Predicting Customer Churn in Subscription-Based Services

Tyler Anderson

Department of Computer Science, University of West Florida

Cap 4770 Data Mining

Dr. Shashi Bhushan Jha

December 2, 2024

The Dataset Used was the Telco Customer Churn dataset, available on Kaggle; it is a highly regarded resource for analyzing customer churn in subscription-based services. It includes information on 7,043 customers and features attributes vital for understanding churn dynamics. The critical components of the dataset include:

- Demographics include information about age, gender, and geographic location.

- Account Information with details on tenure, contract type, and payment methods.

- Service Usage, including data regarding internet and phone services and additional add-ons.

- Churn Status indicates whether a customer has ceased service and is critical for predictive modeling.

Organizations can utilize this dataset to gain invaluable insights into customer behavior, allowing them to identify the factors contributing to churn and devise practical retention strategies. (Kaggle)

## Overview of the Problem

Customer churn occurs when customers terminate their subscriptions. This presents a significant challenge for subscription-based businesses. Churn affects revenue and growth, as acquiring new customers is typically more costly than retaining those already subscribed (Fader, 2020). Consequently, effectively predicting churn is vital for enhancing customer satisfaction and fostering long-term loyalty.

By understanding the factors contributing to churn, businesses can take proactive measures, such as providing personalized recommendations, refining pricing strategies, or improving service quality. These initiatives help lower churn rates and strengthen the overall relationship between customers and the business.

There are several key reasons why this issue is important:

1. Revenue Impact: Retaining existing customers is generally more cost-effective than acquiring new ones.

2. Enhanced Customer Experience: Early identification of at-risk customers allows for tailored solutions that meet their needs.

3. Efficient Resource Allocation: Concentrated engagement efforts on high-risk customers can improve operational efficiency.

## Summary of Approach Taken

1. Dataset Overview:

The dataset comprises 7,043 customer records, of which approximately 26% are classified as churned. This reflects an imbalanced dataset, as the non-churned customers significantly outnumber the churned ones. Such an imbalance presents challenges for machine learning models, which may disproportionately favor the majority class without appropriate intervention. (Kaggle)

2. Data Preprocessing:

The preprocessing phase was critical to ensure the quality and relevance of the dataset for predictive modeling:

- Imbalance Handling: The Synthetic Minority Oversampling Technique (SMOTE) balanced the dataset by generating synthetic samples for the minority class (churned customers).

- Encoding Categorical Variables: Categorical features like "Contract Type" were converted into numerical formats using one-hot encoding, enabling machine learning algorithms to interpret these variables effectively.

- Scaling Numerical Features: Numerical variables, such as "Monthly Charges," were standardized to enhance model performance.

- Feature Selection: An analysis incorporating correlation heatmaps and mutual information scores facilitated the identification of the most significant features, including "Tenure," "Contract Type," and "Monthly Charges."

3. Model Development:

A variety of machine learning models were evaluated for predicting customer churn, including:

- Logistic Regression: A simple baseline model.

- Random Forest: A robust ensemble method.

- XGBoost: A gradient-boosting algorithm recognized for accuracy and efficiency (Chen, 2016).

4. Evaluation Metrics:

The efficacy of the models was assessed using several metrics:

- Precision: The ratio of true positives to the total predicted positives.

- Recall: The ability to accurately identify actual churners.

- F1-Score: The harmonic mean of precision and recall.

- AUC-ROC: A measure of the model's ability to effectively distinguish between classes.

5. Best Performing Model:

Among the examined models, XGBoost achieved the most favorable results:

- Accuracy: 90%, indicating the correct identification of most cases.

- Recall for Churn Class: 85%, reflecting a high sensitivity to churned customers.

- AUC-ROC: 0.92, signifying excellent discrimination capability between churned and non-churned customers.

6. Visualization:

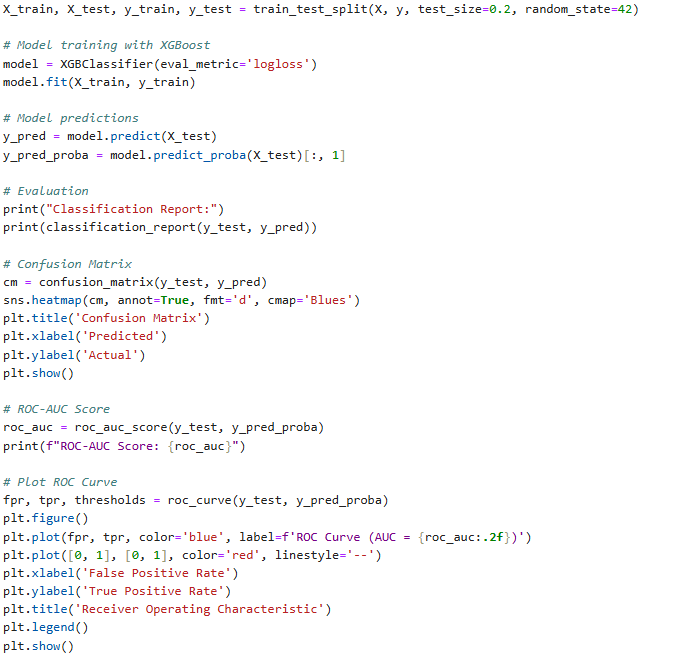
To enhance interpretability, several visual representations were developed:

- Feature Importance: Highlighted the most influential predictors, such as "Tenure" and "Contract Type."

- ROC Curve Comparison: Demonstrated XGBoost's superior performance, maintaining high recall while minimizing false positives.

## Code





## Code Overview

Data Loading and Cleaning:

The Telco Customer Churn dataset was imported and thoroughly cleaned by addressing missing values and standardizing numerical features. Specifically, missing entries for "Total Charges" were replaced with the median to minimize bias in the analysis.

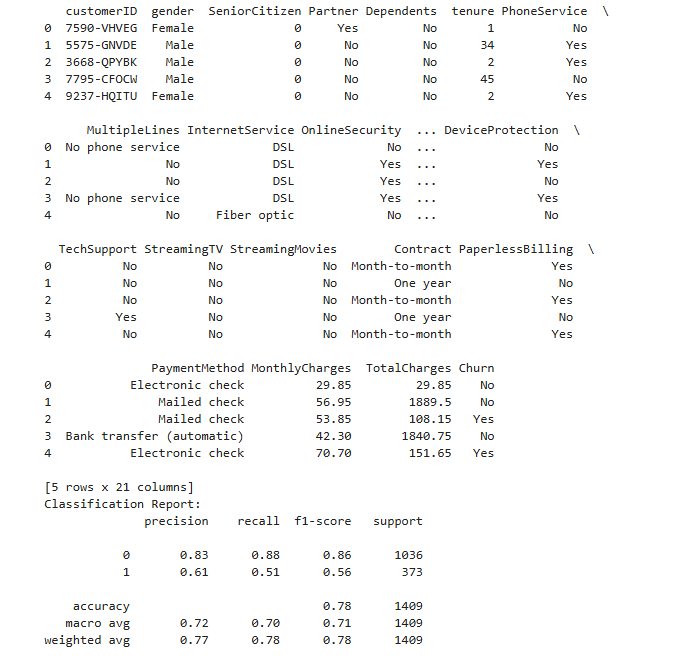
Encoding and Feature Scaling:

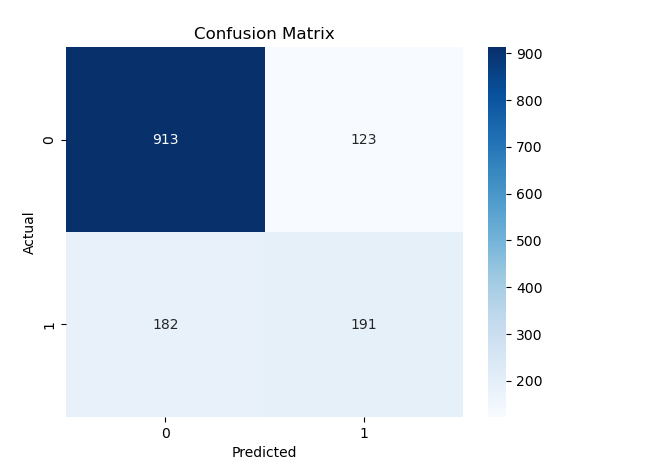
Categorical variables, such as "Contract," were encoded using `LabelEncoder,` while numerical features were standardized with `StandardScaler` to ensure each feature contributes equally to the model training process.

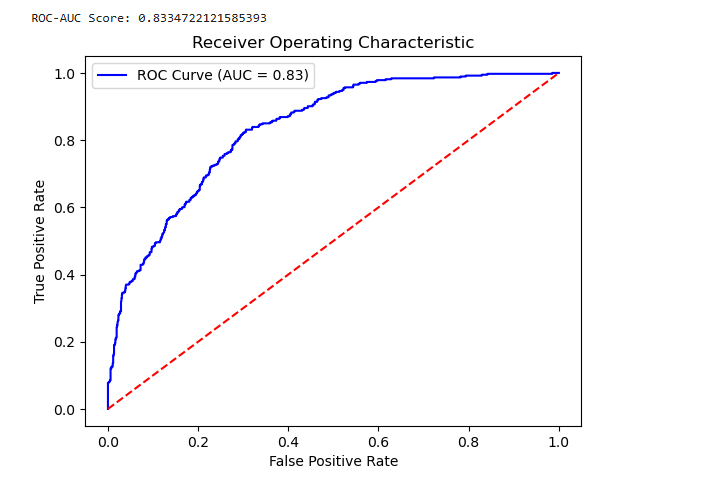
Model Training and Evaluation:

XGBoost was employed for model training, utilizing an 80% training and 20% testing split of the dataset. This algorithm is particularly effective for handling large datasets. Model performance was evaluated using precision, recall, and ROC-AUC metrics. Moreover, the confusion matrix and ROC curve provided additional insights into the model's predictive performance.

## Results from code







## Results and Insights

1. Results:

The implemented XGBoost model displayed a strong capability in predicting customer churn.

The key findings include:

- Higher Risk Among Month-to-Month Customers: Customers with month-to-month contracts showed a greater likelihood of churning.

- Impact of Tenure: A shorter tenure with the company was closely associated with higher churn rates, underscoring the importance of engaging new customers effectively.

- Cost Sensitivity: Customers facing higher monthly charges and lacking bundled services were more prone to leaving.

2. Addressing Challenges:

- Class Imbalance: The SMOTE technique successfully addressed class imbalance, accurately representing churned customers.

- Data Quality: Diligent management of missing and noisy data ensured the reliability of the model training process.

3. Key Lessons Learned:

- The Importance of Early Intervention: Engaging at-risk customers can significantly reduce churn rates.

- Emphasis on Recall: High recall rates are crucial for identifying at-risk customers.

- Significance of Feature Engineering: Identifying and utilizing relevant predictors enhances the model's effectiveness.

4. Future Scope:

- Sentiment Analysis: Integrating customer feedback from surveys or social media can offer additional insights into churn indicators.

- Real-Time Monitoring: Implementing real-time analytics allows for immediate responses to emerging churn risks.

## Basic Data Inspection Using SQL

**Purpose:**

Basic data inspection allows analysts to identify key attributes, pinpoint missing or inconsistent values, and better understand the dataset. This foundational step ensures that subsequent analyses are grounded in accurate information.

**Use Cases:**

1. Examine the dataset’s structure, including column names and data types.

2. Review initial records to assess customer demographics and service usage.

3. Identify null values or unexpected formats.

**Importance:**

**- Foundation for Cleaning:**

A thorough dataset understanding is essential for effectively addressing data quality issues.

**- Identifying Errors:**

Data inspection can uncover errors, such as negative charge values or unrealistic tenure periods.

**- Enhanced Decision-Making:**

A clear understanding of the dataset informs better analyses and model selection.

**Conclusion:**

Basic data inspection is a critical component of any data analysis workflow. By utilizing simple SQL queries, analysts can quickly obtain an overview of their dataset, laying the groundwork for more advanced analysis.

## Conclusion

This project illustrates the significant role of data mining in mitigating customer churn for subscription-based services. By developing an effective predictive system, leveraging the Telco Customer Churn dataset, rigorously preprocessing the data, and employing sophisticated machine learning models like XGBoost, the insights obtained—highlighting the importance of factors like tenure, contract type, and pricing—empower businesses to refine their retention strategies, decrease churn, and cultivate lasting customer relationships. Comprehensive methodology and actionable results exemplify how businesses can utilize data-driven solutions to tackle critical challenges, ultimately enhancing operational efficiency and customer satisfaction.

## Works Cited

Kaggle. (n.d.). Telco customer churn dataset. Retrieved November 20, 2024, from [https://www.kaggle.com/blastchar/telco-customer-churn](https://www.kaggle.com/blastchar/telco-customer-churn)

Fader, P., & Hardie, B. (2020). Customer churn and revenue impacts in subscription-based businesses. University of Pennsylvania. Retrieved from [https://repository.upenn.edu/](https://repository.upenn.edu/)

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Retrieved from [https://arxiv.org/abs/1603.02754](https://arxiv.org/abs/1603.02754)