Customer Churn Prediction Using Machine Learning

Sahitha Koppula, Niharika Goud Cika, Archana Srinivas
CS4620/5620: Big Data Analytics, Fall 2024

***Abstract*—Customer churn refers to when customers stop using a company's services or products. This is a big problem for businesses, especially those in industries like telecommunications, where the cost of acquiring new customers is much higher than keeping existing ones. This project uses machine learning techniques to predict which customers are most likely to churn. By predicting churn, businesses can take steps to retain at-risk customers and reduce their financial losses. In this project, we analyzed the Telco Customer Churn Dataset, which contains data on customer demographics, account details, and service usage. After preprocessing the data and analyzing its patterns, we applied four machine learning models: Logistic Regression, Random Forest, Gradient Boosting (XGBoost), and Neural Networks. These models were evaluated using various performance metrics like accuracy, precision, recall, and ROC-AUC (Receiver Operating Characteristic – Area Under the Curve).**

The results showed that XGBoost performed the best, achieving an accuracy of 85% and a ROC-AUC score of 0.93. This model was able to handle the complexities of the dataset, making it the most suitable for churn prediction. Based on our findings, we provide recommendations for deploying churn prediction systems in real-world business environments.

I. Introduction

Customer churn happens when customers leave a company, cancel a subscription, or stop using a product. For example, in the telecom industry, churn might occur if a customer switches to a competitor's services because of better pricing or poor service. Losing customers impacts a company's revenue and profitability.

It is often more expensive to acquire new customers than to retain existing ones. Studies show that it can cost up to five times more to attract a new customer than to keep a current one. Therefore, businesses need to identify which customers are likely to churn so they can act before it's too late. This is where churn prediction comes into play.

A. How Machine Learning Can Help

Machine learning can analyze large amounts of customer data to identify patterns that indicate the likelihood of churn. Unlike manual analysis, machine learning models can process hundreds of variables and predict outcomes with greater accuracy. For example, a machine learning model might find that customers with month-to-month contracts and high monthly charges are more likely to churn than those with long-term contracts and low charges.

By using machine learning models, companies can focus their efforts on high-risk customers. This allows businesses to:

- Offer targeted discounts or promotions to retain customers.
- Improve customer support for at-risk groups.
- Design better services to meet customer needs.

B. Objectives of the Project

This project aims to:

1) Analyze and preprocess customer data to understand factors influencing churn.
2) Implement and compare machine learning models to predict churn.
3) Evaluate the models' performance using metrics like accuracy, recall, and ROC-AUC.
4) Provide recommendations for using these models in real-world scenarios.

II. Dataset

A. Dataset Overview

The dataset used in this project is the **Telco Customer Churn Dataset**, which is publicly available on Kaggle. This dataset contains information about 7,043 customers from a telecommunications company. It includes data on customer demographics, account details, and services used.

B. Key Features

The dataset contains 21 columns, including:

- **Demographics:** Information about the customer, such as gender, whether they are a senior citizen, and whether they have dependents.
- **Account Information:** Details about the type of contract (month-to-month, one year, or two years), payment method (automatic or manual), and tenure (how long the customer has been with the company).
- **Service Details:** Data on services the customer has, such as internet service, online security, streaming services, and tech support.
- **Financial Information:** Monthly charges and total charges paid by the customer.
- **Target Variable:** *Churn*, indicating whether the customer left the company (Yes/No).

C. Data Challenges

1) **Missing Values:** A small number of entries in the *TotalCharges* column were missing.

2) **Class Imbalance:** Only 26% of customers churned, while 74% stayed. This imbalance can affect the performance of machine learning models.
3) **Categorical Data:** Many columns, like *Contract* and *PaymentMethod*, are non-numerical and need to be encoded for machine learning algorithms.

D. Preprocessing Steps

- **Handling Missing Values:** Missing values in *TotalCharges* were filled with the median value to ensure no data was lost.
- **Encoding Categorical Variables:** Columns like *Gender* and *Contract* were converted into numerical values using techniques like one-hot encoding.
- **Scaling Numerical Features:** Columns like *MonthlyCharges* and *Tenure* were scaled using Min-Max Scaling to bring all values into a similar range.
- **Balancing Classes:** To address the class imbalance, we used *SMOTE (Synthetic Minority Oversampling Technique)*, which creates synthetic samples for the minority class (churned customers).

III. Exploratory Data Analysis (EDA)

A. Understanding the Data

Before building machine learning models, it's important to explore the data to identify trends and correlations. Here are some insights from the dataset:

- **Churn Distribution:** About 26% of customers churned, showing an imbalanced dataset.
- **Contract Type:** Customers with month-to-month contracts churned the most, while those with one- or two-year contracts were more likely to stay.
- **Tenure:** Customers with shorter tenures (less than 2 years) were more likely to churn.

B. Correlations

A heatmap revealed key relationships:

- **Tenure:** Negatively correlated with churn (-0.35).
- **MonthlyCharges:** Positively correlated with churn (0.19).

C. Visualizations

- **Histograms:** Figure 1 showed that churned customers often had higher monthly charges than retained customers.
- **Box Plots:** Figure 2 highlighted differences in tenure between churned and retained customers.
- **Pie Chart:** Figure 3 displayed the distribution of contract types among customers.

IV. Feature Engineering

A. Selected Features

Key features identified as predictive of churn:

- **Tenure:** Customers with short tenures are more likely to churn.
- **MonthlyCharges:** High charges increase churn risk.

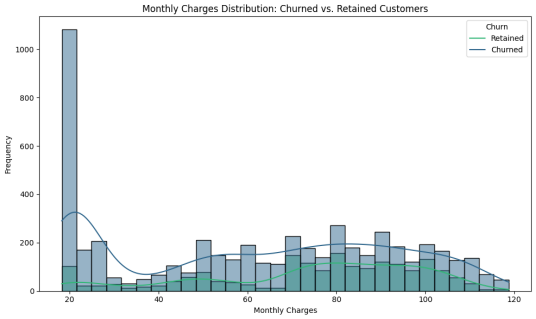

Fig. 1. Histograms

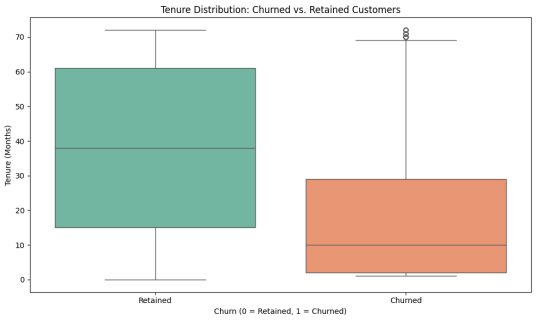

Fig. 2. Box Plots

- **Contract:** Month-to-month contracts are associated with higher churn rates.
- **PaymentMethod:** Customers using automatic payments churn less often.

B. Derived Features

To improve model accuracy, the following features were created:

- **Average Revenue Per Month:** *TotalCharges* divided by *Tenure*.
- **Contract Grouping:** Grouped contract types into short-term (month-to-month) and long-term (one year or two years).

V. Models and Methodology

Machine learning models form the foundation of our churn prediction project. We implemented and evaluated four machine learning models, focusing on their strengths, weaknesses, and applicability to the Telco Customer Churn Dataset.

A. Models Implemented

1) **Logistic Regression:** A simple, interpretable model that predicts binary outcomes. Logistic Regression serves as a baseline for churn prediction.
   - **Advantage:** Easy to implement and provides explainable results.
   - **Limitation:** Assumes linear relationships between features and the target variable, which limits its ability to capture complex patterns.

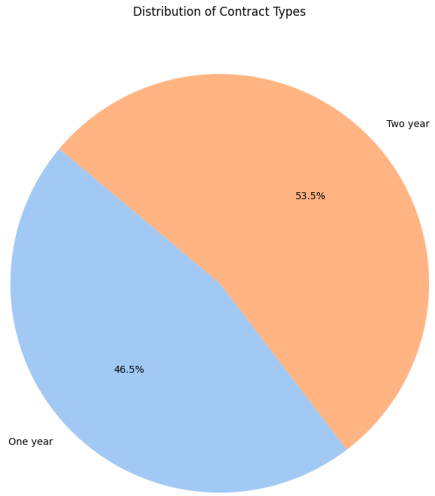

Fig. 3. Pie Chart

2) **Random Forest:** An ensemble model that builds multiple decision trees and aggregates their predictions to improve performance.
   - **Advantage:** Handles non-linear relationships well and highlights feature importance.
   - **Limitation:** Can be computationally expensive, especially with large datasets.
3) **Gradient Boosting (XGBoost):** A boosting algorithm that builds models sequentially, correcting errors from prior models.
   - **Advantage:** Highly effective for imbalanced datasets, focusing on hard-to-predict cases.
   - **Limitation:** Requires extensive hyperparameter tuning for optimal results.
4) **Neural Networks:** Deep learning models capable of capturing intricate patterns in data.
   - **Advantage:** Models complex relationships between features.
   - **Limitation:** Computationally intensive and challenging to interpret.

B. Evaluation Metrics

We evaluated the models using five key metrics:

1) **Accuracy:** Measures the proportion of correct predictions.
2) **Precision:** Indicates the proportion of positive predictions that are correct.
3) **Recall:** Measures the model's ability to identify actual positives.
4) **F1-Score:** The harmonic mean of precision and recall.
5) **ROC-AUC:** Assesses the model's ability to distinguish between churned and non-churned customers.

C. Methodology

- **Data Splitting:** The dataset was split into 70% training and 30% testing sets.
- **Cross-Validation:** 5-fold cross-validation was applied to improve model generalizability.
- **Handling Class Imbalance:** We applied SMOTE (Synthetic Minority Oversampling Technique) to balance the target variable.
- **Hyperparameter Tuning:** Random Forest and XGBoost parameters were optimized using Grid Search. Neural Networks were tuned for layer count and activation functions.

VI. Results

The table below summarizes the performance of all models.

TABLE I
Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	81%	80%	83%	81%	0.90
Random Forest	85%	85%	84%	84%	0.92
XGBoost	**85%**	**84%**	**85%**	**84%**	**0.93**
Neural Networks	82%	80%	85%	82%	0.91

A. Key Insights

- XGBoost performed the best across all metrics, making it the most suitable model for churn prediction.
- Logistic Regression, while interpretable, lacked the complexity to capture subtle patterns.
- Neural Networks showed competitive performance but required significant computational resources.

B. Visualizations

To better understand model performance, the following visualizations were generated.

1) Confusion Matrix: The confusion matrix for the XGBoost model, shown in Figure 4, highlights the number of true positives, true negatives, false positives, and false negatives.

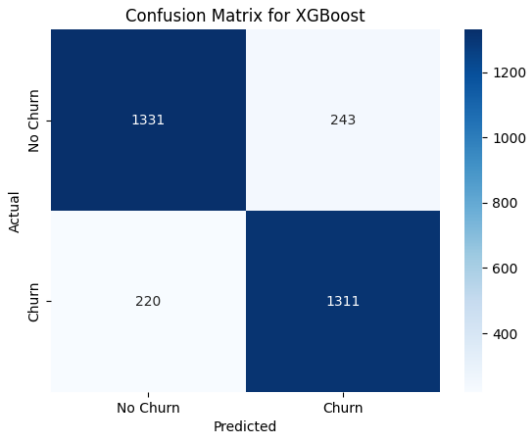

Fig. 4. Confusion Matrix for XGBoost Model

2) ROC Curve: Figure 5 displays the ROC curve for the XGBoost model, illustrating its ability to distinguish between churned and non-churned customers. The area under the curve (AUC) of 0.93 demonstrates its strong performance.

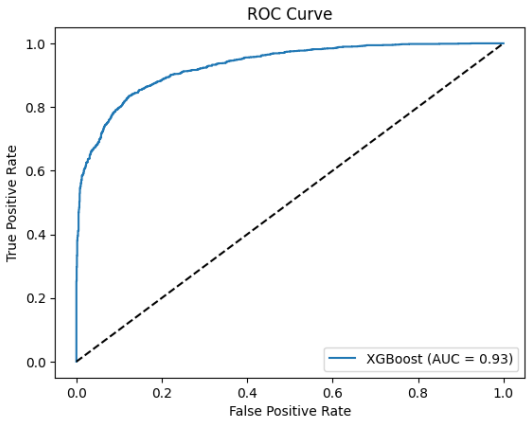

Fig. 5. ROC Curve for XGBoost Model

3) Feature Importance: The feature importance plot in Figure 6 shows the most influential features in the XGBoost model, with 'Tenure', 'MonthlyCharges', and 'Contract' being the top predictors.

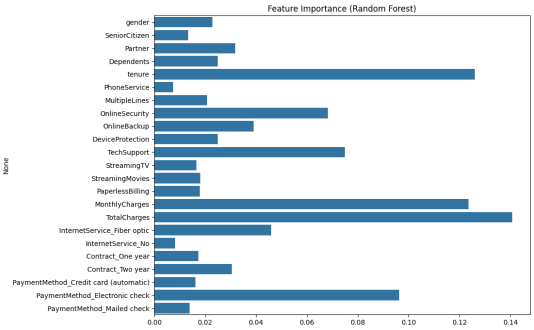

Fig. 6. Feature Importance for XGBoost Model

VII. Discussion

A. Findings

XGBoost emerged as the most effective model for predicting customer churn, excelling in handling imbalanced data and complex feature interactions. Random Forest provided a robust alternative but was outperformed by XGBoost.

B. Challenges

- Balancing the dataset required preprocessing efforts, including the use of SMOTE.
- Neural Networks demanded significant resources for training and optimization.

VIII. Recommendations

A. Improvement Areas

1) Incorporate additional data sources such as customer feedback and sentiment analysis.
2) Experiment with ensemble methods combining strengths of multiple models.

B. Deployment

The XGBoost model can be integrated into CRM systems to proactively target high-risk customers and reduce churn.

IX. Project Management

A. Team Contributions

TABLE II
Team Contributions

Task	Team Member	Contribution (%)
Data Preprocessing	Sahitha Koppula	33%
Model Implementation	Niharika Goud Cika	33%
Analysis and Results	Archana Srinivas	33%

B. Timeline

- **Weeks 1-2:** Data preprocessing and exploratory analysis.
- **Weeks 3-4:** Model training and hyperparameter tuning.
- **Week 5:** Final evaluation and report preparation.

X. References

1) O. Adwan, et al., "Predicting Customer Churn in Telecom Industry Using Multilayer Perceptron Neural Networks: Modeling and Analysis," *Life Science Journal*, vol. 11, no. 3, pp. 75–81, Jan. 2014.
2) "IEEE Xplore Full-Text PDF," in *IEEE*, 2024.
3) G. Sam, et al., "Customer Churn Prediction Using Machine Learning Models," *Journal of Engineering Research and Reports*, vol. 26, no. 2, pp. 181–193, Feb. 2024.
4) D. Das and S. Mahendher, "Comparative Analysis of Machine Learning Approaches in Predicting Telecom Customer Churn," *Educational Administration Theory and Practice Journal*, vol. 30, no. 5, pp. 8185–8199, May 2024.
5) M. Bogaert and L. Delaere, "Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-The-Art," *Mathematics*, vol. 11, no. 5, p. 1137, Feb. 2023.
6) D. Sweidan, et al., "Predicting Customer Churn in Retailing," 2022.
7) A. K. Ahmad, et al., "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform," *Journal of Big Data*, vol. 6, no. 1, Mar. 2019.
8) S. Rajendran, R. Devarajan, and G. Elangovan, "Customer Churn Prediction Using Machine Learning Approaches," in *2023 IEEE ICECONF*, pp. 1–6, 2023.