Machine Learning project:

Ridge Regression for Spotify Song Popularity prediction

Kseniya Yerameichyk
kseniya.yerameichyk@studenti.unimi.it

06/03/2024

1. Introduction

The objective of this project is to explore the effectiveness of Ridge Regression in predicting song popularity based on various features provided in the Spotify Tracks Dataset from Kaggle. To perform it, a custom Ridge Regression model in Python was developed from scratch, exploring both the analytical solution via the normal equation and the iterative approach of gradient descent. The aim of implementing both methods is to compare their performance in terms of computational efficiency and predictive accuracy.

The project involves preprocessing the dataset to ensure its suitability for Ridge Regression modeling. This includes handling missing and duplicated values, encoding categorical features, and scaling numerical attributes.

The performance of the Ridge Regression model is explored in two scenarios:

- The utilization of only the numerical features of the dataset.
- The utilization of both numerical and categorical features after encoding the categorical data.

To assess the model performance was used 5-fold cross-validation. The model was executed both with and without cross-validation to compare its predictive performance under different validation strategies. Root Mean Squared Error (RMSE) and R-squared score (R^2-score) were utilized as metrics for evaluation.

The findings of this project will provide valuable insights into the effectiveness of Ridge Regression in the context of music recommendation systems and will contribute to the ongoing exploration of machine learning techniques for personalized music discovery.

2. Implementation of Custom Ridge Regression Models

2.1 Analytical Approach

The normal equation provides an analytical solution for calculating the coefficients of a linear regression model. For Ridge Regression the normal equation is modified to incorporate the regularization parameter alpha, which aim is to penalize large coefficient values to prevent overfitting.

The normal equation for Ridge Regression can be expressed as:

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

where:

- X is the feature matrix. Each row represents a sample and each column represents a feature (including the bias term)
- y is the target vector
- w is the vector of coefficients to be estimated
- α is the regularization parameter, controlling the strength of regularization. Can take only non-negative values.
- I is the identity matrix (with zero first element to take the bias term into consideration)

In its fit method the custom Ridge Regression implementation calculates the coefficient values using the closed-form expression of the normal equation, considering both the augmented feature matrix and the regularization term. In the predict method it predicts target values using the trained coefficient values and the augmented feature matrix.

2.2 Iterative Approach

In contrast to the analytical solution provided by the normal equation, the iterative approach for implementing Ridge Regression utilizes the stochastic gradient descent (SGD) algorithm. This approach is particularly beneficial for large-scale datasets and allows for efficient optimization of the Ridge Regression model parameters.

The gradient descent method involves iteratively updating the model parameters to minimize the cost function and can be represented with the following equation, which is implemented in the model fit method:

$$w_{new} = w - \eta(2X^T(y - Xw) + 2\alpha w)$$

After the model is fitted to the training data, the predict method of the model is used to make predictions on new data samples.

3. Exploratory Data Analysis

3.1 Data description

The Spotify Tracks Dataset is a comprehensive collection of Spotify tracks samples. Each track in the dataset is associated with various audio features, providing valuable insights into the characteristics of the music.

The dataset consists of twenty columns, each providing valuable information about the tracks:

- *track_id*: The unique Spotify ID for each track.
- *artists*: Names of the artists who performed the track, separated by semicolons if there are multiple artists.
- *album_name*: The name of the album in which the track appears.
- *track_name*: The name of the track.
- *popularity*: A measure of the track's popularity, ranging from 0 to 100.
- *duration_ms*: The duration of the track in milliseconds.
- *explicit*: Indicates whether the track contains explicit lyrics.
- *danceability*: Describes how suitable a track is for dancing.
- *energy*: Represents the intensity and activity of the track.
- *key*: The key in which the track is.
- *loudness*: The overall loudness of the track in decibels.
- *mode*: Indicates the modality (major or minor) of the track.
- *speechiness*: Detects the presence of spoken words in the track.
- *acousticness*: Measures the confidence of whether the track is acoustic.
- *instrumentalness*: Predicts whether a track contains no vocals.
- *liveness*: Detects the presence of an audience in the recording.
- *valence*: Describes the musical positiveness conveyed by the track.
- *tempo*: The overall estimated tempo of the track in beats per minute.
- *time_signature*: The estimated time signature of the track.
- *track_genre*: The genre to which the track belongs.

The dataset contains 114.000 rows, each representing a distinct track from the Spotify platform. During the initial data preprocessing stage, it was observed that one of the records contained missing values for essential columns such as artists,

album_name, and track_name. Given that these fields are crucial for track identification and analysis, the record was considered corrupted and subsequently removed from the dataset to maintain data integrity.

Furthermore, it was discovered that the dataset contained 450 duplicated rows. Duplicate entries can introduce biases and inaccuracies in analytical models. Therefore, the entry was removed from the dataset to ensure accurate analysis and modeling.

The explicit field within the dataset originally contained values of "true" and "false" to denote whether a track has explicit lyrics. For standardization purposes and ease of interpretation, these values were modified to binary representations of 1 and 0, respectively, where 1 indicates explicit lyrics and 0 indicates non-explicit lyrics.

3.2 Numerical features analysis

This section analyzes the numerical features within the dataset to understand their distribution and identify potential issues. The list of the numerical features of the dataset contains the following columns: popularity, duration_ms, explicit, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and time_signature.

An interesting peculiarity was observed in the loudness field. The majority of the values were negative. In audio processing, loudness represents the perceived intensity of sound, with positive values indicating sounds louder than a reference level and negative values indicating quieter or softer sounds. This finding suggests that the dataset may be skewed towards quieter audio samples.

There are 163 rows were with a time_signature value of 0 in the dataset. In music theory, time signatures specify the number of beats per bar. A value of 0 typically indicates non-musical audio content, such as audiobooks, white noise, or speech recordings.

To gain further insights into the numerical features, histograms were created to visualize their distribution. Danceability and tempo show the distributions close to normal. The distribution of popularity showed a high concentration of entries at 0. However, excluding these zeros, the remaining data shows a distribution close to normal. Speechiness, liveness and duration_ms display right-skewed distributions. This means a larger proportion of data points lie towards the lower end of the spectrum, with a longer tail towards higher values. As mentioned earlier, the

loudness distribution was left-skewed. This confirms the observation of predominantly negative values, indicating quieter audio samples.

A correlation matrix was constructed to explore the relationships between the numerical features. The target variable popularity did not show strong correlations with any of the features. The highest positive correlation was observed between loudness and energy (0.76), indicating that louder music generally has higher energy. Conversely, the strongest negative correlation was found between acousticness and energy (-0.73), suggesting less acoustic music tends to have higher energy levels. A moderate positive correlation existed between valence and danceability (0.48), implying potentially happier music might be associated with dance music characteristics. A moderate negative correlation was observed between loudness and acousticness (-0.59), suggesting quieter music often leans towards being more acoustic. A negative correlation was also found between loudness and instrumentalness (-0.43), indicating quieter music may have a greater presence of vocals.

During the analysis were noticed some outliers in different features of the data. It is important to mention that these outliers, while different from the usual data points, fit within the unique characteristics of audio dataset. For example, in audio data, outliers might represent rare or extreme sound patterns with distinctive acoustic qualities. Although they stand out from the rest of the data, these outliers offer valuable insights into the diverse nature of audio tracks.

3.3 Categorical features analysis

The dataset contains five categorical features: track_id, artists, album_name, track_name, and track_genre. Since the track_id feature does not contribute any meaningful information to the analysis of popularity prediction, it can be safely removed from the dataset. The artists feature consists of 31.437 unique artists. "The Beatles" emerges as the most frequent artist, appearing 279 times in the dataset. For the album_name feature 46.589 unique album names were observed. The most frequently occurring album name is "Alternative Christmas 2022" appearing 195 times. The track_name feature shows even greater diversity, comprising 73.608 unique track names. The top track name "Run Rudolph Run" occurs 151 times. The track_genre feature categorizes tracks into 114 unique genres. The most frequent genre, acoustic, appears 1000 times, indicating a considerable prevalence of acoustic recordings within the data.

4. Feature Engineering

4.1 Categorical features Target Encoding

Initially, the artists field contains a list of artists separated by semicolons. The dataset was transformed by creating separate track rows for each artist in the initial list. This transformation ensures that each artist associated with a track is treated as an individual observation, allowing the influence of each artist on the target variable to be captured independently.

For categorical variables track_name, artists, and track_genre was performed target encoding using both smoothed and unsmoothed techniques. Smoothed target encoding accounts for the variability in target frequencies across different categories and helps prevent overfitting.

The optimal smoothing parameter values for the TargetEncoder were determined by employing a 5-fold cross-validated custom Ridge Regression model. This involved iteratively testing different smoothing values and evaluating their performance based on mean squared error (MSE). The smoothing value that showed the best performance was selected as the optimal choice for target encoding. For track_name, the optimal smoothing value was determined to be 175. For artists, the optimal smoothing value was determined to be 85. For track_genre, the encoding smoothing optimal choice resulted in similar MSE values for every smoothing value tested. Therefore, this field was target encoded without smoothing, while the other two were encoded with both optimal smoothing value and without it.

The encoded categorical features were removed from the dataset. The entries were then grouped by track id, and the encoded values of artists and track_genre were aggregated by averaging across the grouped entries. This aggregation ensures that each track is represented by a single set of encoded values, capturing the collective influence of artists and track genres associated with each track.

4.2 New features generation

During the feature engineering step were generated new features based on the existed categorical ones. A new feature called num_genres represents the number of genres associated with each track. This feature provides insights into the diversity of musical styles represented within each track. Another newly introduced feature is num_artists which denotes the number of artists contributing to each track. The artists_freq feature quantifies the frequency of occurrence of each artist across the dataset. This helps to recognize the impact of popular or frequently appearing artists

on the modeling results. Similarly, the album_freq feature measures the frequency of occurrence of each album name within the dataset.

5. Model Execution

The target vector y duplicates the popularity values extracted from the initial dataset. Additionally, two feature matrices X were created: one containing smoothed target-encoded values for artists and track names, and the other consisting of unsmoothed values for the same fields.

Before model execution, MinMaxScaler was applied to the feature matrices to normalize their values. The MinMaxScaler was set to scale features within the range of 0 to 1.

5.1 Numerical features only

The selected numerical features include duration_ms, explicit, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, and time_signature.

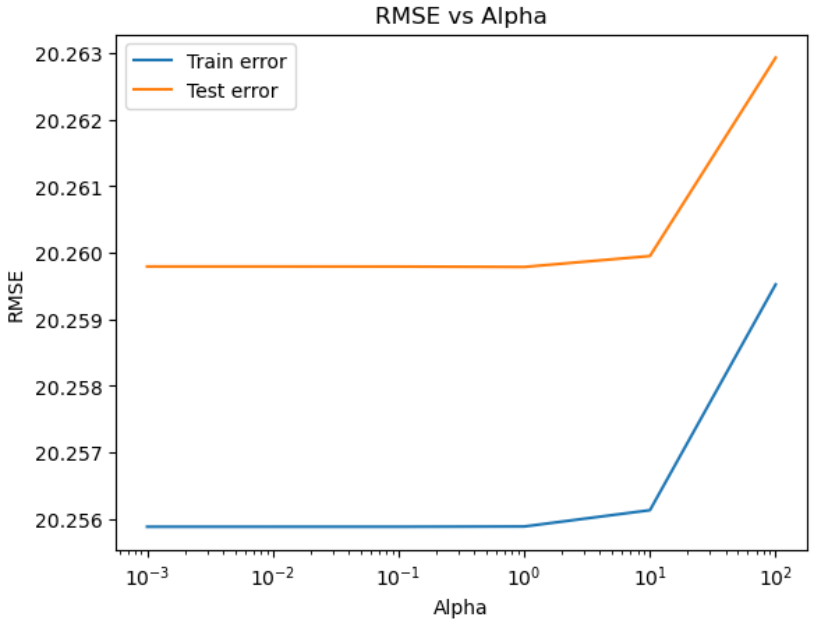
Initially, model was executed using RidgeRegNormalEquation, a custom analytical ridge regression implementation. A set of alpha values were used is the following 0.001, 0.01, 0.1, 1, 10 and 100. The model was tried both with and without cross-validation. Secondly, the model was executed using iterative ridge regression implementation RidgeRegGradientDescent. The table below summarizes the findings of the analysis:

	Best alpha	RMSE	R^2-score
Normal equation (no cv)	0.001	20.328	0.034
Normal equation (with 5-fold cv)	1.0	20.260	0.033
SGD (no cv)	0.001	20.512	0.016

Table 1. Numerical only features model execution results

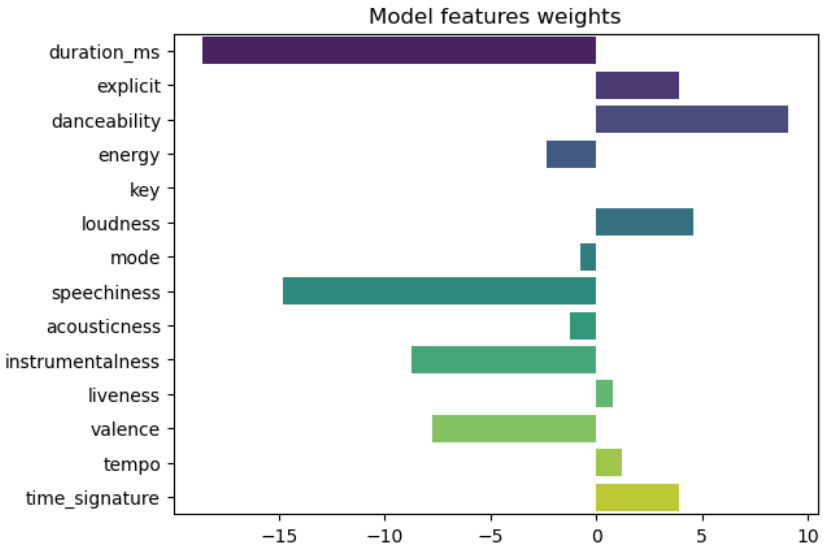
Based on the results it can be concluded that the analytical ridge regression using the normal equation shows better results in terms of RMSE and R^2-score compared to stochastic gradient descent (SGD) based approach, indicating that the iterative optimization process may not be as effective for this specific dataset. The best performing model among the tested configurations was the normal equation-based ridge regression with 5-fold cross-validation, which achieved the lowest

RMSE of 20.260 and a R^2-score of 0.033, using an alpha value of 1.0. In the plot below it is seen a small drawdown for the chosen best alpha value.



Picture 1. RMSE vs Alpha plot for numerical features model execution

Feature importance analysis reveals that the most influential feature selected by the model is duration_ms, followed by speechiness, instrumentalness, valence, and danceability. Other features make relatively smaller contributions to the model predictions. Notably, a key feature has a coefficient of zero, indicating its insignificance for the model performance.



Picture 2. The numerical features model weights

In summary, the R^2-score being close to zero suggests that the model ability to explain the variance in the target variable is quite poor. Additionally, the model demonstrated relatively high RMSE. These findings suggest that relying solely on numerical features may not be sufficient for accurately predicting the target variable, and further improvements may be necessary to enhance model performance.

5.2 Numerical and categorical features (not smoothed target encoding)

For the second model execution numerical features were combined with generated and encoded categorical features, without applying smoothing for target encoding. Were employed custom ridge regression models (the same as for the numerical features model only) with various alpha values and compared their performance with and without cross-validation. The table below shows the obtained results:

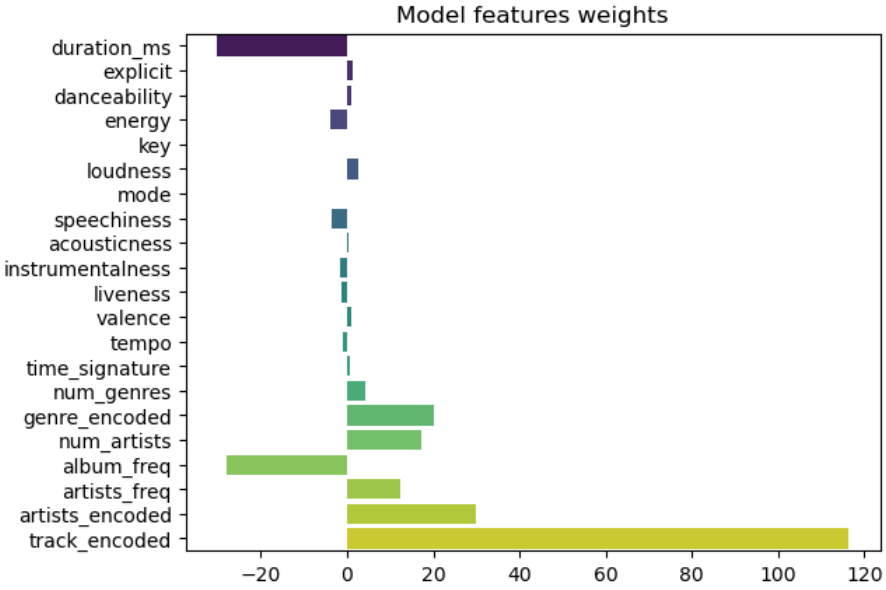
	Best alpha	RMSE	R^2-score
Normal equation (no cv)	0.001	11.553	0.688
Normal equation (with 5-fold cv)	0.1	11.577	0.684
SGD (no cv)	0.001	15.105	0.466

Table 2. Numerical and categorical (no smoothing) features model execution results

Integrating encoded categorical features significantly improved the model ability to explain the variance in the target variable, as evidenced by the increase in R^2-score from 0.034 (previous model) to 0.688 (current model). The normal equation-based ridge regression, particularly without cross-validation, showed the best results with an RMSE of 11.553. Cross-validation introduced a minimal performance change with an RMSE increase of only 0.024. According to the cross-validated model run the best value of the regularization term alpha is 0.1. Even with stochastic gradient descent (SGD), despite exhibiting higher RMSE values, the R^2-score remained relatively high, suggesting the model still captured a significant portion of the variance.

Feature weights provide insights into the relative importance of each feature in predicting the target variable. The track_encoded feature has the highest weight at approximately 118, indicating a strong positive influence on the model predictions. This suggests that the specific track plays a significant role in determining the target variable. The feature artists_encoded holds the second-highest

weight at around 30, implying a moderate positive influence. The artists identity contributes less significantly than the specific track itself. The album_freq, with a weight of -25, is the third most important feature but with a negative coefficient. This suggests that higher frequency of an album appearing in the dataset is associated with a lower popularity value of the target variable. Numerical features, except for duration_ms, have relatively low weights, indicating a minimal impact on the model predictions. In addition to the key, also the mode has a coefficient of zero, indicating its non-importance for the popularity prediction. The duration_ms feature has a weight that is still considerably smaller than the top categorical features, suggesting some, but limited, influence on the popularity value.



Picture 3. The numerical and categorical (not smoothed) features model weights

The obtained results highlight the positive impact of incorporating categorical features on the model ability to fit the data and explain the target variable. While the normal equation-based ridge regression without cross-validation achieved the highest R^2-score, further evaluation is necessary to determine the optimal model and hyperparameter configuration.

5.3 Numerical and categorical features (smoothed target encoding)

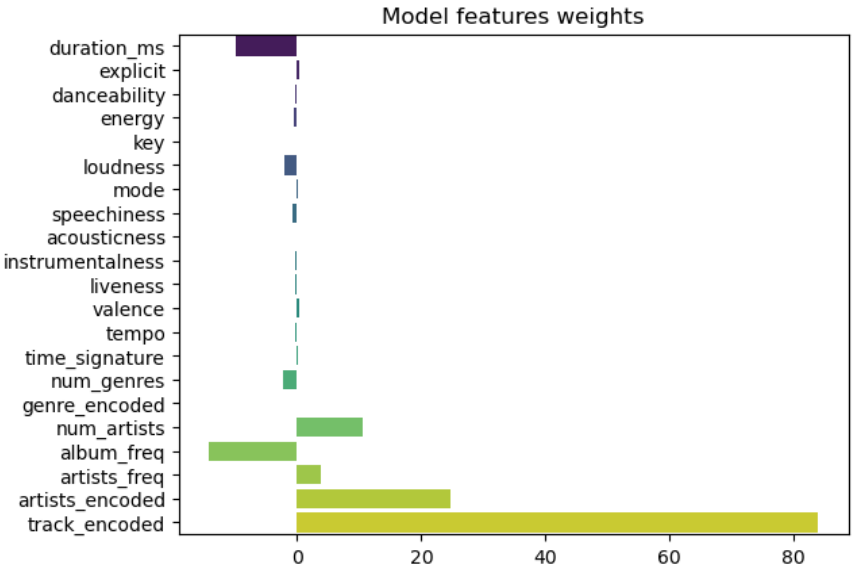
Further experimentation explored the impact of applying smoothing to the track_name and artists target encoding. This approach demonstrated a model with superior performance compared to the previous best model without smoothed encoding. The table below shows the obtained results:

	Best alpha	RMSE	R^2-score
Normal equation (no cv)	1.0	7.802	0.858
Normal equation (with 5-fold cv)	0.1	7.816	0.856
SGD (no cv)	0.001	10.452	0.745

Table 3. Numerical and categorical (with smoothing) features model execution results

The model with smoothed target encoding achieved a significantly lower RMSE of 7.802 compared to 11.553 in the previous model. This signifies a substantial reduction in prediction error. The R^2-score also increased significantly to 0.858, indicating a stronger fit of the model to the data and a better explanation of the variance in the target variable. The optimal alpha value remains consistent with the alpha value that was determined as the best in the previous cross-validation process for the model.

Like the previous model, the track_encoded feature still holds the highest weight at approximately 85. This reinforces its crucial role in predicting the track popularity. The weight for duration_ms has decreased compared to the previous model, indicating a reduced significance but remains the most impactful numerical feature. Key, acousticness, and genre_encoded features now have zero weights in this model. This suggests that their influence has been negated by other features.



Picture 4. The numerical and categorical (smoothed) features model weights

The introduction of smoothed target encoding demonstrably enhanced the model performance, evidenced by the substantial reduction in RMSE and significant increase in R^2-score. While the track_encoded feature remains the most influential, weight changes and zero weights for some features suggest a potential shift in the model emphasis on different factors for prediction.

6. Conclusion

The project investigated the effectiveness of Ridge Regression in predicting song popularity on the Spotify Tracks dataset. One of the key findings of the project is the significance of incorporating categorical features. Integrating encoded information from categorical features like artists and track_name led to a significant improvement in model performance compared to relying solely on numerical features. This demonstrates the importance of capturing these characteristics for accurate popularity prediction.

The application of smoothing to the target encoding process further enhanced the model ability to predict popularity. The model with smoothed target encoding achieved a substantially lower RMSE (7.802) and a considerably higher R^2-score (0.858) compared to the model without smoothing. This suggests that smoothing mitigates the influence of outliers and potential biases in the target variable, leading to more accurate predictions.

The analysis of feature weights revealed the critical role of the track_encoded feature in predicting song popularity, consistently holding the highest weight across all models. This implies that the specific track itself significantly influences its popularity. Furthermore, other features like artists_encoded and duration_ms also demonstrated positive influences on popularity prediction, while some features, such as acousticness and genre_encoded exhibited varying importance depending on the specific model configuration.

In conclusion, this project demonstrates the potential of Ridge Regression, combined with effective feature engineering and target encoding techniques, for predicting song popularity. While this study showed promising results, further exploration incorporating other machine learning models, advanced feature engineering strategies, and hyperparameter tuning could potentially lead to even more accurate and robust music popularity prediction models.

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my/our work. I/We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.