1]Hasan Mhowwala [1]San Jose State University

Analyzing User Behavior on a Streaming Platform: An Empirical Approach using SEMMA

[

September 30, 2023

Abstract

The burgeoning significance of streaming platforms has shaped contemporary entertainment paradigms. This paper seeks to discern patterns and behaviors associated with user interactions on these platforms using the SEMMA (Sample, Explore, Modify, Model, Assess) methodology. Delving into this dataset provides insights into user preferences and offers a predictive model for subscription renewals, integral for retention strategies in such platforms.

1 Introduction

Streaming platforms have undergone an unprecedented rise in user adoption. Gleaning insights from user behavior on these platforms can offer enhanced user experiences and foster improved retention strategies. This research harnesses the SEMMA methodology to analyze and predict user behaviors concerning subscription renewals.

2 SAMPLE: Data Overview

Our dataset, comprising 1,000 records and 8 distinctive columns, encapsulates varied user interactions on the platform, such as viewing durations, device preferences, and subscription renewal behaviors. A preliminary analysis revealed a predominant age bracket of 2544, a proclivity towards movie consumption, and the simultaneous use of 2 to 3 devices by many users.

3 EXPLORE: In-depth Analysis

Graphical visualizations elucidated intricate relationships within the data. For instance, longer subscriber tenure appears to be positively correlated with the likelihood of subscription renewals.

4 MODIFY: Data Preparation

To prepare for machine learning applications, we underwent data transformation processes. Categorical attributes, including age demographics and preferred content types, were transformed using one-hot encoding techniques. Furthermore, numerical attributes underwent standardization to ensure homogeneity in scale.

5 MODEL: Predictive Analysis

Our focal objective was to prognosticate the propensity of users renewing their subscriptions. A binary classification approach was chosen, leveraging the Logistic Regression algorithm. Post segmenting the data into training (80

6 ASSESS: Model Evaluation

Further evaluation using the Precision-Recall Curve exhibited an average precision score of 0.850. Both the ROC and Precision-Recall curves provided a comprehensive perspective of the model's performance in diverse classification scenarios. The elevated average precision score is indicative of the model's prowess in achieving high precision and recall concurrently.

7 Conclusion

Employing the SEMMA framework enabled a structured and comprehensive analysis of the streaming platform dataset. This research underscores the value of methodical exploration, meticulous data preparation, and rigorous model evaluation. It provides an exemplar for aspiring data scientists, illuminating the importance of a methodological approach to data analytics in real-world scenarios.

8 Acknowledgments

Gratitude is extended to all individuals and entities contributing to the dataset's acquisition and providing tools for this research.