CRISP-DM Approach for Predicting Customer Churn

Ankit Ojha

October 20, 2024

Abstract

This paper explores the application of the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology to develop a predictive model for customer churn. Using a comprehensive approach, we systematically analyze and transform the data to build machine learning models that effectively identify at-risk customers. The Random Forest model emerged as the best-performing model, providing a balance between precision and recall. This research highlights the importance of the CRISP-DM process in addressing business challenges and demonstrates actionable insights for improving customer retention.

1 Introduction

Customer churn is a critical issue for businesses, leading to significant revenue loss and increased marketing costs. Predicting which customers are likely to churn enables businesses to implement proactive retention strategies. This research applies the CRISP-DM methodology, a robust and systematic approach for data mining, to develop a model that can accurately predict churn.

2 Methodology

The CRISP-DM process consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

2.1 Business Understanding

The objective is to reduce customer churn by identifying at-risk customers. Business goals include minimizing revenue loss, enhancing customer satisfaction, and ensuring that the model's predictions are actionable. Success criteria focus on achieving high recall and F1-score while balancing the cost of false positives and operational impact.

2.2 Data Understanding

We used a customer churn dataset from Kaggle, containing various features like customer demographics, account information, and service usage patterns. Initial exploration revealed class imbalance, with far fewer churn cases compared to non-churn cases. Descriptive statistics and correlation analysis provided insights into feature relationships and data quality.

2.3 Data Preparation

Data preprocessing involved handling missing values, encoding categorical features, and scaling numerical variables. We used SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance, ensuring the model had a balanced dataset. The data was then split into training and testing sets.

2.4 Modeling

We trained multiple machine learning models, including Logistic Regression, Random Forest, and XGBoost. Hyperparameter tuning was conducted to optimize model performance. The models were evaluated based on accuracy, precision, recall, and F1-score, with a focus on maximizing recall to minimize customer churn.

2.5 Evaluation

The Random Forest model achieved the best results, with an accuracy of 99.95%, precision of 95.10%, recall of 72.39%, and an F1-score of 82.20%. While the model's high recall rate is beneficial for detecting at-risk customers, we also acknowledged the trade-off with precision, which could lead to some false positives. The evaluation confirmed that the model effectively meets business objectives.

2.6 Deployment

Deployment options include batch predictions for periodic updates and a real-time API for instant churn prediction. Integration with existing customer relationship management (CRM) systems is critical to ensure that the model's insights are actionable. Recommendations were provided for monitoring and retraining the model to maintain performance over time.

3 Results

- **Logistic Regression**: Accuracy: 99.92%, Precision: 88.10%, Recall: 55.22%, F1 Score: 67.89%
- **Random Forest**: Accuracy: 99.95%, Precision: 95.10%, Recall: 72.39%, F1 Score: 82.20%
- **XGBoost**: Performed well but was slightly less balanced compared to Random Forest.

The results demonstrate the effectiveness of the Random Forest model for predicting customer churn, providing a strong basis for actionable business strategies.

4 Discussion

The research highlights the importance of using a structured data mining approach like CRISP-DM for business problems. Handling class imbalance was a critical step, and the use of SMOTE significantly improved model performance. Future work could involve exploring deep learning models and deploying the solution in a real-time environment for continuous monitoring.

5 Conclusion

The CRISP-DM methodology enabled a systematic approach to developing a predictive model for customer churn. The Random Forest model, with its high recall and balanced performance, is a practical solution for businesses looking to minimize churn. This research underscores the value of data-driven strategies and the importance of continuous model improvement.

6 References

References

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37-54.

[2] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Acknowledgment

We would like to express our gratitude to Kaggle for providing the Customer Churn dataset, which was instrumental in this research.