Predicting Customer Churn Using CRISP-DM: A Data-Driven Approach

Your Name

November 17, 2024

Abstract

Customer churn poses a critical challenge for businesses, leading to significant revenue losses and increased customer acquisition costs. This paper demonstrates a step-by-step application of the CRISP-DM methodology to predict customer churn using the Telco Customer Churn dataset. By leveraging machine learning models, we achieved 81% accuracy and identified key factors influencing churn, such as internet service type and total charges. The insights derived from this study can help businesses develop targeted retention strategies.

1 Introduction

Customer churn, the phenomenon where customers discontinue using a company's services, directly impacts profitability and operational efficiency. Understanding and predicting churn is crucial for businesses to develop effective retention strategies.

This research applies the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** methodology, a widely adopted framework for structured data mining, to build a predictive model for customer churn. Using the **Telco Customer Churn dataset** from Kaggle, we aim to:

- Build a machine learning model with at least 80% accuracy.
- Identify key features influencing churn to provide actionable business insights.

2 Methodology

2.1 CRISP-DM Overview

CRISP-DM is a six-phase methodology for data mining:

- **Business Understanding**: Define objectives and success criteria.
- **Data Understanding**: Explore the dataset and assess data quality.

- **Data Preparation**: Clean, transform, and encode the data for modeling.
- **Modeling**: Train and test machine learning models.
- **Evaluation**: Assess model performance and insights.
- **Deployment**: Prepare the model for real-world applications.

2.2 Dataset Description

The **Telco Customer Churn dataset** contains 7,043 rows and 21 columns, capturing customer demographics, subscription details, and churn status. The target variable is `Churn` (Yes/No), with a class imbalance of 26% churners and 74% non-churners.

3 Data Preparation

3.1 Preprocessing Steps

Key preprocessing steps included:

- **Handling Missing Values**: Imputed missing values in the `TotalCharges` column with the median.
- **Encoding Categorical Features**: Applied one-hot encoding to features like `InternetService` and `PaymentMethod`.
- **Scaling Numerical Features**: Standardized `tenure`, `MonthlyCharges`, and `TotalCharges` to ensure uniform scaling.
- **Removing Irrelevant Columns**: Dropped `customerID` as it provided no predictive value.

3.2 Final Dataset

The preprocessed dataset contains 7,043 rows and 31 features, ready for modeling.

4 Modeling

4.1 Models Tested

Three machine learning models were used:

- Logistic Regression: A simple yet effective classification algorithm.
- Random Forest: A robust ensemble learning method.
- XGBoost: A gradient boosting algorithm known for high performance.

4.2 Results

The **Logistic Regression** model achieved the best performance with:

- **Accuracy**: 81%
- **Precision**: 67%
- **Recall**: 56%
- **F1-Score**: 61%
- **ROC-AUC**: 0.8448

Other models (Random Forest and XGBoost) performed slightly below Logistic Regression.

4.3 Feature Importance

Feature importance analysis from Logistic Regression revealed the top factors influencing churn:

1. **InternetService_Fiber optic**: Customers with fiber optic internet were more likely to churn.
2. **TotalCharges**: Higher total charges correlated with lower churn rates.
3. **PaperlessBilling_Yes**: Customers with paperless billing showed higher churn tendencies.

5 Evaluation

5.1 Confusion Matrix

The Logistic Regression model correctly predicted the majority of `No Churn` customers but showed moderate recall for `Churn` customers, highlighting room for improvement in identifying high-risk individuals.

5.2 Limitations

The recall for churners was only 56%, indicating that further techniques, such as oversampling or ensemble learning, could enhance the model's ability to detect churn.

6 Business Insights

6.1 Key Findings

- **Internet Service**: Customers with fiber optic internet are at higher risk of churning, suggesting the need for targeted retention campaigns.
- **Billing Preferences**: Customers using paperless billing may face challenges that lead to churn, necessitating improved communication and support.
- **Total Charges**: Customers with higher total charges churn less, indicating loyalty among long-term customers.

6.2 Recommendations

- Offer customized retention offers to customers using fiber optic internet.
- Address customer concerns about paperless billing through enhanced support.
- Prioritize engagement with new customers to build long-term relationships.

7 Conclusion and Future Work

This study successfully applied the CRISP-DM methodology to predict customer churn, achieving an accuracy of 81%. The analysis provided actionable insights to guide retention strategies, such as focusing on high-risk customers and addressing issues related to internet services and billing preferences. Future work could involve exploring advanced techniques like SMOTE for better recall and integrating the model into real-time business systems.

8 References

- Kaggle: Telco Customer Churn Dataset. `https://www.kaggle.com/blastchar/telco-customer-c`
- CRISP-DM Framework. IBM SPSS Modeler Documentation. `https://www.ibm.com/docs/en/spss-modeler`
- Logistic Regression Documentation. scikit-learn. `https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`