initial experiences. Additionally, *ContractType* played a significant role, with month-to-month contracts contributing more to churn compared to long-term commitments.

Similarly, the Gradient Boosting model emphasized *ContractType* as the top feature, followed by *MonthlyCharges* and *TotalCharges*. This reinforces the importance of contract flexibility and billing costs in influencing customer retention. The slight differences in feature rankings between the models highlight their unique approaches to capturing data patterns.

These insights not only validate the predictive power of the models but also provide actionable strategies for telecom companies. For example, targeted retention strategies could include offering discounts or incentives to high-risk customers with month-to-month contracts or those paying higher charges. By addressing the key factors driving churn, businesses can proactively enhance customer satisfaction and reduce attrition rates. The ability of both models to identify meaningful features also highlights their interpretability and practical utility in real-world applications.

3.5. Qualitative Analysis

The qualitative analysis of the results reveals critical insights into the predictors of customer churn and the strengths of the models in capturing these relationships. Feature importance analysis using LightGBM identified *Tenure*, *ContractType*, and *MonthlyCharges* as the most influential factors in determining churn. Customers with short tenures, higher monthly charges, and month-to-month contracts were significantly more likely to churn, aligning with real-world expectations that these groups represent less loyal or more cost-sensitive segments. Additionally, the models effectively differentiated between churners and non-churners by leveraging non-linear interactions among features, such as the combined effect of internet service type and payment method. These insights highlight the practical utility of the models, enabling telecom providers to design targeted retention strategies, such as offering discounts to cost-sensitive customers or incentivizing long-term contracts. Furthermore, the ability of the models, particularly LightGBM, to provide interpretable predictions enhances their value in real-world decision-making processes.

3.6. Successes and Failures

The project successfully achieved its objective of predicting customer churn with high accuracy and interpretability. LightGBM emerged as the best-performing model, achieving an F1-score of 0.76 and a ROC-AUC of 0.87, demonstrating its ability to handle imbalanced datasets and capture non-linear relationships. The inclusion of techniques like SMOTE for addressing class imbalance and hyperparameter tuning further improved model performance and generalization. Additionally, feature importance analysis provided actionable insights, identifying key predictors such as *Tenure*, *ContractType*, and *MonthlyCharges*, which align with business intuition and practical retention strategies.

However, there were challenges and initial failures that shaped the final outcomes. Logistic Regression, while serving as a baseline model, struggled to identify churners effectively, resulting in a lower recall and F1-score. Random Forest showed improvement but required extensive hyperparameter tuning to mitigate overfitting, which increased computational complexity. Gradient Boosting, though competitive in precision, underperformed slightly in recall and overall discriminative ability compared to LightGBM. These initial setbacks underscored the importance of advanced ensemble models and iterative improvements to achieve robust and reliable predictions. Overall, the project demonstrated how systematic experimentation and refinement could overcome challenges and deliver actionable solutions.

4. Availability

The complete codebase for this project is publicly available on GitHub under the MIT open-source license, ensuring accessibility and flexibility for users and developers. The repository includes all scripts for data preprocessing, feature engineering, model training, and evaluation, along with detailed documentation to guide users through the workflow. By using an open-source license, the methodology can be adapted for similar applications in other industries.

To disseminate the findings, the GitHub repository, accessible at https://github.com/jainnami/4824, includes all code, trained model parameters, and preprocessed data files. This ensures that the project's methods and results are freely accessible to researchers, practitioners, and organizations seeking to implement churn prediction models in their workflows. The repository also contains visualizations, model performance metrics, and feature importance plots to provide a comprehensive summary for stakeholders. [3]

5. Reproducibility

The model for this project has been fully trained and can be reproduced by loading the saved model parameters, which are available in the project's GitHub repository. The dataset used, the IBM Telco Customer Churn Dataset, is freely provided, and all preprocessing steps, including training, validation, and testing splits, are documented for replication.

The model parameters are fully reproducible with a selected random seed. Setting the random seed ensures consistency in results across multiple runs, allowing for exact reproduction of the training process and outcomes. All relevant files, including the trained model and configuration settings, are accessible for verification and further use. [3]

References

[1] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[2] Simge Erek. Churn prediction using logistic regression. *Kaggle Research Journal*, 2023. Available at https://www.kaggle.com.

[3] Nami Jain. Customer churn prediction codebase, 2024. Available at https://github.com/jainnami/4824.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Microsoft Research*, 2017. Available at https://github.com/microsoft/LightGBM.

[5] Yajun Liu, Jingjing Fan, Jianfang Zhang, and Xinxin Yin. Research on telecom customer churn prediction based on ensemble learning. *Journal of Intelligent Information Systems*, 2022.

[6] Cheng Wang, Congjun Rao, Fuya Hu, and Xinping Xiao. Risk assessment of customer churn in telco using fclcnn – lstm. *Expert Systems with Applications*, 2024.