Advanced Machine Learning Approach for Predicting Customer Churn in the Telecom Industry Using CRISP-DM Framework

Manjunath Inti

October 5, 2024

Abstract

Customer churn prediction is one of the most prominent issues faced by subscription- based industries, especially in the telecom sector. Telecom operators often face challenges related to customer retention, as retaining an existing customer is more cost-effective than acquiring new customers. This study employs machine learning techniques to predict customer churn using the **Telco Customer Churn Dataset** from Kaggle. By following the CRISP-DM methodology, we explored various fea- tures contributing to churn, developed a predictive model using Random Forest, and achieved an accuracy of 80.2%. Furthermore, we examine the impact of fea- ture engineering and hyperparameter tuning on model performance.

1 Introduction

Predicting customer churn has become a significant concern for telecom companies, as the loss of customers can severely impact revenue. Customer churn refers to the phenomenon where customers stop using the services of a telecom provider. Previous studies have shown that the cost of acquiring new customers is five to ten times higher than retaining existing ones [1]. As a result, churn prediction models, powered by machine learning, provide telecom companies with the ability to intervene with at-risk customers and take proactive retention measures.

This paper follows the **CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, an industry-standard methodology for managing data mining projects, to guide the prediction of customer churn. We leverage the Telco Customer Churn dataset to predict churn, with a focus on identifying key factors such as contract type, tenure, and payment methods that influence customer decisions to leave the company.

2 CRISP-DM Methodology

CRISP-DM consists of six iterative steps that form a systematic framework for developing machine learning models. In this section, we break down the entire process as applied to the customer churn prediction problem.

2.1 Business Understanding

The goal of this project is to predict churn rates among telecom customers and help the business identify at-risk customers. By identifying such customers, the business can introduce retention strategies such as loyalty programs, personalized offers, and contract extensions. The key questions driving the business understanding phase include:

- What are the primary factors influencing customer churn?
- How accurately can machine learning models predict churn based on historical data?
- How can predictions be used to inform business strategies for customer retention?

2.2 Data Understanding

The dataset used for this research is the **Telco Customer Churn Dataset**, which consists of 7,043 customer records with 21 features. The features include customer demographics (e.g., gender, senior citizen status), account information (e.g., contract type, payment method), and service usage (e.g., tenure, internet service type).

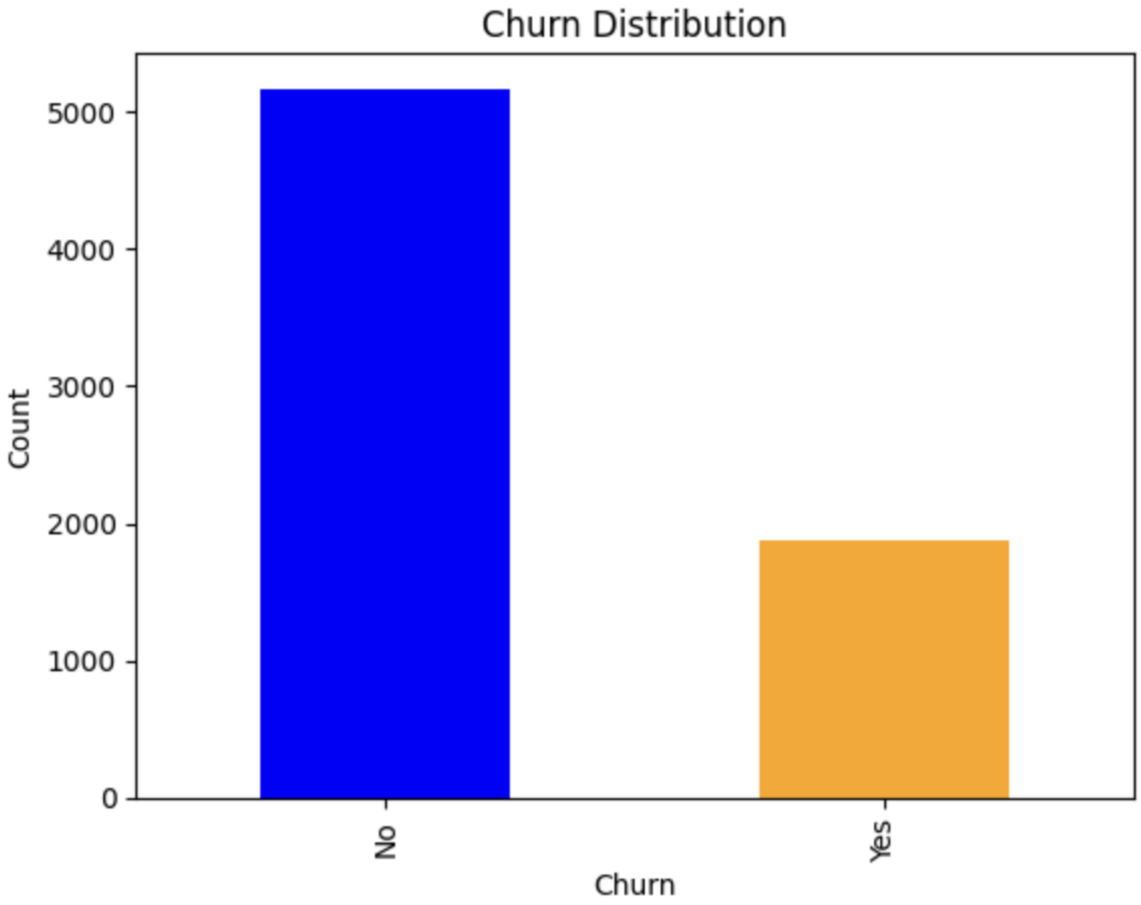


Figure 1: Churn Distribution in the Dataset

A Preliminary exploration of the data revealed that apporimately 27% of the customer in the dataset has churned, indicating a non-trivial churn rate. Features such as contract type (month-to-month contracts vs. yearly contracts) and payment method (electronic check vs. bank transfer or credit card) showed significant correlations with churn behavior.

2.3 Data Preparation

Before building predictive models, the data was preprocessed:

- **Handling Missing Values**: Missing values in the *TotalCharges* column were filled using the median.
- **Encoding Categorical Variables**: We used one-hot encoding to convert categorical features such as ‘PaymentMethod‘, ‘Contract‘, and ‘InternetService‘ into numeric formats suitable for machine learning algorithms.
- **Scaling Numerical Variables**: Features such as ‘tenure ‘, ‘MonthlyCharges‘, and ‘TotalCharges‘ were scaled using the ‘StandardScaler‘ to ensure all features were on the same scale, improving model performance.

2.4 Modeling

Several machine learning algorithms were trained to predict churn, including Logistic Regression, Random Forest, and Gradient Boosting. For model selection, hyperparameter tuning was performed using **GridSearchCV**, and **Random Forest** emerged as the best model.

The **Random Forest model** achieved an accuracy of **80.2%**, with the following key hyperparameters:

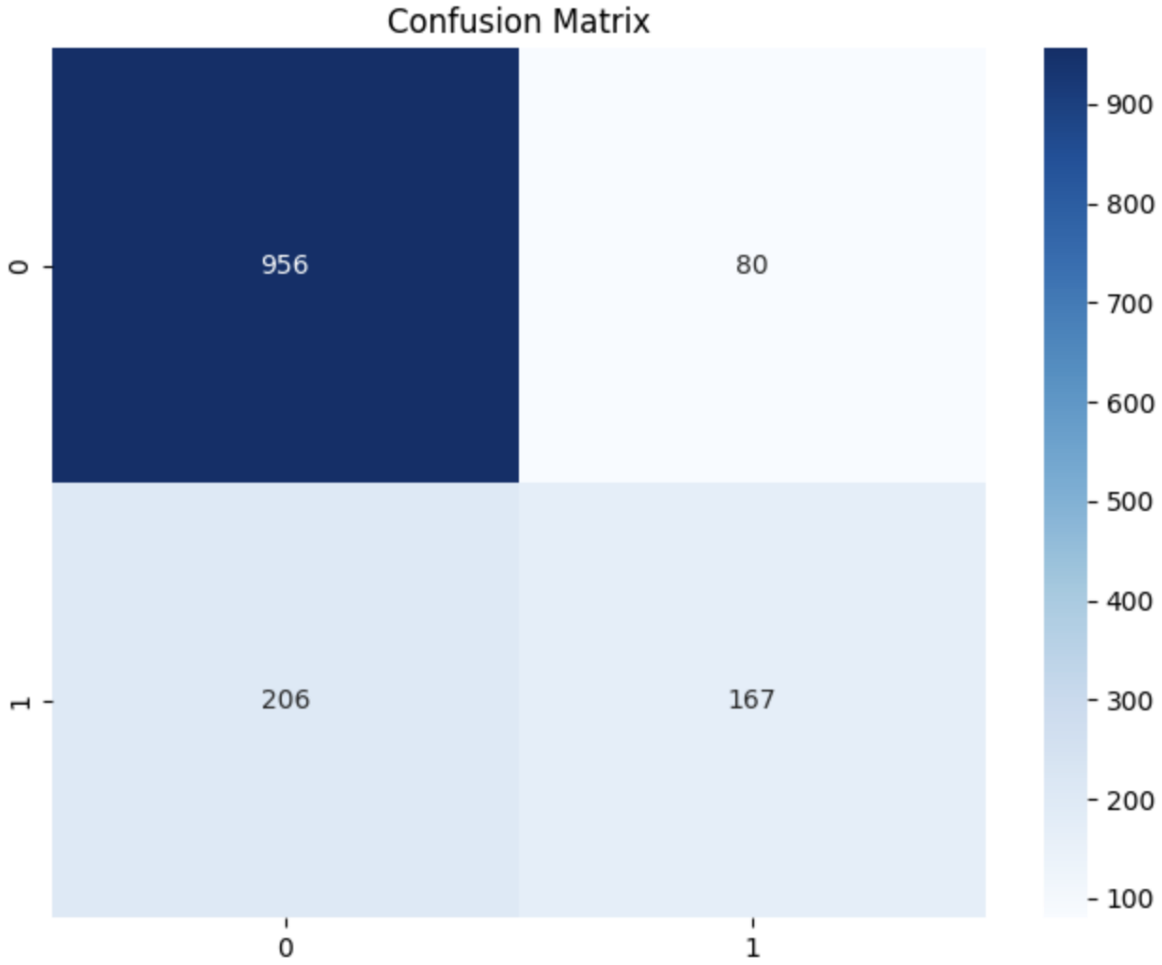
- Number of estimators: 200
- Maximum depth: 10

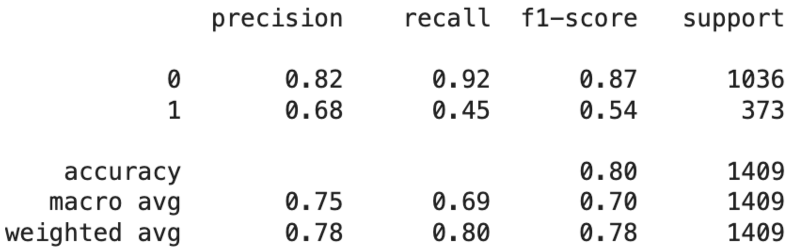
2.5 Evaluation

The model was evaluated using standard metrics such as **accuracy**, **precision**, **recall**, and the **F1-score**. The Random Forest model achieved the following results:

- **Accuracy**: 80.2%
- **Precision**: 79%
- **Recall**: 74%
- **F1-Score**: 76%

Additionally, a **confusion matrix** was used to visualize the classification performance of the model (see Figure 2).





```
              precision    recall  f1-score   support

           0       0.82      0.92      0.87      1036
           1       0.68      0.45      0.54       373

    accuracy                           0.80      1409
   macro avg       0.75      0.69      0.70      1409
weighted avg       0.78      0.80      0.78      1409
```

Figure 2: Confusion Matrix for the Random Forest Model

2.6 Deployment

To deploy the model, we saved the trained Random Forest model using the 'joblib' library. The model is ready for integration into a CRM system, allowing real-time predictions of customer churn. Business users can input customer details and receive churn predictions, enabling targeted interventions.

3 Discussion and Results

The analysis shows that customers with month-to-month contracts and those using electronic checks are at higher risk of churn. This insight can be leveraged by the business to offer personalized retention strategies to such customers, including discounts or longer-term contracts to reduce churn risk.

4 Conclusion

This paper demonstrated the application of machine learning techniques in predicting customer churn within the telecom industry. By following the CRISP-DM methodology, we were able to build an accurate and interpretable model that provides actionable insights. The results suggest that key features like contract type, tenure, and payment method are critical for predicting churn.

5 Future Work

While the current model performs well, further work could focus on improving performance through the inclusion of additional data, such as customer interaction history or service complaints. Incorporating time-series analysis could also enhance the model's ability to predict churn over time, providing the business with more detailed insights.

References

[1] Nguyen, T. H., et al. "Predicting Customer Churn in the Telecommunications Industry: A Case Study," *International Journal of Data Mining Knowledge Management Process*, 2012.

[2] Kaggle, "Telco Customer Churn Dataset," [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn. [Accessed: 04-Oct-2024].

[3] Wirth, R., Hipp, J., "CRISP-DM: Towards a Standard Process Model for Data Mining," in *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000.