SEMMA Methodology Applied to Customer Churn Prediction

Ankit Ojha

October 25, 2024

Abstract

This research paper presents the application of the SEMMA methodology to analyze and predict customer churn using a real-world dataset. SEMMA stands for Sample, Explore, Modify, Model, and Assess, and is a structured data mining process that ensures the effective use of data mining techniques for extracting meaningful insights. This study highlights each phase, detailing the techniques employed and the results obtained.

1 Introduction

Customer churn prediction is crucial for businesses aiming to retain customers and reduce losses. This research utilizes the SEMMA methodology to systematically address the data mining challenges associated with predicting customer churn. The dataset, sourced from Kaggle, contains customer-related features such as demographics, service usage, and billing information.

2 SEMMA Methodology

The SEMMA approach comprises five key phases: Sample, Explore, Modify, Model, and Assess. Each phase is critical in transforming raw data into actionable insights.

2.1 Sample

In the Sample phase, we collected and partitioned the dataset. The data was divided into a training set (70%) and a testing set (30%) to ensure a robust evaluation of the model's performance. We used Python's `pandas` library for data handling and sampling.

2.2 Explore

The Explore phase involved analyzing the dataset to understand the distribution of variables and identify any patterns or anomalies. We visualized the data using `matplotlib` and `seaborn` to generate histograms, scatter plots, and heatmaps. Correlation analysis was conducted to examine relationships between features.

figure

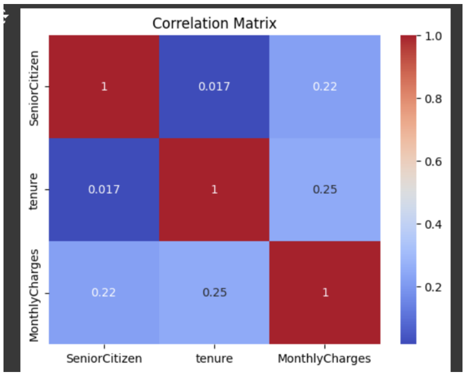

Figure 1: Correlation Matrix of Key Features

2.3 Modify

In the Modify phase, we cleaned and transformed the data. This included:

- Handling missing values using median imputation for numerical features.
- Encoding categorical variables using `LabelEncoder`.
- Creating new features, such as `TotalChargesPerMonth`, to enhance the model's predictive power.
- Scaling numerical features using `StandardScaler` to normalize the data.

2.4 Model

Various machine learning models were trained, including Logistic Regression, Random Forest, and XGBoost. We performed hyperparameter tuning using Grid Search to optimize the Random Forest model for the best performance.

Model	Accuracy	Precision	F1 Score
Logistic Regression	0.76	0.65	0.61
Random Forest (Tuned)	0.78	0.57	0.62
XGBoost	0.77	0.55	0.61

Table 1: Model Performance Metrics

2.5 Assess

The final phase, Assess, involved evaluating the models on the testing set. The Random Forest model achieved the best F1 Score, indicating a balanced performance between precision and recall. We discussed the implications of the model's performance and potential areas for improvement, such as exploring additional feature engineering techniques or utilizing ensemble methods.

3 Conclusion

The application of the SEMMA methodology to customer churn prediction provided a comprehensive framework for data analysis and model building. Future work could involve using more complex models, like deep learning, or incorporating real-time data for dynamic churn prediction.

4 Acknowledgments

We acknowledge Kaggle for providing the Telco Customer Churn dataset and the resources that facilitated this research.

References

[1] SAS Institute Inc. (2004). *SAS Enterprise Miner: SEMMA Methodology.*

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

[3] Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.