UNIVERSITÀ
DEGLI STUDI
DI MILANO

Statistical Methods for Machine Learning

Project (Ridge Regression)

Professor Nicolò Cesa-Bianchi

Prepared by:

Chan Zhong Ping, Jeffrey (V10153)

1. Setting up Kaggle API and Dataset

The Kaggle dataset “Spotify Tracks Dataset” was last accessed and downloaded on 8 June 2023.

```
import os
os.environ['KAGGLE_USERNAME'] = "XXXXX"
os.environ['KAGGLE_KEY'] = "XXXXX"
!mkdir spotifydata
!kaggle datasets download -d maharshipandya/-spotify-tracks-dataset -p
spotifydata
!unzip spotifydata/-spotify-tracks-dataset.zip -d spotifydata
```

Upon setting up the API key, the file is downloaded and a new directory “spotifydata” was created to unzip all the files into a single directory. The directory can be accessed through `spotifydata/dataset.csv`.

2. Data Pre-processing

2.1 Loading of Data into Dataframe

The downloaded dataset is loaded onto a pandas dataframe, with the features separated into numerical and categorical features. The numerical features are: ‘duration_ms', ‘danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo', and 'time_signature’, while the categorical features are: ‘key', 'explicit', 'mode', ‘time_signature’ and 'track_genre'. Although the variables ‘key’, ‘time_signature’ and ‘mode’ take on numerical values, they are classified as categorical features as these features have categorical characteristics. This dataset consists of 114k different rows, with some missing values. As the missing values are very minimal, the rows with empty values can be dropped. Additionally, the variables, namely ‘track_id’, ‘artists’, ‘album_name’, ‘track_name’ and number column, which do not affect the popularity of the song are dropped.

2.2 Encoding Categorical Variables

Once the missing data and irrelevant columns are dropped, the categorical features have to be encoded using an appropriate method. This experiment uses a variation of one-hot encoding in order to turn the categorical variables into numerical values. The standard one-hot encoding turns every category into a new column with binary values, however this has proven to be inappropriate for this particular dataset because of the large number of values in track_genre. By performing one-hot encoding, a total of (n*m) new values have to be created, where n is the number of rows and m is the number of columns.

The variation used in this experiment consolidates the entire column of genres and groups them into two main categories – the songs are either rap and hip-hop or others (Nair, 2023). This reduces the amount of memory usage significantly, as m is now reduced to 2, which allows for better memory management. The remaining categorical variables are then encoded as well.

2.3 Combining Numerical and Categorical Variables

Once encoded, the numerical and categorical variables are converted from a dataframe to a two-dimensional array. The two types of variables are then concatenated to form a combined array of independent variables which will be used for model building.

3. Ridge Regression Function

The defined Ridge Regression function takes in three parameters: X, y and alpha and returns the coefficient values of the trained model. The Ridge Regression formula is given as

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

Where:

- $\widehat{\beta}$ is the vector of coefficient estimates of the model
- X is the matrix of the input features
- X^T is the transposed matrix of X
- α is the regularization parameter, where a larger regularization parameter implying a stronger penalty on the cost function which results in a more constrained model
- I is the identity matrix which is used together with the regularization parameter to scale the penalty, reducing impacts of large coefficients and chances of overfitting
- y is the vector of output values, or the dependent variable

After defining the function, the dataset is then split into two – training set and validation/test set. The train ratio is set to 0.8, where 80% of the data is used for training the model whereas 20% of the data is used for testing. By using the random module, random indices are generated and shuffled before applying the split onto the dataset. This ensures the samples chosen are completely random and non-overlapping; this also means that the results will be different each time this cell is executed. Random sampling is one of the sampling methods for regression models, and is chosen to prevent bias in the modelling process. To get consistent results, the line `random.seed()` can be inserted to use a set seed at the start of the notebook. In this experiment, the seed was set to 1203 to ensure a consistent result.

4. Model Training

4.1 Numerical Features only

After splitting the data into test and training sets, the alpha value must be determined. As alpha is a regularization parameter, there is no fixed value; the optimal value varies depending on the dataset and model. The Ridge Regression seeks to minimize the empirical loss through the introduction of alpha, which is determined in this model through a range of values between 1 to 10 (excluding), with a step count of 1. This generates a total of 9 alpha values which can be placed in the regression model.

The goal of performing multiple regressions using different alpha values is to find the optimal trade-off between fitting the data well and reducing the complexity of the model. The alpha value allows for control over the bias-variance tradeoff, as well as improving the model's ability to generalize unseen data. One measurement of risk estimate is Mean Squared Error, or MSE, which measures the average squared difference between the predicted and actual values. The formula is as given:

$$MSE \; = \; \frac{1}{n}\Sigma(y_i - \widehat{y_i})^2$$

The mean MSE calculated in this experiment was 495.09, and generally the value does not deviate much across the different alpha values. This shows that the regularization parameter alpha does not seem to play a very important role in the regression consisting of only numerical features. This is further supported by the fact that MSE is the lowest when alpha is 0, which simplifies the model into a linear regression:

$$\widehat{\beta} = \; (X^T X)^{-1} X^T y$$

$$\widehat{\beta} = \; Ay$$

Where A is a matrix of size n x m.

4.2 Numerical and Categorical Features

In this section, the numerical features are combined with the encoded categorical features from the dataset. The data was concatenated in cell 2.3 of the notebook. The process of using multiple alpha values and calculating the MSE is repeated in this process. In this experiment, the average MSE decreased to 482.44, and similar to the experiment for numerical features only, the value does not deviate significantly. Once again, in this experiment it appears that the value of alpha does not play such a significant role. However, the slight decrease in MSE shows that combining numerical and categorical features makes for a better model than just having numerical features. This is intuitive, as one can predict that minimally, the genre of music would affect the popularity of a song.

5. 5-fold Cross-Validation

More generally known as K-fold cross-validation, this method starts by partitioning the dataset into K different subsets (also known as folds). In this experiment, the dataset of n rows is partitioned into 5 subsets

$$D_1 = \{(x_1, y_1), \ldots, \{(x_{\frac{n}{5}}, y_{\frac{n}{5}})\}$$

$$D_2 = \{(x_{\frac{n}{5}+1}, y_{\frac{n}{5}+1}), \ldots, \{(x_{\frac{2n}{5}}, y_{\frac{2n}{5}})\}$$

$$D_3 = \{(x_{\frac{2n}{5}+1}, y_{\frac{2n}{5}+1}), \ldots, \{(x_{\frac{3n}{5}}, y_{\frac{3n}{5}})\}$$

$$D_4 = \{(x_{\frac{3n}{5}+1}, y_{\frac{3n}{5}+1}), \ldots, \{(x_{\frac{4n}{5}}, y_{\frac{4n}{5}})\}$$

$$D_5 = \{(x_{\frac{4n}{5}+1}, y_{\frac{4n}{5}+1}), \ldots, \{(x_n, y_n)\}$$

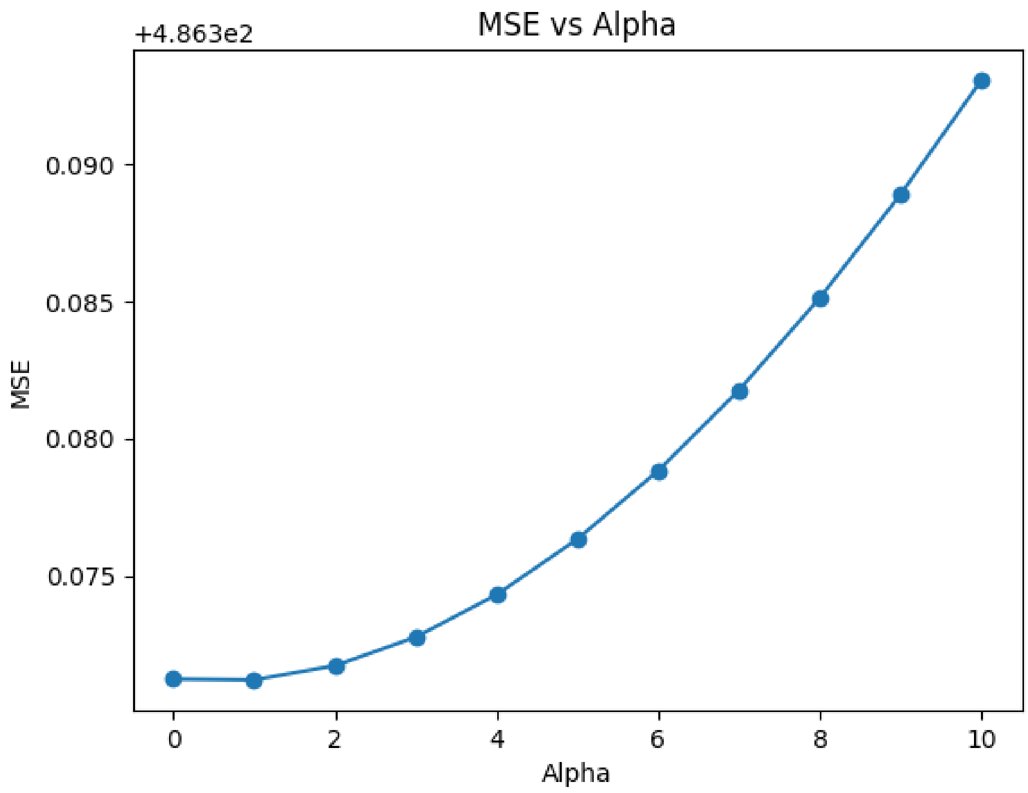
The first set D_1 is treated as the test set, while the remaining k-1 sets are used as the training set. In each iteration, the test set changes to each of the remaining folds, ensuring that each of the 5 subsets is used as the test set exactly once. Ridge regression is then performed on the training set using the function defined previously.

In the extreme case of K = n, this is called Leave-One-Out-Cross-Validation (LOOCV). As the name suggests, the dataset is split into the number of observations in the dataset, less one value. This is used when the dataset provided is small due to the high computational power required. The model generally has lower bias at the expense of high variance, but is more accurate with lower test error. In this experiment, the normal K-fold cross-validation is used as the Spotify dataset is sufficiently large.

For this experiment, the dataset is split into 5, meaning the regression is run 5 times in total. The code iterates through a set of alpha values, ultimately choosing the best alpha value among the list through finding the minimum average MSE.

$$l_S^{CV}(\alpha) = \frac{1}{K}\sum MSE$$

Due to the random shuffling of the test and training indices, the results and the respective MSE will be different for each shuffle, as different subsets of the training data is used to train the model for each fold. Based on this experiment, the best alpha is 1. A graph is plotted for the average MSE against alpha values for better visualisation as shown.



However, even upon performing 5-fold cross-validation, it appears the variation between the loss values is quite small. The variation between the errors is smaller than the first decimal point, which does not give much usefulness in comparing between the alpha values. Despite the small variation, the best alpha value is still assumed to be 1.

6. Final Model

6.1 Coefficient Determination

Assuming alpha taking a value of 4 does indeed produce a significantly smaller error. The value is put in the regression function once again to determine the respective coefficients from the optimal model. From the output, there are a few key takeaways from the model:

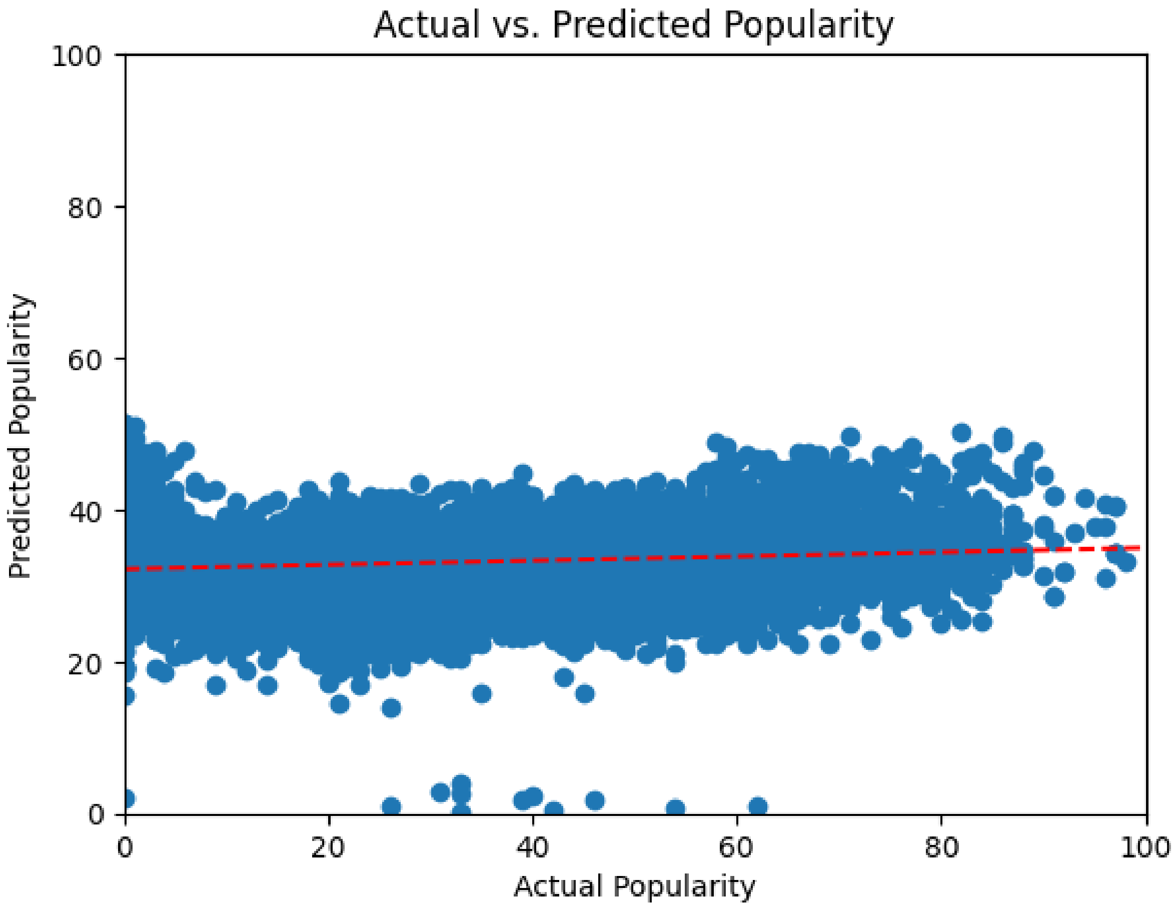
1. Duration of the song has a very small negative value. This value is small enough that it can be said that the duration of the song has insignificant effects on the popularity of the song. However, it is possible that as the duration of the song increases significantly, it can have significant effects on the popularity of the song.
2. Danceability and is_pop_or_hiphop had the largest positive coefficients. This means

that the higher the danceability and the song being either pop or hip-hop significantly increases the popularity of the song. One possible reason could be because pop and hip-hop music are commonly used for dance choreographies

3. Speechiness and instrumentalness and valence have the largest negative coefficients. This means that people generally preferred calmer music, music with less speech and less instruments.
4. Out of the four different types of time_signatures, the one with highest coefficient was 4. This means that songs with time signatures of 4 beats per measure were likely to be more popular. This would make sense, as 4 is the most commonly used time signature, also known as the default time signature (Apple, 2022).

6.2 Measuring Model Effectiveness

The most straightforward way to measure the accuracy and effectiveness of the trained model is to compare the predicted and actual outcome of the model. In this section, the graph of predicted popularity was plotted against actual popularity.



Apart from the few outliers, it can be seen that the graph for the predicted popularity and actual popularity does not follow a diagonal line with gradient 1. There appears to be an underprediction of values for the popular songs and over-prediction of values for the less popular songs; the gradient for the graph is a very small value close to 0, resulting in an almost horizontal line. This shows that the ridge regression model may not be very accurate in predicting the popularity of songs.

Model Criticism

Throughout the process of training the model, there are some aspects which may result in lowered accuracy of the final model.

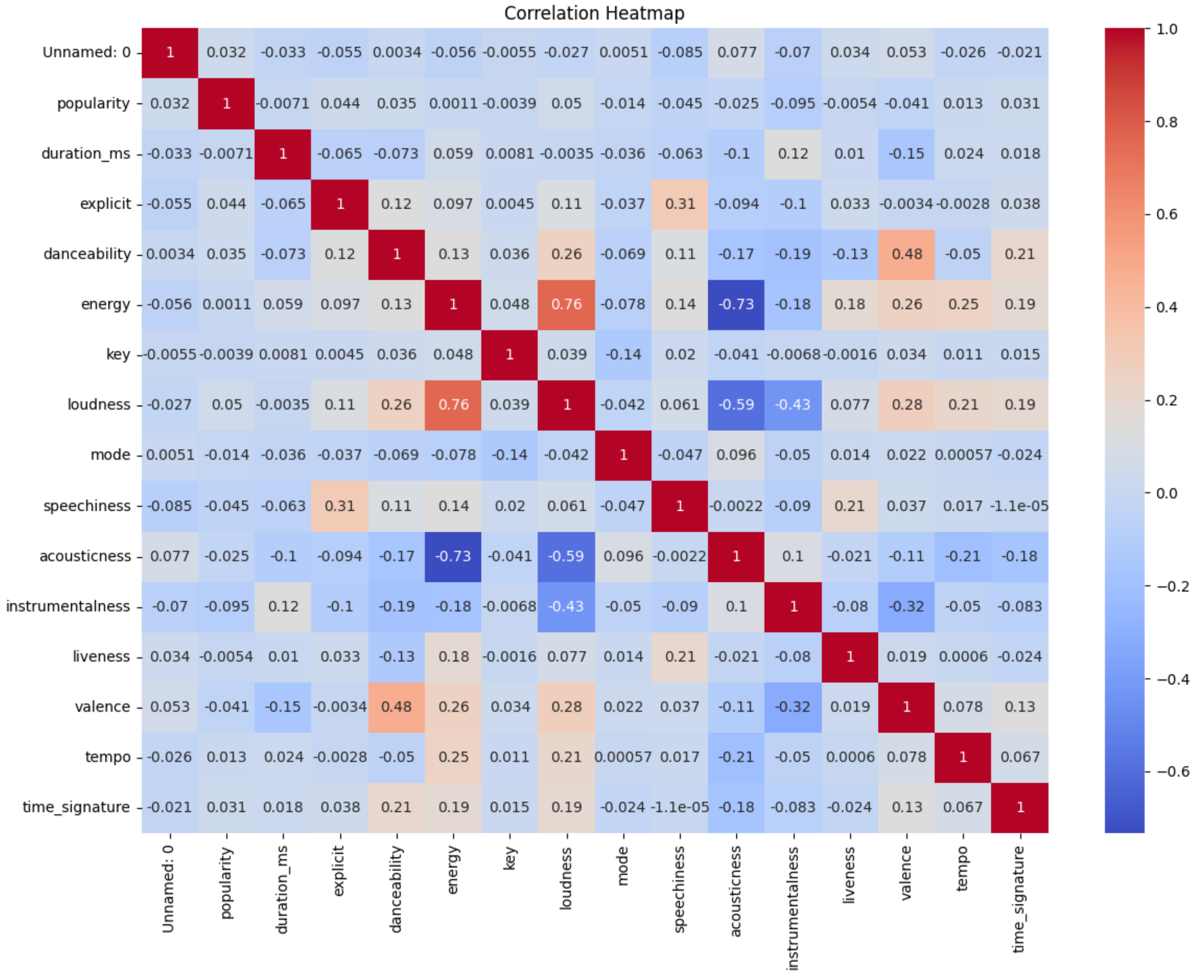
- Improper encoding: The encoding method used in this experiment is a variation of one-hot encoding. Due to the large number of categories within a single variable for the music genre, it is computationally inefficient to perform one-hot encoding and create a new column for every single genre.

  In this experiment, the number of genres was shrinked from 114 (1000 values per genre) to 2, with 2000 values for one genre and 112,000 values for the other. This severely limits the information that is usable to train the model, as it is very likely that there are some genres that contribute significantly to the popularity of songs which are underrepresented in the model.

  An alternative encoding method is to use label encoding. Label encoding only creates a single new column and assigns a numerical label to each category. This is not computationally exhaustive, but may inaccurately represent certain categories by assigning disproportionate weights. Annex A shows the difference between label encoding and one-hot encoding.
- Duplicated tracks: According to the dataset description, there are 89,741 unique track IDs, compared to 114,000 values in the dataset. This means that around 20% of the data are duplicates. However, the values do not match the number of unique track names, thus it is uncertain as to whether these duplicated tracks provide unique information that is insightful towards contributing to the model
- Normalizing variables: The numerical variables in this dataset are of different scales. This may result in extra weight being placed on variables with a large scale. For example, the duration_ms variable takes values 100,000 and above, compared to speechiness which takes values between 0 to 1. However, normalization changes the distribution and range of data, and makes it difficult to compute the final model as these same pre-processing techniques have to be applied to the inputs in the final model.

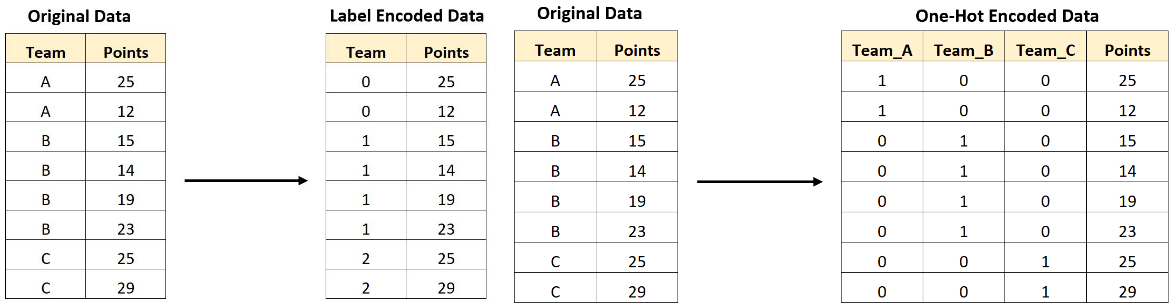
6.3 Checking Correlation of Variables

Ridge regression primarily works best when there is a certain degree of multicollinearity between the variables. The current model underperforms when predicting the popularity of songs. This could be attributed to having variables with minimal correlation. The correlation heatmap is therefore plotted to check on the correlation of the variables.



According to the correlation heatmap, the only variables that have a substantial correlation (>0.5) are between loudness, acousticness and energy. This shows that there is not much multicollinearity between the independent variables, therefore indicating that ridge regression may not be the most suitable model for this dataset. Some alternative regression techniques include XGB Regressor, Random Forest or Lasso Regression.

Annex A: Label encoding vs One-Hot encoding



Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Reference: (Zach, 2022) https://www.statology.org/label-encoding-vs-one-hot-encoding/

References

Apple. (2022, June 14). *Understanding the 4/4 time signature*. Soundbrenner. https://www.soundbrenner.com/blog/understanding-the-4-4-time-signature/#:~:text=It’s%20the%20most%20frequently%20used,%E2%80%9Ccommon%20time%E2%80%9D%20as%20well.

MaharshiPandya. (2022, October 22). *Spotify tracks dataset*. Kaggle. https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

Nair, A. (2023, February 1). *Identifying drivers of Spotify song popularity with Causal ML*. Medium. https://towardsdatascience.com/identifying-drivers-of-spotify-song-popularity-with-causal-ml-934e8347d2aa

Zach. (2022, August 8). *Label encoding vs. One hot encoding: What’s the difference?* Statology. https://www.statology.org/label-encoding-vs-one-hot-encoding/

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.