Customer Churn Prediction using Machine Learning

Challa Thulasi
COMP 7118-001 – Data Mining

Abstract

Customer churn prediction is a critical task for telecom service providers seeking to retain subscribers and reduce revenue loss. This paper presents a supervised learning framework that addresses the challenges of imbalanced data, heterogeneous feature types, and noise in the Telco Customer Churn dataset. The proposed solution integrates SMOTE-ENN for class rebalancing and noise reduction, followed by a comparative evaluation of multiple machine learning classifiers including K-Nearest Neighbors (KNN), Random Forest (RF), XGBoost, Logistic Regression, and AdaBoost. Hyperparameter tuning is conducted using GridSearchCV with cross-validation to optimize model performance. Evaluation metrics prioritize recall and F1-score to minimize false negatives, crucial in churn scenarios. Experimental results demonstrate that KNN, when paired with proper preprocessing and resampling, achieves superior performance with a recall of 98.74% and F1-score of 0.9728. A lightweight Gradio-based interactive interface was developed to facilitate model interpretability and deployment for business stakeholders. This study confirms that robust preprocessing combined with ensemble and instance-based methods can effectively mitigate churn and support proactive retention strategies in telecommunications.

1 Introduction

Customer churn—defined as the loss of existing customers to competitors—is a critical business concern for service-oriented industries, particularly the telecommunications sector. Retaining customers is substantially more cost-effective than acquiring new ones, with studies estimating that acquiring a new customer can cost five to seven times more than retaining an existing one. Therefore, early identification of churn-prone customers allows organizations to take proactive steps to retain them through personalized offers, improved services, or targeted outreach.

Churn prediction, however, presents several challenges. The data is often imbalanced, with a relatively small proportion of customers who actually churn, making it difficult for standard classifiers to detect the minority class. Additionally, the heterogeneous nature of the data—which includes numerical, categorical, and binary attributes—complicates the modeling process. Furthermore, missing values, feature redundancy, and noise add layers of complexity to the task.

In this project, we address these challenges using a supervised machine learning pipeline. The goal is to build a robust and interpretable churn prediction model using real-world customer data from the IBM Telco Customer Churn dataset. By exploring a combination of preprocessing techniques, class balancing strategies, and multiple machine learning algorithms, we aim to improve predictive accuracy and recall, thereby enabling telecom providers to better allocate retention resources and reduce customer attrition.

2 Related Work

Customer churn prediction has been widely explored in the domains of customer relationship management (CRM), marketing analytics, and data mining. Traditional statistical techniques, such as logistic regression, have been extensively used due to their simplicity and interpretability. However, with the emergence of large-scale customer data and complex patterns in churn behavior, machine learning approaches have gained prominence.

Ensemble methods like Random Forest and Gradient Boosting have shown superior performance in churn prediction tasks by capturing non-linear relationships and interactions between features [Verbeke et al., 2012]. K-Nearest Neighbors (KNN), despite its simplicity, has been used effectively for churn classification in scenarios where local decision boundaries dominate.

A significant challenge in churn modeling is the class imbalance problem, where the number of churned customers is substantially lower than non-churned ones. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) [Chawla et al., 2002] and its extensions (e.g., SMOTE-ENN) have

been employed to oversample the minority class while eliminating noisy samples from the majority class, thereby improving classification accuracy and recall.

Recent studies have also emphasized the importance of cost-sensitive learning and feature engineering to improve churn prediction outcomes [Bahnsen et al., 2015]. In parallel, tools like Gradio have enabled the deployment of machine learning models via interactive user interfaces, enhancing the interpretability and accessibility of churn analytics for non-technical stakeholders [Abid et al., 2019].

Building upon this prior work, our approach integrates data preprocessing, advanced resampling techniques, hyperparameter tuning, and an interactive GUI to create an end-to-end churn prediction pipeline suitable for real-world business deployment.

3 Methodology

3.1 Dataset and Preprocessing

The dataset used in this study is the IBM Telco Customer Churn dataset, which includes 7,043 customer records and 21 features, encompassing customer demographics, service details, billing information, and payment methods. The target variable `Churn` is binary (Yes/No), indicating whether a customer left the service.

Data preprocessing involved handling missing entries in the `TotalCharges` column, which were either imputed or removed based on completeness. Categorical features were transformed into numeric representations using one-hot encoding, and the `customerID` attribute was dropped as it held no predictive power. A new binary flag was introduced to indicate internet service availability. Furthermore, the `TotalCharges` attribute was binned to segment customers into value tiers for later interpretability.

3.2 Addressing Class Imbalance

Churn prediction suffers from severe class imbalance, where non-churning customers significantly outnumber churners. To mitigate this, we used SMOTE-ENN, a hybrid technique combining Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN). SMOTE generates synthetic minority class samples, while ENN removes ambiguous or noisy majority samples, leading to a cleaner, more balanced training set.

3.3 Model Training and Hyperparameter Tuning

We evaluated six machine learning models. All models were trained using 5-fold cross-validation and tuned using GridSearchCV for hyperparameter optimization. Below we describe the configuration and outcomes of each model:

Decision Tree

A Decision Tree classifier was trained with a maximum depth of 10 and a minimum of 20 samples required at each leaf node. This simple, interpretable model achieved an accuracy of 91% and a recall of 94%.

Confusion Matrices for Each Model

Table 1 to Table 6 summarize the confusion matrices for each classification model. The rows represent actual class labels and columns represent predicted labels (0 = No Churn, 1 = Churn).

Table 1: Confusion Matrix — Decision Tree

Actual \ Predicted	0	1
0 (No Churn)	453	65
1 (Churn)	33	602

Random Forest

The Random Forest model used 200 estimators, a maximum depth of 15, and default values for minimum sample split and leaf size. The model demonstrated strong generalization with an accuracy of 94.71%, precision of 93.48%, and recall of 97.17%.

Table 2: Confusion Matrix — Random Forest

Actual \ Predicted	0	1
0 (No Churn)	475	43
1 (Churn)	18	617

XGBoost

XGBoost was configured with a learning rate of 0.2, 200 estimators, and a maximum depth of 7. A subsampling rate of 0.9 was applied to reduce overfitting. The model achieved 94.36% accuracy and maintained high F1 and recall scores.

Table 3: Confusion Matrix — XGBoost

Actual \ Predicted	0	1
0 (No Churn)	480	38
1 (Churn)	27	608

AdaBoost

AdaBoost, using 200 weak learners and a learning rate of 0.2, showed solid results with 92.54% accuracy and 96.54% recall. It served as a strong baseline among boosting methods.

Table 4: Confusion Matrix — AdaBoost

Actual \ Predicted	0	1
0 (No Churn)	454	64
1 (Churn)	22	613

Logistic Regression

Logistic Regression was regularized using an L2 penalty with a regularization strength parameter $C = 10$. Despite its simplicity, the model yielded competitive performance with 93.15% accuracy and 94.80% recall.

Table 5: Confusion Matrix — Logistic Regression

Actual \ Predicted	0	1
0 (No Churn)	473	46
1 (Churn)	33	602

K-Nearest Neighbors (KNN)

The KNN model, using Manhattan distance and $k = 3$, achieved the best performance across all models with an accuracy of 96.96% and recall of 98.74%. Its high sensitivity makes it ideal for capturing churn signals in the data.

3.4 Interactive User Interface

To enhance practical utility, we developed a Gradio-based web interface allowing users to upload datasets, select models, perform predictions, and view evaluation metrics. This makes the solution accessible to business analysts and stakeholders without coding expertise.

Table 6: Confusion Matrix — K-Nearest Neighbors

Actual \ Predicted	0	1
0 (No Churn)	491	27
1 (Churn)	8	627

4 Results

4.1 Model Performance Summary

Table 7 compares the accuracy, precision, recall, and F1-score of the trained models. Among these, the K-Nearest Neighbors (KNN) classifier outperformed others with the highest recall and F1-score, making it most effective for minimizing customer churn.

Table 7: Performance Comparison of Models

Model	Accuracy	Recall	F1-Score
KNN	0.9696	0.9874	0.9728
Random Forest	0.9471	0.9717	0.9529
XGBoost	0.9436	0.9575	0.9493
Logistic Regression	0.9315	0.9480	0.9384
AdaBoost	0.9254	0.9654	0.9345
Decision Tree	0.9100	0.9400	0.9200

4.2 Top 5 Important Features

Feature importance plays a critical role in interpreting model behavior. Based on the KNN model (supported by Random Forest's feature importance rankings), the top five features influencing churn prediction were: TotalCharges, Tenure, MonthlyCharges, Partner (Yes), Dependents (Yes)

These features reflect customer engagement, contract duration, and financial activity — all of which are strong indicators of retention or churn risk.

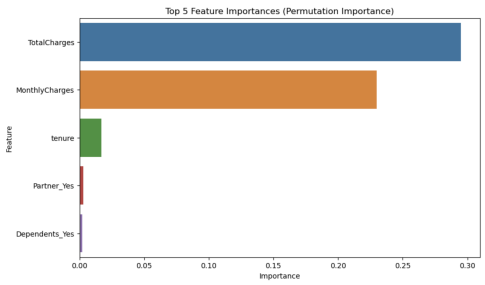

Figure 1: Top 5 Features Contributing to Churn Prediction

4.3 Exploratory Data Analysis (EDA)

EDA was conducted to uncover hidden patterns and key indicators of churn behavior. Visualizations provided actionable business insights and informed feature engineering. Key observations include:

- Customers with month-to-month contracts and fiber optic internet service exhibited higher churn rates.
- Electronic check users were more likely to churn compared to users with automatic or bank payments.
- Customers with high monthly charges and low total charges had a higher probability of churn, suggesting dissatisfaction or lower service commitment.

Figure 2 illustrate these findings.

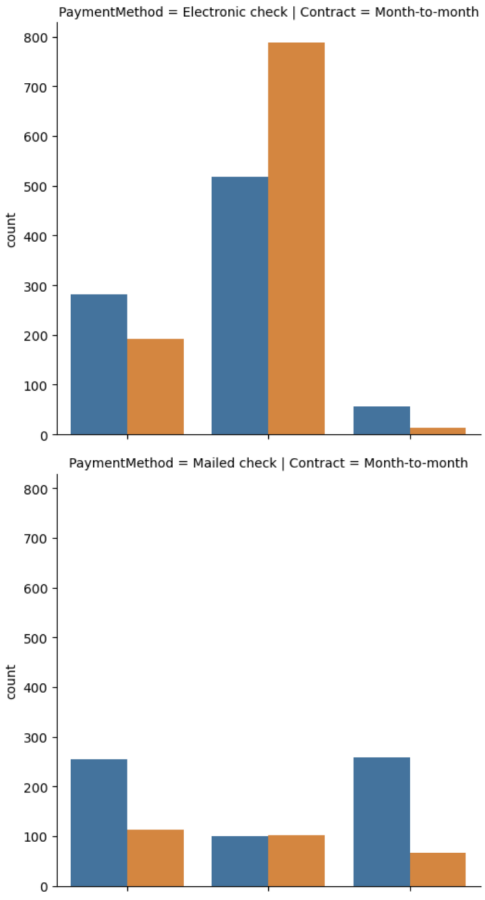

Figure 2: Distribution of Payment method and contract type vs. Churn

4.4 User Interface for Churn Prediction

To make the solution accessible for non-technical stakeholders, an interactive GUI was developed using the Gradio framework. This web-based interface allows users to:

- Upload new customer datasets.
- Perform live churn prediction using the trained models.
- View key EDA insights and select the best-performing model.

Figure 3 shows a snapshot of the deployed GUI on Hugging Face Spaces.

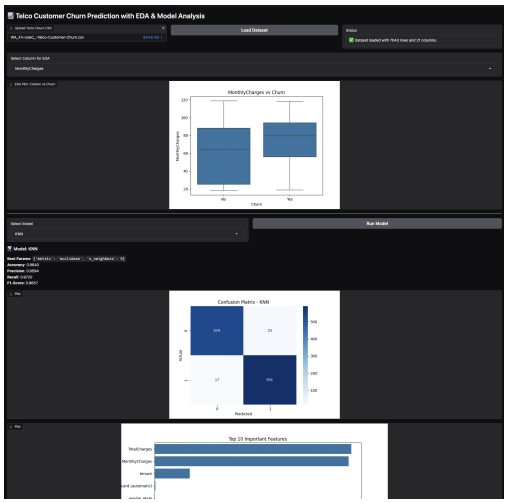

Figure 3: Interactive Gradio UI for Customer Churn Prediction

5 Conclusion

This study presented an end-to-end machine learning pipeline for predicting customer churn in the telecommunications sector using the IBM Telco Customer Churn dataset. We addressed key challenges including class imbalance, heterogeneous feature types, and data noise by employing SMOTE-ENN resampling, robust preprocessing, and multiple classification algorithms.

Among the evaluated models, the K-Nearest Neighbors (KNN) classifier achieved the highest predictive performance, with an accuracy of 96.96% and recall of 98.74%. These results demonstrate the effectiveness of instance-based learning for capturing nuanced churn patterns. Ensemble models such as

Random Forest and XGBoost also showed strong performance, offering a good balance between recall and precision.

Exploratory Data Analysis (EDA) revealed critical churn indicators such as high monthly charges, low total charges, and specific contract and payment types, providing actionable insights for retention strategies. Furthermore, a Gradio-based interactive interface was developed, enabling users to explore EDA, select models, and perform predictions seamlessly—making the system suitable for business deployment.

In summary, our approach effectively combines data preprocessing, resampling, algorithmic tuning, and user-centered interface design to deliver a practical churn prediction solution. Future work may explore cost-sensitive learning, real-time streaming inference, and integration with CRM platforms for proactive customer engagement.

Acknowledgements

Instructor: Dr. Xiaofei Zhang
Course: COMP 7118-001 – Data Mining

References

Abubakar Abid, Ahmed Abdalla, J. Zeng, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild, 2019.

Alejandro Correa Bahnsen, Djamila Aouada, and Björn Ottersten. Example-dependent cost-sensitive logistic regression for credit scoring. In *2015 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 263–269. IEEE, 2015.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Wouter Verbeke, David Martens, and Bart Baesens. Predictive modeling for churn in telecommunications: Revisiting the value of customer segmentation. *Expert Systems with Applications*, 39(10): 9986–9995, 2012.