An In-depth Analysis of Song Popularity Using the SEMMA Methodology on the Spotify 2023 Dataset

Sri Vinay Appari

September 29, 2023

Contents

1 Abstract

This research paper delves deep into the intricacies of the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology applied to data science, specifically targeting the prediction of song popularity. By employing the Spotify 2023 dataset as a foundation, this paper provides a thorough step-by-step analysis, emphasizing the importance of each phase in the SEMMA methodology.

2 Introduction

The music industry has, for decades, been intrigued by the enigma of song popularity. Historically, determining a song's success was more art than science. However, in the era of data, predictive analytics offers the possibility of forecasting song popularity with significant accuracy. Utilizing the rich dataset from Spotify for 2023, this research takes a structured approach using the SEMMA methodology, which stands as a testament to methodical data analysis.

3 Methodology

The SEMMA methodology, a systematic approach to data analysis, consists of five pivotal stages.

3.1 Sample

Sampling is not just about data collection but ensuring the gathered data is representative of the entire population. For this study, a comprehensive dataset from Kaggle encapsulating various song attributes from Spotify in 2023 was used. This dataset promises diversity and depth, crucial for any analytical exploration.

```
import pandas as pd
spotify_data = pd.read_csv('/mnt/data/spotify-2023.csv')
```

3.2 Explore

Exploratory Data Analysis (EDA) is the bedrock of any data-driven research. It's not merely about gleaning basic statistics but understanding the essence of the data—its trends, patterns, and anomalies.

```
# Fundamental statistics and missing value analysis
basic_stats = spotify_data.describe(include='all')
```

```
missing_values = spotify_data.isnull().sum()
```

With 953 unique tracks and a plethora of features ranging from artist count to musical properties, our initial foray unveiled missing data points in the 'key' and 'in_shazam_charts' columns. This discovery underscores the importance of thorough EDA before any predictive modeling.

3.3 Modify

Any data, regardless of its source, is seldom clean. The modification phase focuses on transforming the raw data into a format suitable for modeling. This includes handling missing data, and potential outliers, and ensuring all data is in a numerically interpretable format.

```
# Data transformation and encoding
spotify_data['key'].fillna(key_mode, inplace=True)
spotify_data['in_shazam_charts'].fillna(shazam_median,
    inplace=True)
```

3.4 Model

At the heart of predictive analytics, modeling is where raw data gets transformed into actionable insights. Leveraging the power of various machine learning algorithms, this research aimed to predict song popularity, defining "hit" songs based on their streaming percentile.

```
# Machine learning model training
from sklearn.tree import DecisionTreeClassifier
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)
```

3.5 Assess

While modeling provides predictions, assessment validates their accuracy. This phase employed a suite of metrics to evaluate model performance, ensuring robustness and reliability.

4 Discussion and Recommendations

The high accuracy achieved by tree-based models, while promising, comes with caveats. Potential overfitting, a common pitfall with complex models, warrants further valida-

tion, possibly through techniques like cross-validation. Moreover, while the current feature set was identified based on correlation, domain-specific insights could lead to the discovery of other influential features, further enhancing model accuracy.

5 Conclusion

Through the structured lens of the SEMMA methodology, this research illuminated the potential of data-driven approaches in predicting song popularity. With rigorous exploration, meticulous data modification, robust modeling, and thorough assessment, the research underscores the essence of structured data analysis.

6 References

1. SEMMA: A Comprehensive Guide, SAS Institute.
2. Top Spotify Songs 2023 Dataset. Available at: `https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023`.