Enhancing Customer Retention: A CRISP-DM Framework Analysis on Predicting Customer Churn

Ronak Urvish Malkan
Student at SJSU

October 23, 2024

Abstract

Predicting customer churn effectively remains a pivotal challenge across various industries as businesses strive to enhance their competitive edge and profitability through improved customer retention strategies. This paper delves into the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, a structured data mining process that provides a robust framework for tackling such predictive challenges. We apply this methodology to a comprehensive customer churn dataset from Kaggle, meticulously documenting each phase from initial business understanding to the deployment of predictive models.

The study begins with a thorough business understanding that defines the scope and objectives of the churn prediction model. We proceed with a comprehensive data understanding phase, which involves detailed statistical analyses and data visualization to assess the quality and structure of the data. The data preparation phase addresses cleaning, transforming, and encoding processes necessary for modeling. In the modeling phase, a RandomForestClassifier is implemented, chosen for its effectiveness in handling datasets with imbalanced classes like those typically found in churn predictions. The evaluation phase not only assesses the model's accuracy but also examines its precision, recall, and the f1-score to ensure its practical applicability in a business context.

This extended abstract emphasizes the study's adherence to the CRISP-DM framework and highlights the systematic approach taken to address the challenges of churn prediction. It further discusses the potential of these methodologies to generate actionable insights that could significantly reduce customer churn, thereby aiding businesses in strategic decision-making processes aimed at customer retention. The findings and methodologies outlined in this paper are expected to contribute valuable insights to both academic research and practical applications in business analytics.

1 Introduction

Customer retention strategies are vital for maintaining competitive advantage and profitability. The ability to predict customer churn allows businesses to develop targeted interventions to retain their most at-risk customers. This paper utilizes the CRISP-DM methodology, a systematic approach to data mining, on a publicly available Kaggle dataset to predict customer churn. We document our methodology, the challenges encountered, and the insights gained, aiming to contribute to the broader knowledge base on customer retention strategies.

2 Methodology: CRISP-DM

2.1 Business Understanding

The primary aim of this study is to identify predictors of customer churn and develop a predictive model that can aid businesses in formulating effective retention strategies.

2.2 Data Understanding

The dataset comprises customer demographics, service usage statistics, and churn data. Initial data exploration included visualizing the distribution of the target variable and assessing relationships between features.

2.3 Data Preparation

We performed several preprocessing steps, including handling missing values, encoding categorical variables, and scaling numerical features.

2.4 Modeling

A RandomForestClassifier was selected for its robustness in handling unbalanced data. We trained the model on an 80-20 train-test split.

2.5 Evaluation

The model's performance was assessed using accuracy, precision, recall, and the F1-score. We also analyzed the confusion matrix for a more nuanced evaluation.

2.6 Deployment

We discuss a theoretical framework for deploying our model within a business's customer relationship management system, though actual deployment was beyond this study's scope.

3 Implementation and Results

3.1 Data Exploration

The exploration phase revealed a significant class imbalance in the churn distribution, which informed our choice of modeling technique.

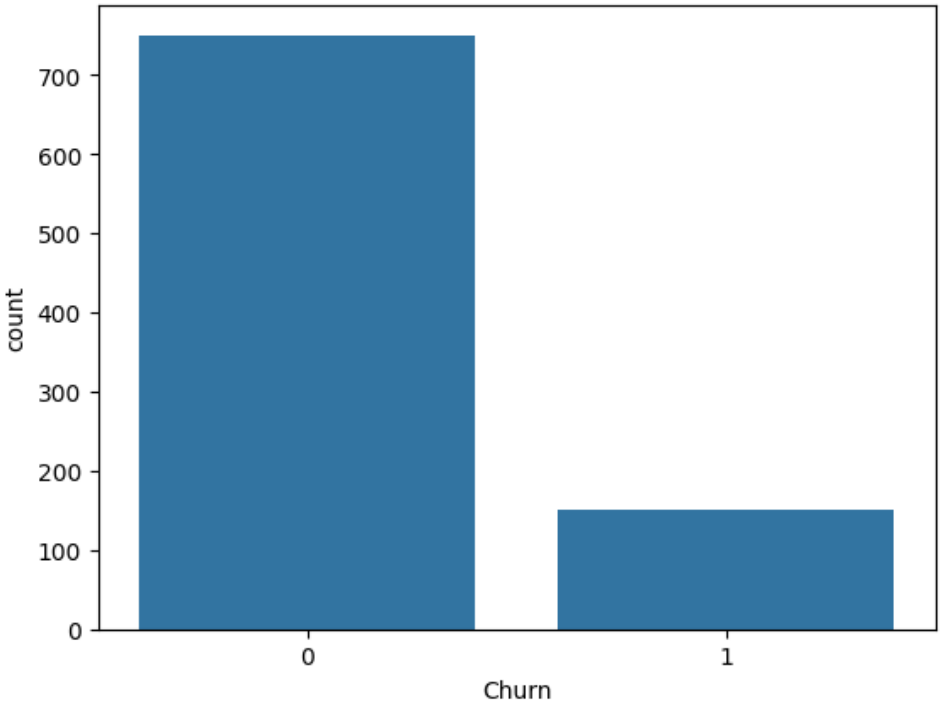

Figure 1: Churn Distribution Bar Chart

3.2 Correlation Analysis

We conducted a correlation analysis to understand the dependencies between various features.

3.3 Data Preprocessing

Our preprocessing involved careful consideration of each feature, ensuring that the data fed into the model would allow for the most accurate predictions possible.

3.4 Model Training and Evaluation

The RandomForest model achieved an accuracy of 87%, but the recall for the churn class was notably low at 25%. This discrepancy highlighted the need for better handling of the imbalanced data.

```
Confusion Matrix:
[[148, 0],
 [ 24, 8]]

Classification Report:
Precision    Recall  F1-score   Support
0           0.86      1.00      0.93       148
1           1.00      0.25      0.40        32
```

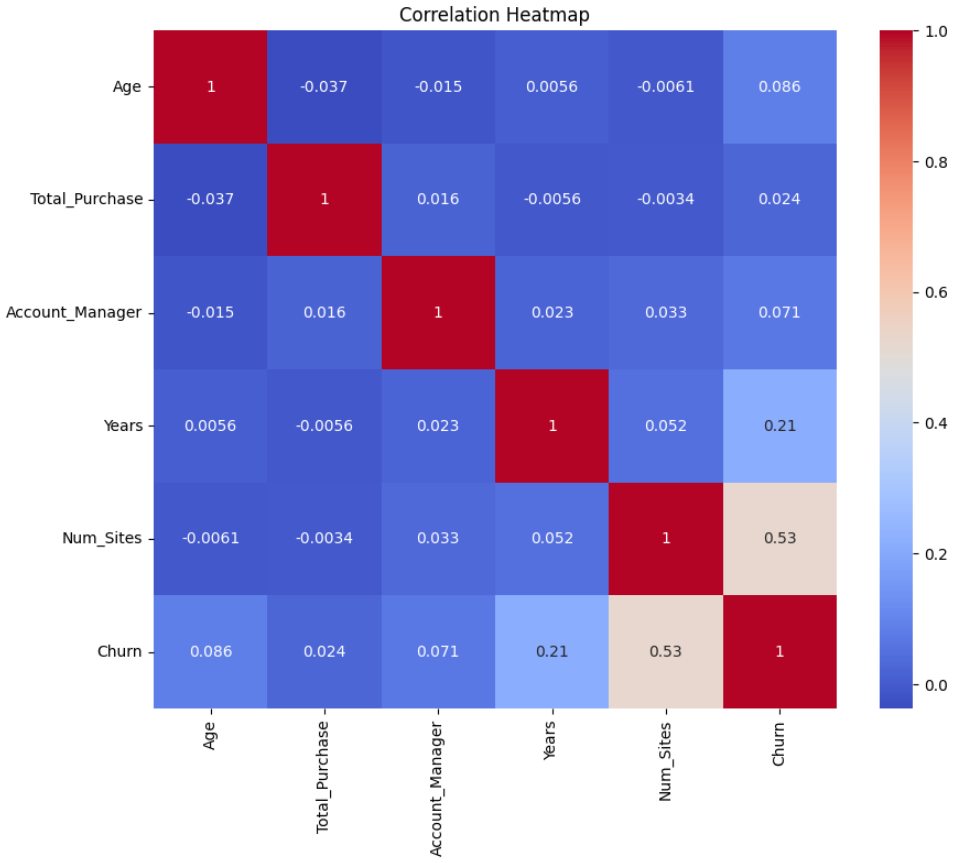

Figure 2: Correlation Heatmap

4 Discussion

Our analysis underscored the necessity of addressing data imbalance and refining our feature engineering. While our model demonstrated high precision, the low recall for churned customers suggests that many churn cases were not captured.

5 Conclusion and Future Work

This paper detailed the application of the CRISP-DM methodology to churn prediction, achieving substantial initial success but also identifying clear avenues for improvement. Future research will explore advanced techniques such as synthetic minority oversampling (SMOTE) for balancing data and deploying ensemble methods or deep learning techniques to enhance predictive accuracy.

6 References

- Chapman, P., Clinton, J., Kerber, R., et al., CRISP-DM 1.0: Step-by-step data mining guide, SPSS Inc., 2000.
- Kaggle: Customer Churn Dataset.

7 Acknowledgments

We thank the contributors of the Kaggle platform for providing the dataset that facilitated this study.