Predicting Song Popularity: Exploring Machine Learning Techniques and Spotify Data Analysis

ABSTRACT

This paper aims to investigate the factors contributing to song popularity and develop a predictive model using machine learning techniques. The dataset utilized in this study is sourced from kaggle.com and consists of 11,450 songs spanning from 1986 to 2023. Data analysis was first conducted and multiple regression models, such as linear regression, ridge regression, Lasso regression, decision tree regression, random forest regression, and principal component analysis were explored. We also plan to employ feature selection techniques to identify the most influential variables affecting song popularity. The outcomes of this research endeavor will provide insights into the predictors of song popularity and offer a predictive model to estimate the popularity level of new songs.

KEYWORDS

Music analytics, Data mining, Feature engineering, Machine learning, Predictive modeling

1 Introduction

Music is an integral part of people's lives as well as a profitable commercial field. Understanding what makes a song popular is a fascinating endeavor. In this project, we aim to explore the factors influencing song popularity and predict the popularity of new songs using machine learning techniques.

We will utilize a dataset from kaggle.com, sourced from Spotify's API, consisting of 11,450 songs spanning from 1986 to 2023 obtained from Spotify's API. The dataset comprises 37 features, including both numerical and categorical variables related to song characteristics, artist information, and release details. This dataset provides a wealth of information, including tempo, genre, key, mood of the lyrics and other intrinsic factors to the music itself, as well as the artist who released the music and release information.

Our approach involves employing regression models such as linear regression, ridge regression, Lasso regression, decision tree regression, random forest regression, and principal component analysis. We also utilize feature selection techniques to identify the key variables impacting song popularity.

By predicting song popularity, this research can offer valuable insights for the music industry, aiding in decision-making processes for song production, marketing strategies, and music recommendations.

In this paper, we present the data analysis and preprocessing, machine learning methods and results of our exploration of song popularity.

2 Data Processing

2.1 Dataset Introduction

The dataset contains 36 columns with one target 'popularity' and 35 possible features to predict the popularity of the song. The distribution of our target is relatively normal with a mean around 68 and a total range from 44 to 100.

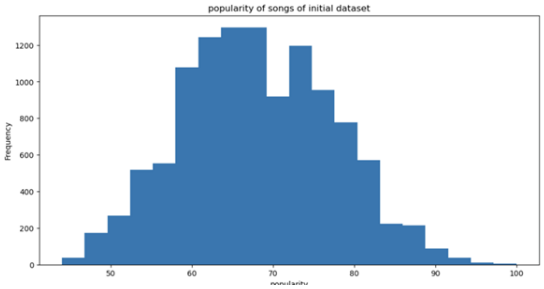
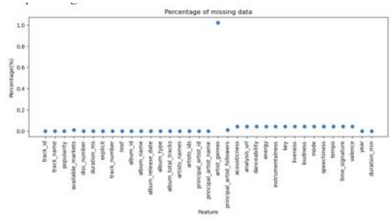


Figure 2.1 distribution of target

2.2 Data Cleaning and Sampling

To ensure the quality and integrity of our data, we performed cleaning and sampling processes. The following steps were taken:

We examined the presence of missing data in the dataset. Most features had complete data, except for 16 features with missing proportions less than 1%. Additionally, we identified 5 samples that had missing values for all evaluations related to the style and characteristic of the song. To mitigate the impact of missing data on our model, we dropped these samples.



feature name	missing data
available_markets	11449/11450
artist_genres	11333/11450
principal_artist_followers	11449/11450
acousticness	11445/11450 same 5 samples
analysis_url	
danceability	
energy	
instrumentalness	
key	
liveness	
loudness	
mode	
speechiness	
tempo	
time_signature	
valence	

Figure 2.2 missing data analysis

Certain features were deemed unnecessary for our analysis or there exist other features serving the same purpose and were dropped from the dataset. These features included the track ID, analysis URL, track name, album ID, album name, artist IDs, principal artist ID, principal artist name, duration in milliseconds, album release date, and time signature.

We retained certain features in their original format as they provided valuable information for our analysis.

Table 2.1 feature kept original

feature name	explanation
track_number	disc number the song belongs to on an album
album_total_tracks	The total number of songs on the album
key	12 keys, The musical key of the song.
loudness	negative, The loudness of the song.
tempo	The tempo of the song
acousticness	The number of followers of the principal artist on Spotify
danceability	An indicator of how danceable the song is.
energy	The perceived energy of the song.
liveness	likelihood that the song was performed live.
mode	2 categories, 0 and 1
speechiness	amount of speech in the song.
valence	A measure of the positivity of the song.
year	The year in which the song was released
duration_min	The duration of the song in minutes

2.3 Feature Encoding and Transformation

To prepare the dataset for modeling, we performed feature encoding and transformation. The technologies we used include Categorical Feature Encoding, Feature Transformation, Feature Scaling.

Table 2.2 Feature Transformation

feature name	encode or transformation
explicit	one hot encoding: True/False
album_type	one hot encoding: Album/Single/Compalation
available_markets	sum the population of all available countries
artists_names	count the number of artists
artist_productivity	number of songs the artists produced
artist_genres	count the number of genres
principal_artist_followers	log transformation: very skewed
instrumentalness	log transformation: very skewed

2.4 Final Dataset Building

we computed the correlation matrix to identify relationships between features. We observed a positive correlation of 0.7 between energy and loudness, indicating that songs with higher energy tend to have higher loudness. Conversely, there was a negative correlation of -0.64 between energy and acousticness, suggesting that songs with higher energy tend to have lower acousticness. The absolute correlation value between 'energy' and 'loudness', 'energy' and 'acousticness' and 'explicit_True' and 'speechness' are higher than 0.5. As a result, we dropped 'energy' and 'explicit'.

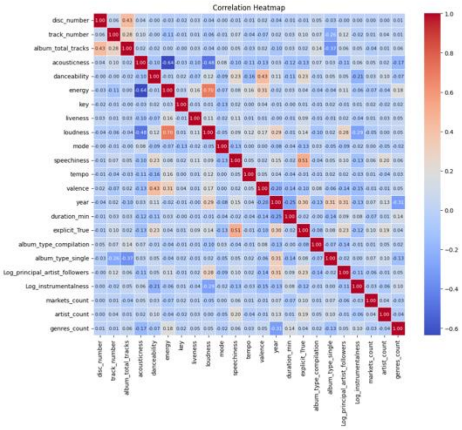


Figure 2.3 correlation

We also implemented a standard scale for numerical features as some of our models are sensitive to the scale of data while others, like tree-based models, are not.

The final dataset consists of 11327 samples with 20 features.

3 Data Analysis

In this section, we present the results of our data analysis, including visualizations and insights derived from the dataset.

3.1 Continuous Numerical Features

We examined the distribution of the continuous numerical features in the dataset. Notably, features such as liveness, acousticness, and tempo were found to be highly skewed to the right. Additionally, the feature markets_count exhibited extreme values, being either very large or very small.

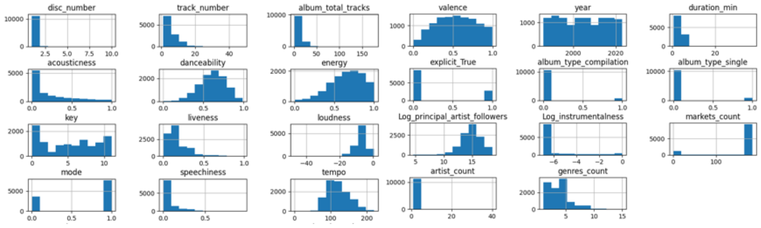


Figure 3.1 Numerical feature distribution

3.2 Boxplot Analysis

To gain further insights, we conducted a boxplot analysis for both categorical and numerical variables.

For the categorical variable disc_number, an outlier was identified with a value of 10.

The variable Explicit showed a slightly higher mean popularity when set to True, while having numerous outliers with high popularity when set to False.

The category Single had the highest mean popularity, while the means for Album and Compilation Album were similar.

However, the 75th percentile was higher for Album, indicating a higher popularity range.

Regarding the mode variable, both modes had similar means, but mode = 1 exhibited a higher number of outliers with high popularity.

An outlier was identified for the variable artist_count, with a value of 40. Generally, the mean popularity increased with more artists involved, but when there were too many artists, the mean popularity decreased compared to a single artist.

The number of genres had an inverse relationship with popularity; as the number of genres increased, the mean popularity decreased. However, the lowest popularity levels seemed to be similar regardless of the number of genres.

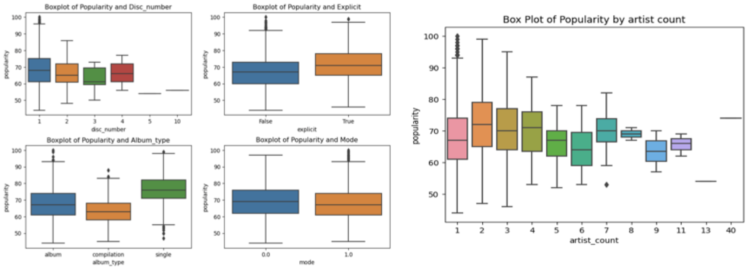


Figure 3.2 Boxplot of categorical features

4 Machine learning models and results

This is a regression problem and we used several regression models and test their performance on our dataset.

(1) Models

Potential models include Linear Regression, Decision Tree, Random Forest, Gradient Boosted Tree and SVM. Within Linear Regression, we compared model with or without norms such as L1, Lasso and Elastic-Net. When using SVM, we utilized 2 kernels, linear and rbf, respectively.

(2) Evaluation Metrics

We used Mean Squared Error (MSE), Mean Absolute Error (AME) and Coefficient of Determination (R^2) to evaluate the performance of our models.

(3) Hyperparameter Tuning

Some of the models have hyperparameters, we use GridSearchCV with a k-fold split of 5 to search the optimal hyperparameters minimizing the mean squared error of our training dataset.

The overall performance of our models can prove that it is feasible to use machine learning methods to predict the popularity of a new-released song.

By comparing the evaluation metrics of models, Random Forest works best with a test MSE of 36.34 and a training MES of 6.89. The result is satisfactory comparing with the range of the data from 44 to 100. The corresponding hyperparameters are max depth of individual tree is 18, max features using is 0.3 with total estimator of 110. XGBoost model also performs well. The linear models have relatively higher training error comparing indicating the relationship between our features and target may not be exactly linear and the ability of linear models to reflect the underlying distribution is limited. The SVM and DNN models also work well having a slightly better performance than linear models. Both linear kernel and rbf kernel have a testing MSE around 41. On the other hand, the Decision Tree can easily overfit the data and have a max test MSE.

Table 4.1 Model results

		Train Result			Test Result		
Model	Hyper parameter	MSE	MAE	R2	MSE	MAE	R2
Linear Regression		39.56	5.06	0.55	41.29	5.21	0.52
Ridge Regression	alpha=0.42	39.56	5.06	0.55	41.27	5.21	0.52
Lasso Regression	alpha=0.022	39.57	5.06	0.55	41.34	5.21	0.52
Elastic-net Regression	alpha=0.022 l1_ratio=1.0	39.57	5.06	0.55	41.34	5.21	0.52
Decision Tree	max_depth=17	3.69	0.84	0.96	68.37	6.33	0.21
Random Forest	max_depth=18 max_features=0.3 n_estimators=110	6.89	2.09	0.92	36.34	4.86	0.58
SVM	kernel= rbf gamma= 0.001 C=464.16	37.63	4.81	0.57	40.75	5.09	0.53
XGBoost	learning_rate=0.1 max_depth=4 max_iter=30	30.25	4.40	0.65	37.19	4.95	0.56
DNN	structure				39.08	5.05	0.55

5 Conclusion

This study investigates the variables that impact the popularity of songs and constructs a predictive model utilizing machine learning methodologies. By utilizing regression models and feature selection techniques, we conducted an analysis on a dataset consisting of 11,450 songs obtained from Spotify's API.

The results of our study unveiled significant associations between the attributes of songs and their level of popularity. There was a positive correlation between higher energy and loudness, and a negative correlation between higher energy and acousticness. By employing feature selection, we were able to concentrate on crucial variables, thereby enhancing the precision of our predictive models.

Random Forest outperformed the other models that were tested by accurately predicting song popularity. The linear models and Support Vector Machines (SVM) also produced satisfactory outcomes. These insights have ramifications for the music industry, facilitating decision-making in song production, marketing, and recommendations.

Our research enhances the field of music analytics by employing advanced algorithms and leveraging Spotify data to accurately forecast the popularity of songs. Additional research can enhance the accuracy of the models and investigate the influence of changing musical trends on popularity.