University of Milan

The Master of Science in “Data Science for Economics”

Machine Learning and Statistical Learning

MACHINE LEARNING PROGECT:

RIDGE REGRESSION ON SPOTIFY TRACKS DATASET

Professor: Nicolò Cesa-Bianchi
Student: Ekaterina Sergeeva

Milan, 2024

Content

1. Introduction

The aim of this project is to predict the popularity of tracks using the Spotify Tracks Dataset, available on Kaggle.

1.1 Dataset:

This dataset encompasses over 100,000 songs, featuring both categorical and numerical attributes. Categorical features include artist names, album titles, and song genres, while numerical features encompass various song characteristics like danceability, acousticness, and duration. Additionally, two binary features indicate whether the song contains explicit lyrics and its modality (major or minor). The target variable, song popularity, is rated on a scale from 0 to 100 based on the total number of plays and their recency.

1.2 Methodology:

To accomplish the task, we employ Ridge regression initially using only numerical features and subsequently incorporating both categorical and numerical attributes. Prior to model training, the dataset underwent appropriate transformations to handle categorical data effectively. Furthermore, to compute reliable risk estimates, we employed 5-fold cross-validation. Additionally, we explored alternative algorithms to assess their performance relative to Ridge regression. Each phase of the methodology will be elaborated upon in subsequent sections.

In this report, we provide a detailed overview of our approach, the transformations applied, the evaluation of the models through cross-validation, and a discussion on the performance of Ridge regression compared to other algorithms tested. Further insights into the code implementation will be presented in the final section.

2. Data Cleaning and Transformation

During the initial data exploration phase, several challenges were identified, necessitating thorough cleansing and transformation to ensure the dataset's suitability for predictive modeling. This section elucidates the strategies employed to rectify these issues.

2.1 Handling Duplicate Tracks

One of the primary challenges encountered was the presence of duplicate tracks. These duplicates arose due to songs being listed both individually and as part of albums or compilations. Consequently, the same song was represented multiple times with identical technical characteristics but varying popularity ratings, posing challenges for subsequent analysis. To address this:

- **Rounding Numerical Columns:** All numerical attributes were rounded to two decimal places to mitigate discrepancies arising from minor variations.
- **Removing Low Popularity Tracks:** Among duplicate tracks with identical technical features, only the entry with the highest popularity was retained. Additionally, tracks with a popularity rating below 10 were removed to ensure dataset reliability.

2.2 Handling Missing Values

To ensure data integrity, missing values were identified and subsequently removed from the dataset.

2.3 Managing Categorical Features

Another significant challenge revolved around handling categorical features, such as artists, albums, and genres. Due to the large number of occurrences, traditional one-hot encoding was deemed impractical. Instead, target encoding was employed, wherein categorical values were replaced with their corresponding average popularity. Further adjustments were made to address specific issues within categorical features:

2.3.1 Resolving Multi-genre Songs

Certain tracks were labeled with multiple genres, leading to ambiguity. To streamline the dataset, only the most frequent genre for each track was retained.

2.3.2 Managing Multi-artist Songs

Tracks featuring multiple artists posed challenges in attributing popularity. To simplify, only the most popular artist associated with each track was retained, based on their average popularity across songs.

2.3.3 Addressing Homonymous Albums

Some albums shared identical titles despite being distinct projects from different artists. By leveraging the first artist listed in the 'artists' column, unique albums were identified, resolving potential ambiguities.

3. Ridge Regression: Exploring Models for Popularity Prediction

To predict track popularity, this project adopts Ridge regression, initially focusing on numerical features and subsequently incorporating both numerical and categorical attributes.

3.1 Ridge Regression and Cross Validation

3.1.1 Ridge Regression: Enhancing Linear Models

Ridge regression is a sophisticated extension of linear regression, designed to mitigate issues like multicollinearity and overfitting. It achieves this by introducing a regularization term into the model, which imposes constraints on model coefficients. This regularization balances bias and variance, controlled by the hyperparameter alpha. The Ridge regression equation takes the form:

$$w = (S^T S + \alpha I)^{-1} S^T y$$

Where:

- α\alpha denotes the regularization parameter.
- S represents the training example matrix.
- y is the vector of labels.
- w is the vector of weights.

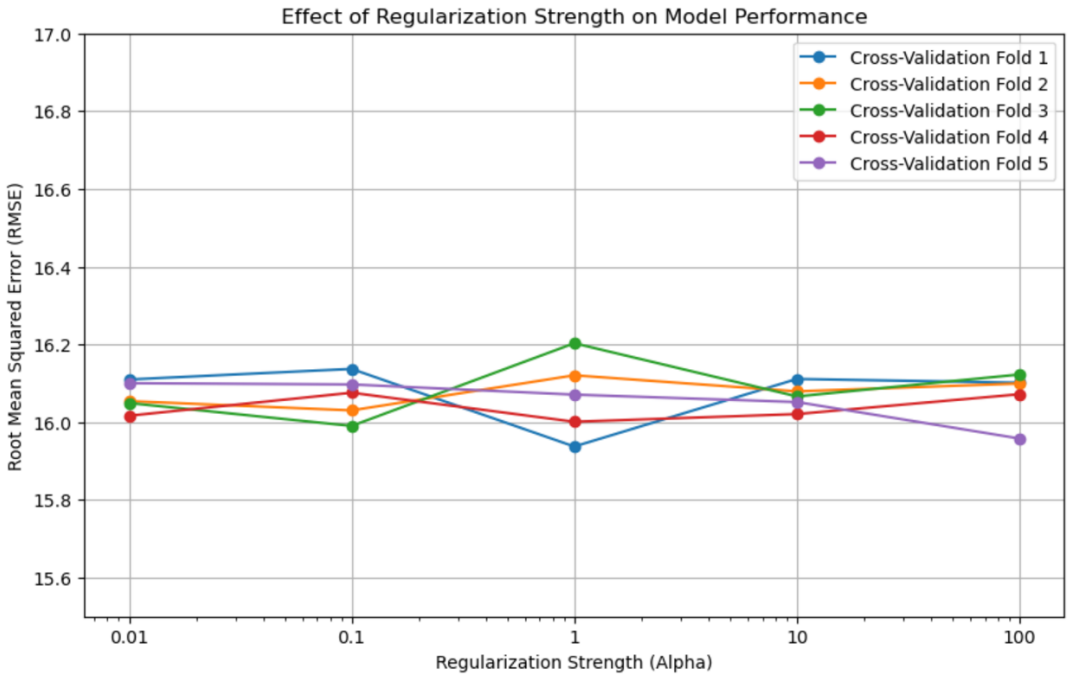
3.1.2 Employing Cross Validation for Robust Evaluation

To estimate model performance effectively, 5-fold cross-validation methodology was adopted. This technique partitions the dataset into five subsets, iteratively training the model on four subsets while validating on the fifth. By rotating the validation subset, this approach provides robust estimates of model performance.

3.2 Ridge Regression Using Numerical Features

The initial model focuses on applying Ridge regression exclusively to numerical features, including the binary 'Explicit' and 'Mode' variables.

The performance of this model is illustrated in the accompanying graph, using Root Mean Square Error (RMSE) as the evaluation metric. Given that the target variable is numeric, accuracy metrics are inappropriate. Instead, RMSE is utilized because it reflects the mean square error of predictions, with the root taken to match the target variable's units. RMSE is chosen over Mean Absolute Error (MAE) due to its higher sensitivity to larger errors, thus providing a more stringent evaluation of model performance.



Graph 1: Application of Ridge regression to numerical features only

Interpretation and Insights:

- Regularization Strength (Alpha): This parameter affects the penalty on the model coefficients.

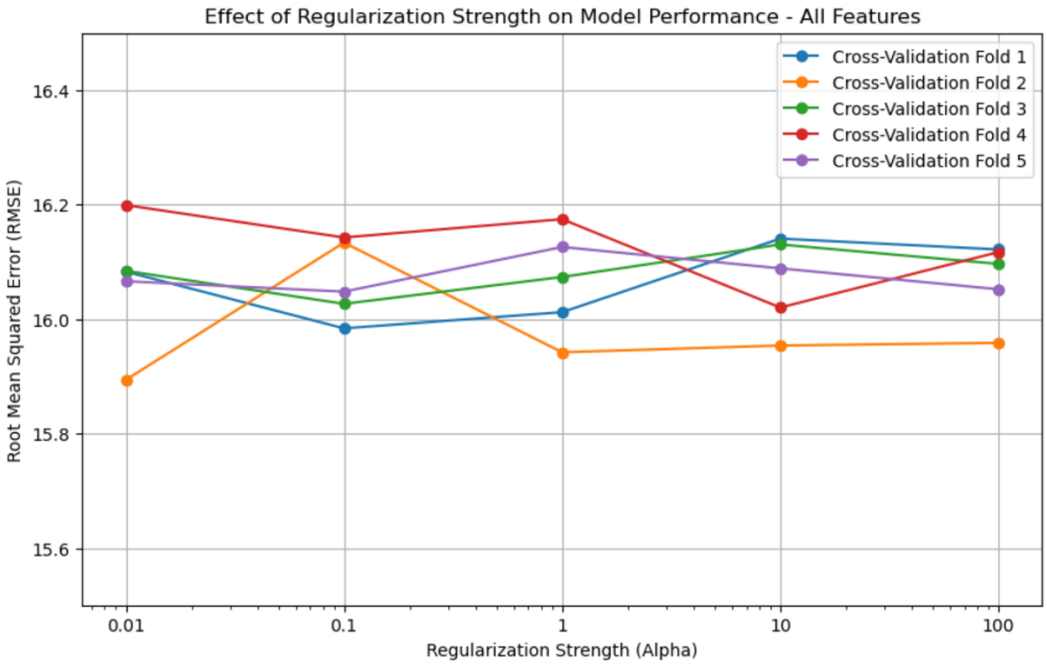
- RMSE Trends: The RMSE remains relatively stable across different alpha values, indicating that the model's performance doesn't significantly vary with regularization strength.

- Model Robustness: The stability across folds suggests that the model's predictions are consistent regardless of the subset of data used.

- Implication for Popularity Prediction: For predicting Spotify track popularity, it suggests that the chosen features are relevant and robust, leading to consistent performance across various regularization strengths.

3.3 Enhanced Ridge Regression with All Features

The subsequent model represents an advancement over the initial iteration, now incorporating all available features, including the three categorical variables: artist, album, and genre.



Graph 2: Application of Ridge regression to all features

Interpretation and Insights:

- Alpha Impact: The RMSE increases significantly at higher alpha values, indicating over-regularization.

- Optimal Alpha: Lower alpha values yield better performance by balancing the trade-off between bias and variance.

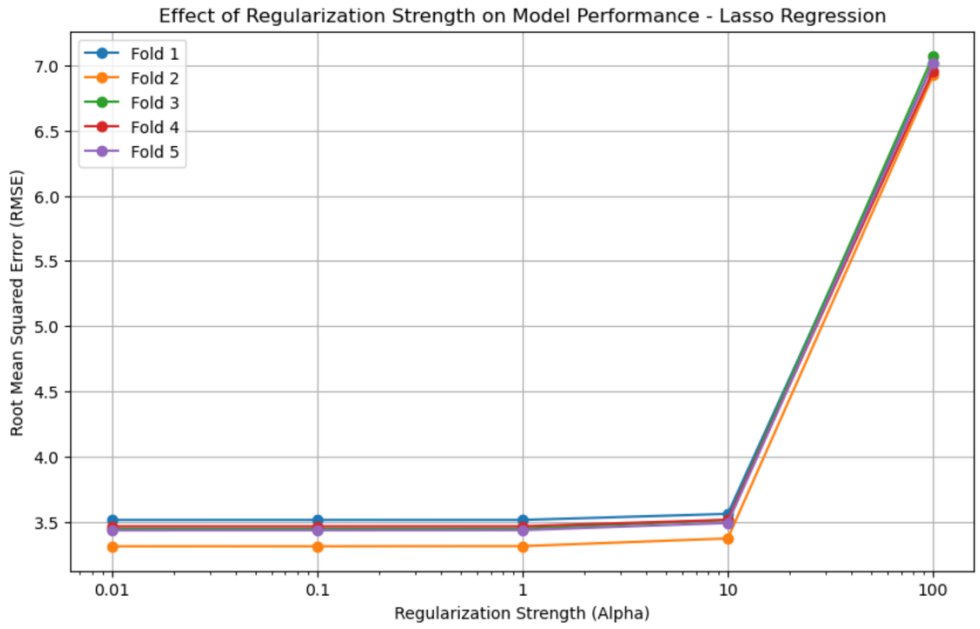
• Implication for Popularity Prediction: An optimal alpha value must be selected to avoid underfitting while ensuring that the model captures relevant patterns in the data without penalizing important features excessively.

4. Additional Models

To benchmark the performance of the model, other algorithms such as Lasso Regression and standard Linear Regression were also implemented on the same dataset. These models were trained using the full set of available features.

4.1 Lasso Regression

Lasso Regression differs from Ridge Regression in a significant way: it can shrink some coefficients to exactly zero. This makes Lasso not only a tool for regression but also a method for feature selection. By potentially eliminating irrelevant features, Lasso can effectively handle multicollinearity and enhance model interpretability.



Graph 3: Lasso Regression on all features

Interpretation and Insights:

• Lasso Regularization: RMSE values increase dramatically at higher alpha values, highlighting the importance of selecting an optimal alpha.

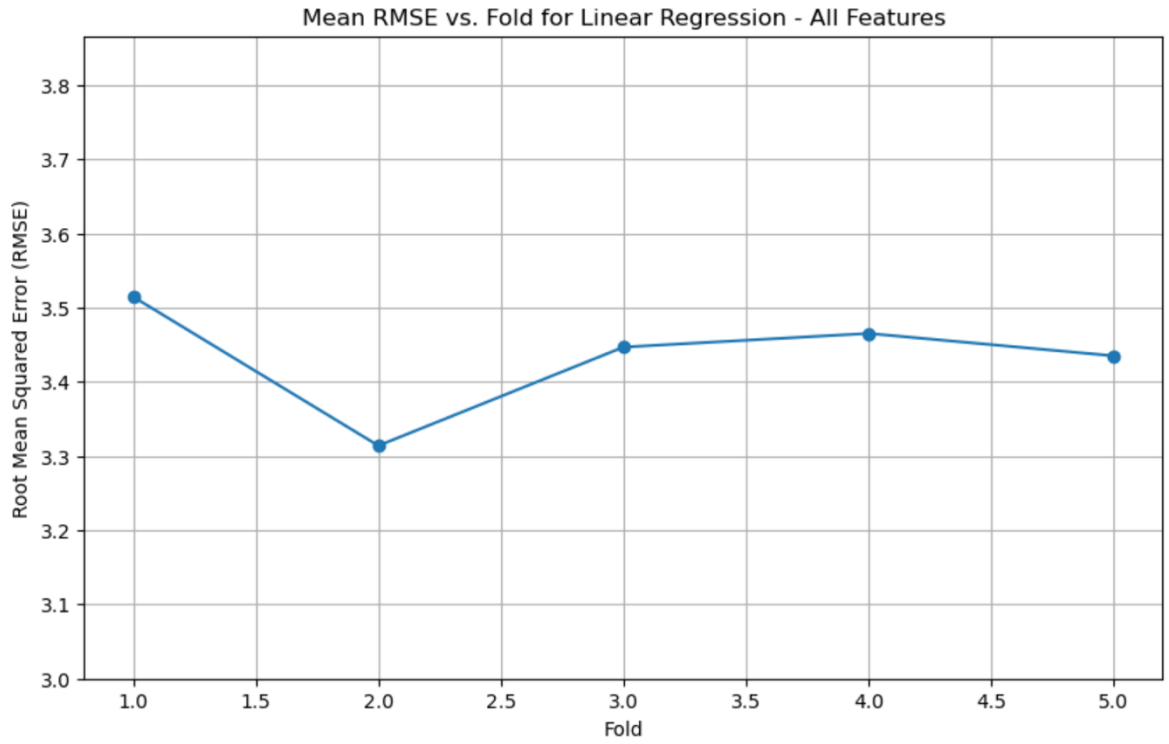
• Feature Selection: Lasso regression helps in selecting the most relevant features by applying penalties to less important ones.

• Implication for Popularity Prediction: Proper tuning of the alpha parameter in Lasso regression is crucial for balancing model complexity and prediction accuracy for Spotify track popularity.

4.2 Linear Regression

To determine if regularization is necessary, we also evaluated a standard Ordinary Least Squares (OLS) Linear Regression model using all available features. The performance results are displayed on the Graph 4.

Unlike the other models, Linear Regression does not involve a regularization parameter, so the graph simply shows the mean RMSE for each fold. The mean RMSE values range between 3.3 and 3.55, which is comparable to the performance achieved by the Ridge Regression model.



Graph 4: Linear Regression on all features

Interpretation and Insights:

• Fold Analysis: The RMSE is fairly consistent across different folds, with minor variations.

• Model Reliability: This consistency indicates that the model generalizes well to different subsets of the data.

• Implication for Popularity Prediction: The uniform performance across folds enhances the confidence that the model can reliably predict track popularity across different data splits.

4.3 Summing up insights on Ridge, Lasso and Linear Regression

Key insights:

1. Model Consistency: Both the linear regression and Lasso regression models show consistent performance across different folds, suggesting that the models generalize well and are reliable for predicting Spotify track popularity.

2. Regularization Tuning: Proper tuning of the regularization strength (alpha) is crucial, especially for Lasso regression, to avoid over-regularization and underfitting. Lower alpha values generally yield better performance.

3. Feature Relevance: The stability in RMSE across different alpha values implies that the selected features are relevant, and the model effectively utilizes them to predict track popularity.

5. Conclusion

The project aimed to predict the popularity of Spotify tracks using the Spotify Tracks Dataset. Here is a summary of the key findings:

1. ***Ridge Regression with Numerical Features Only:***

   The Ridge regression model, when applied solely to numerical features, yielded a high RMSE of 16.2. This indicates that numerical attributes alone are insufficient for accurately predicting track popularity.

2. ***Incorporation of Categorical Features:***

   Introducing categorical features such as artist, album, and genre significantly enhanced the model's predictive accuracy. Using target encoding for these categories reduced the RMSE to around 3.4, representing a five-fold improvement.

3. ***Caveats in the Results:***

   Despite the improvement, these results should be interpreted with caution. The absence of complete album data for most songs, combined with album being the most influential feature, suggests that the model may not perform as well with a more comprehensive dataset.

4. ***Influence of Hyperparameter Alpha:***

   The value of the hyperparameter alpha in Ridge regression did not substantially affect the model's performance. Testing five different alpha values showed negligible changes in the results.

5. ***Comparison with Lasso and Linear Regression:***

   Both Lasso and standard Linear Regression exhibited performance similar to Ridge regression. The choice between these models should depend on the specific objectives of the project. For instance, Lasso can also serve as a feature selection method by shrinking some coefficients to zero.