UNIVERSITÀ DEGLI STUDI DI MILANO

Machine learning Project

Fatemeh Jandari

May 2024

University of Milan

Data Science and Economics

Contents

Table of Figures

Introduction

In this project, a custom ridge regression model is developed within a Python programming environment, excluding the use of pre-existing machine learning libraries. The model is trained over four distinct approaches to data preparation and feature encoding. This report evaluates and compares the performance of the model across these different datasets, highlighting the effectiveness of each method in predicting the popularity of tracks from the Spotify Tracks Dataset. This approach not only adheres to the assignment's requirements but also provides a nuanced understanding of the impact of data preprocessing and model tuning on machine learning outcomes.

The Python code (Jupiter notebook) for this project and the report can be accessed online on my GitHub repository.

Data

The Spotify Tracks Dataset was downloaded from Kaggle as guided in the project assignment note. It consists of 20 columns as described below:

```
#   Column            Non-Null Count   Dtype
--  ------            --------------   -----
0   track_id          113999 non-null  object
1   artists           113999 non-null  object
2   album_name        113999 non-null  object
3   track_name        113999 non-null  object
4   popularity        113999 non-null  int64
5   duration_ms       113999 non-null  int64
6   explicit          113999 non-null  bool
7   danceability      113999 non-null  float64
8   energy            113999 non-null  float64
9   key               113999 non-null  int64
10  loudness          113999 non-null  float64
11  mode              113999 non-null  int64
12  speechiness       113999 non-null  float64
13  acousticness      113999 non-null  float64
14  instrumentalness  113999 non-null  float64
15  liveness          113999 non-null  float64
16  valence           113999 non-null  float64
17  tempo             113999 non-null  float64
18  time_signature    113999 non-null  int64
19  track_genre       113999 non-null  object
dtypes: bool(1), float64(9), int64(5), object(5)
```

Explanatory Data Analysis and Preprocessing

To prepare the data, first it is necessary to drop the records with missing values. Since there are very few records with such values, removing them wouldn't affect the quality of the dataset. As the table below illustrates, there is only one record with missing values in three of its columns.

```
Unnamed: 0          0
track_id            0
artists             1
album_name          1
track_name          1
popularity          0
duration_ms         0
explicit            0
danceability        0
energy              0
key                 0
loudness            0
mode                0
speechiness         0
acousticness        0
instrumentalness    0
liveness            0
valence             0
tempo               0
time_signature      0
track_genre         0
```

Next, we proceed with removing the '*unnamed: 0*' column which is just an indexing reference and has no use for our modelling purposes.

Another matter to be handled is checking for duplicated items in the '*track_id'* column which should only contain unique values. In our case, out of 113999 records, there exists 16641 duplicates. By removing the duplicates and only keeping one record for each instinct track_id, we have 89740 items.

To better understand how different features are related in the dataset, a correlation matrix is generated as shown in .

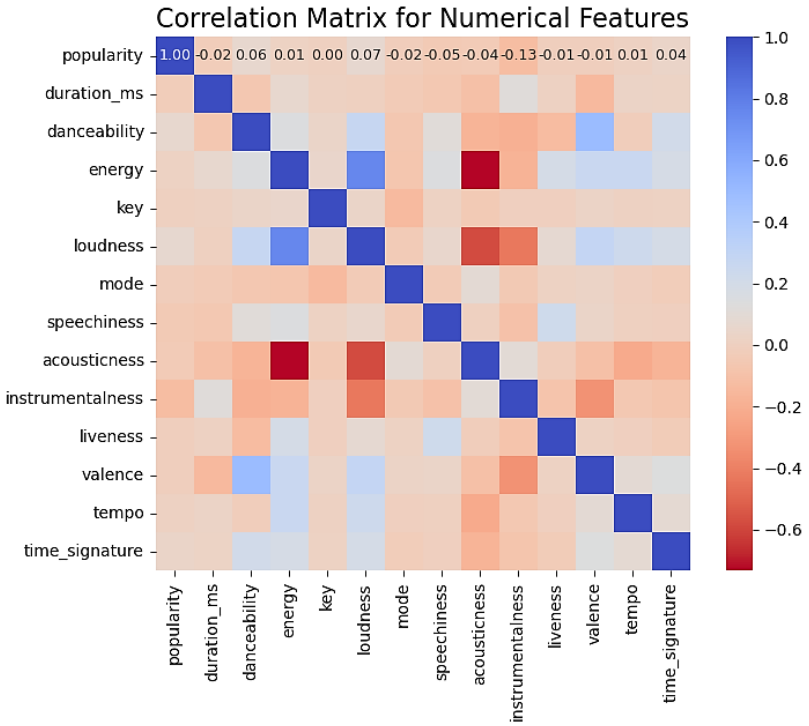

Figure 1 - Correlation matrix

Methodology

The ridge regression model is defined as a class with different functions to be used in the modelling phase on different versions of the dataset.

Ridge Regression Model

The ridge regression algorithm is a linear model. In addition to the classic linear method, it has an alpha hyperparameter that reduces the variance by changing the coefficient estimates. In this project, a class is defined as `RidgeRegression` with the following functions:

Function title	Description
fit()	weights are estimated with the closed-form regression formula
predict()	used to make predictions based on the estimated weight
meanSqr_error()	Estimates the mean square error
R2_performance()	Estimates R^2 to assess the correlation between prediction and reality
CV_Kfold()	Sklearn Kfold method to find best alpha

For tuning purposes, after testing different scales, alpha values between 10^{-2} and 10^{6} where selected for this dataset.

Dataset Partitioning and Encoding

As stated in the project assignment, a subset of the dataset with only the numerical features is prepared along with three versions of the full dataset with different encoding methods for the categorical features. The model is trained, tunned, and compared on these four versions. The three encoding techniques are described as follows:

Target Encoding: This technique replaces each category with the mean of the target variable for that category. It helps capture the relationship between the categorical feature and the target variable, reducing the dimensionality of the data.

Frequency Encoding: This method replaces each category with the frequency (or count) of that category in the dataset. It transforms categorical data into numerical values based on the occurrence of each category, helping algorithms process the data more effectively.

Leave One Out: Like target encoding, this technique replaces each category with the mean of the target variable for that category but excludes the current row when calculating the mean. This reduces the risk of data leakage and overfitting by ensuring that the target value of the row being encoded does not influence the encoding.

It is worth noting that all datasets are split using the sklearn library with the same random state (42) to avoid bios in dataset partitioning.

Results and discussion

In this section, the performance of the ridge regression model over each of the four datasets is presented. At the end, a comparison of all is showcased.

Only Numerical Features Subset

Firstly, the model is trained on the numerical dataset. The results are as follows:

	Training	Test
MSE	408.572	414.725
R^2	0.0321	0.0287

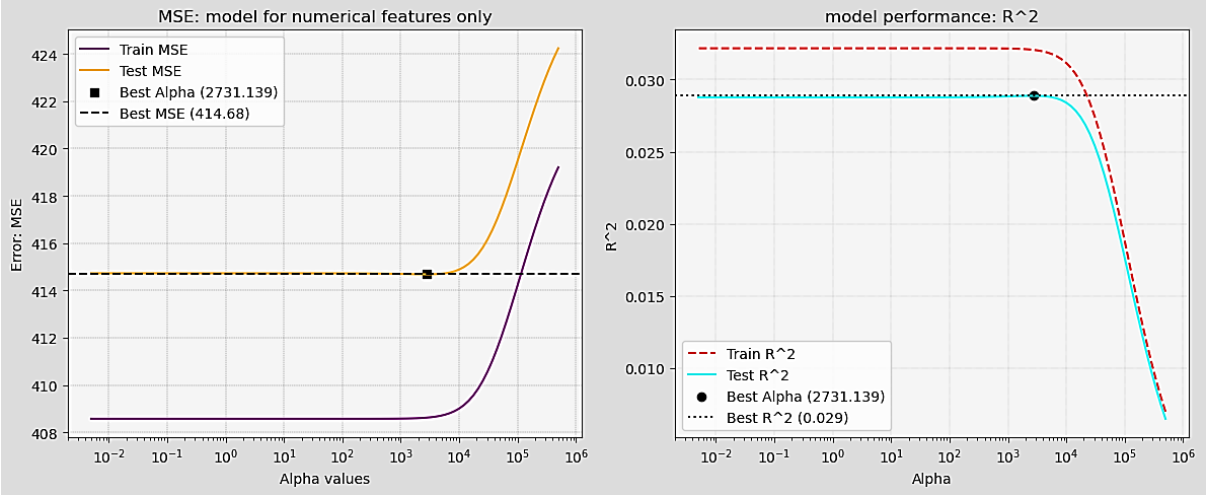

Figure 2 - Performance plots for Numerical features-only subset

Since the error difference between training and test is less than 2%, we assume that there is no overfitting. While the R2 values are considerably low, the performance of the model over numerical features only covers 2.9% of the target variance. However, the 5-fold cross validation estimates an alpha value of 243.13 as the best performing regression over this subset.

Target Encoded Dataset

Our model performed best over this encoding technique. The results can be seen in the following table:

	Training	Test
MSE	82.276	84.454
R^2	0.8051	0.8022

As the error difference on training and test subsets is 2.3%, there is no overfitting in this case as well. As shown in figure x, the initial R^2 value is 80.2% which is a reasonable number. Best alpha value after applying cross validation, suggests 2.32079.

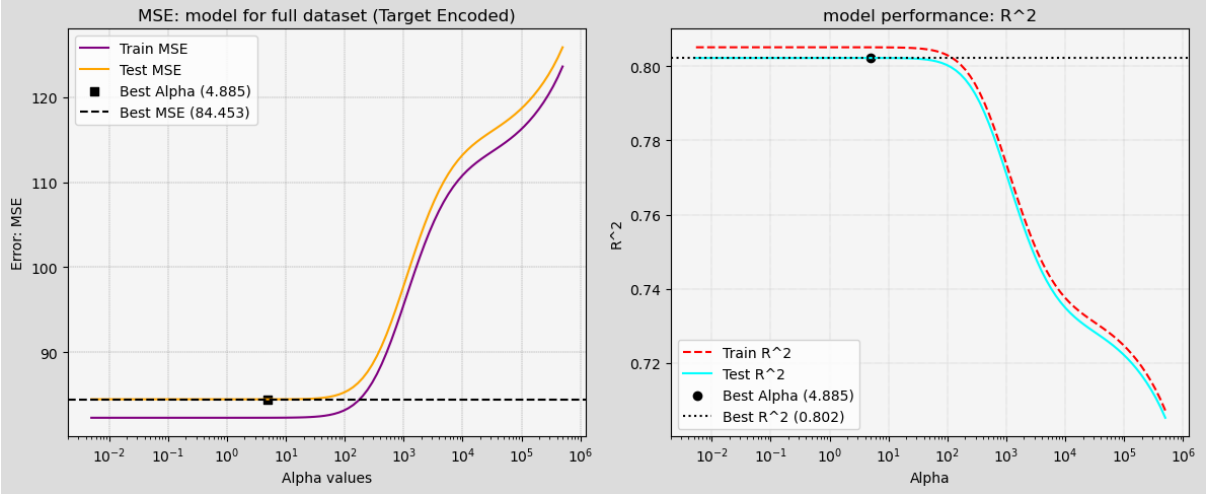

Figure 3 - Target Encoded dataset Performance plot

Frequency Encoded Dataset

Performance measures on test and training subsets are shown in the following table:

	Training	Test
MSE	228.302	233.161
R^2	0.4591	0.4539

The model performance has dropped by 50% on this dataset (best R^2 is 45.4%) compared to the Target encoded method. Still no overfitting as the error difference is 2.1%.

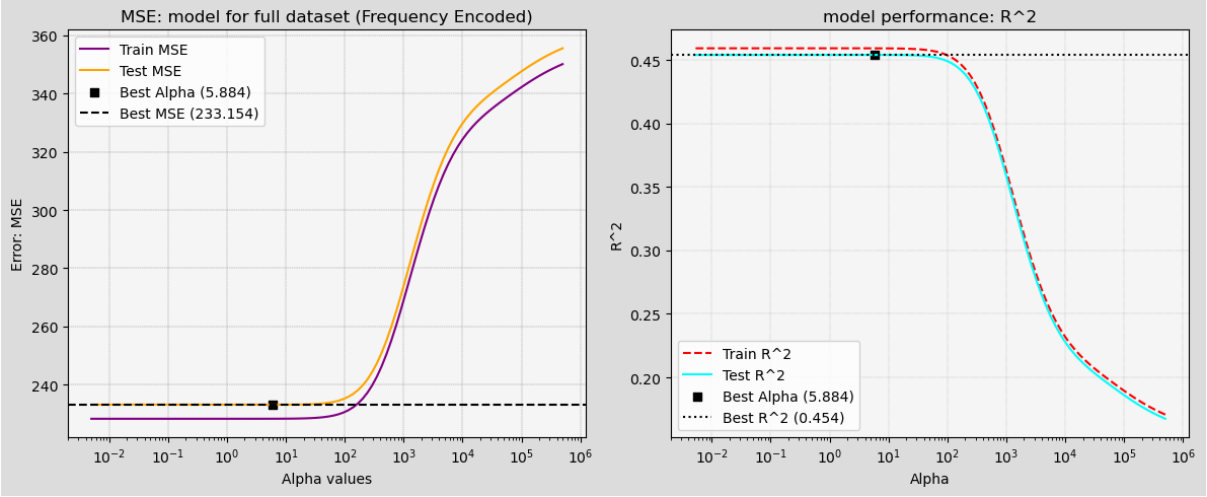

Figure 4 - Performance plot for Frequency Encoded dataset

With 5-fold cross validation, best found alpha value is 2.32079 and the corresponding average MSE is 229.1973.

Leave-One-Out Dataset

This encoding on the dataset performed much better than Frequency and slightly lower than Target encoding with a 69.75% R^2. The performance table is shown below:

	Training	Test
MSE	124.447	129.145
R^2	0.7052	0.6975

Same as the three previous cases, error difference between training and test sets is 3.8% and it suggests that no overfitting is occurring. The 5-fold cross validation resulted in 7.087371 for the best alpha.

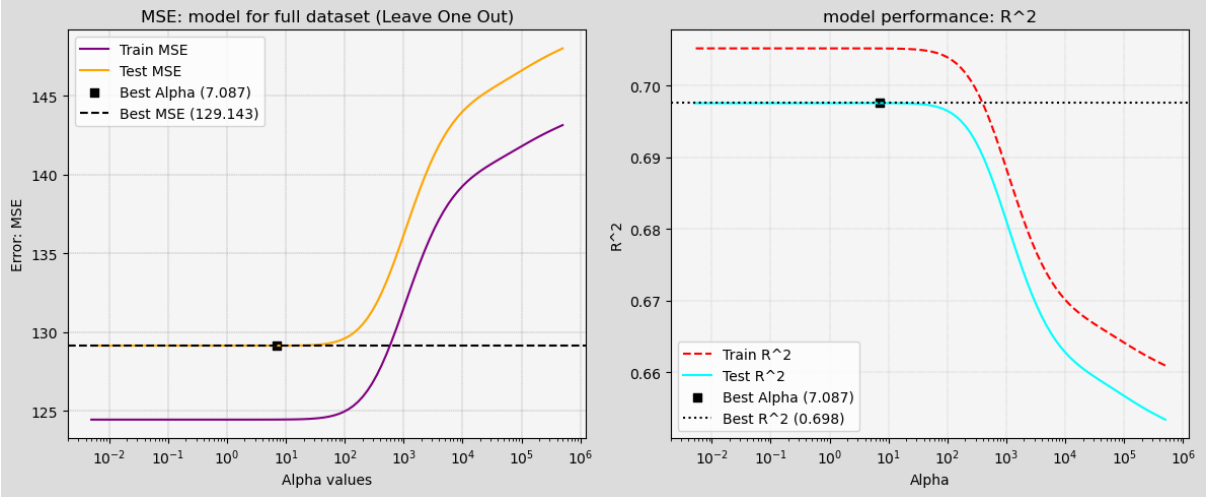

Figure 5 - Performance plot for leave one out encoding dataset

Comparison of the results

The following figure illustrates the training and test error (MSE) for each of the four datasets.

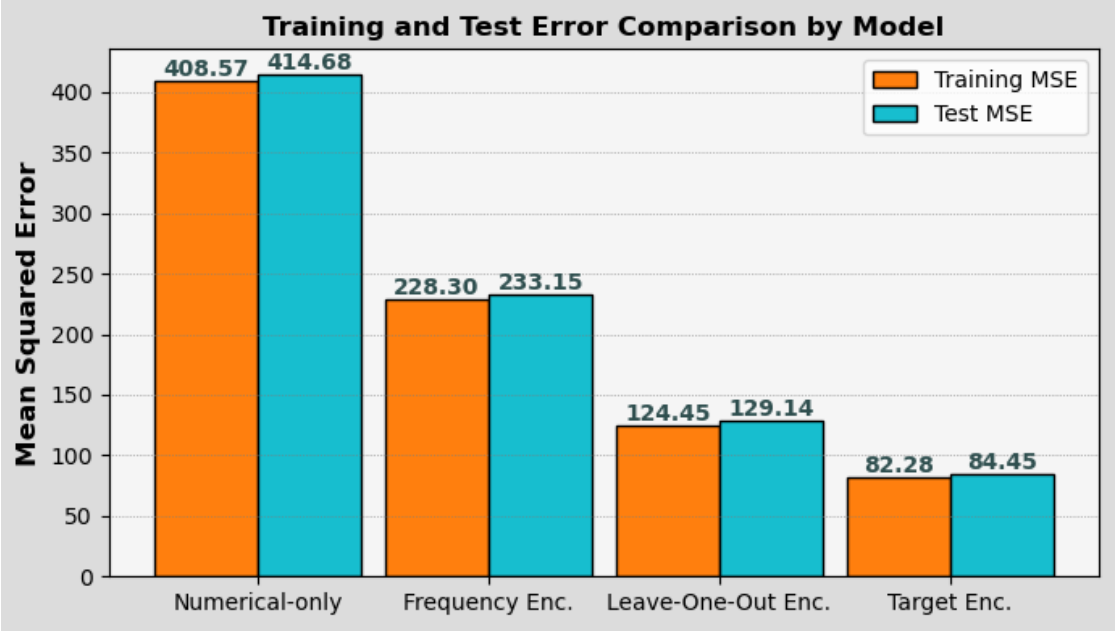

Figure 6 - MSE comparison

As it can be seen in the plot, and as mentioned previously, all models suggest no overfitting.

In the figure below, the best tunned models for the four versions of the dataset are shown. It is evident that the dataset with only numerical features has the lowest performance (R^2) and the Target encoded version is the best performing one.

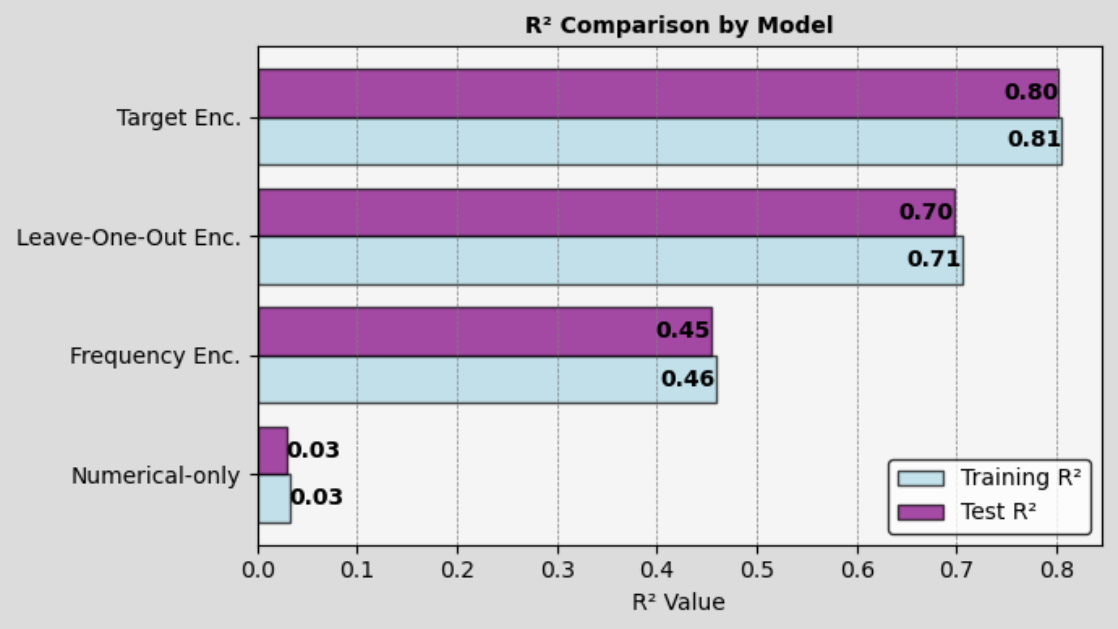

Figure 7 - Performance Comparison

While the leave one out method was expected to have a higher performance compared to the target encoded dataset, it turned out that controlling for data leakage in this method was in fact misinforming for the ridge regression model.

We proceeded with the final figure which showcases the lowest error and best alpha value for all the different approaches.

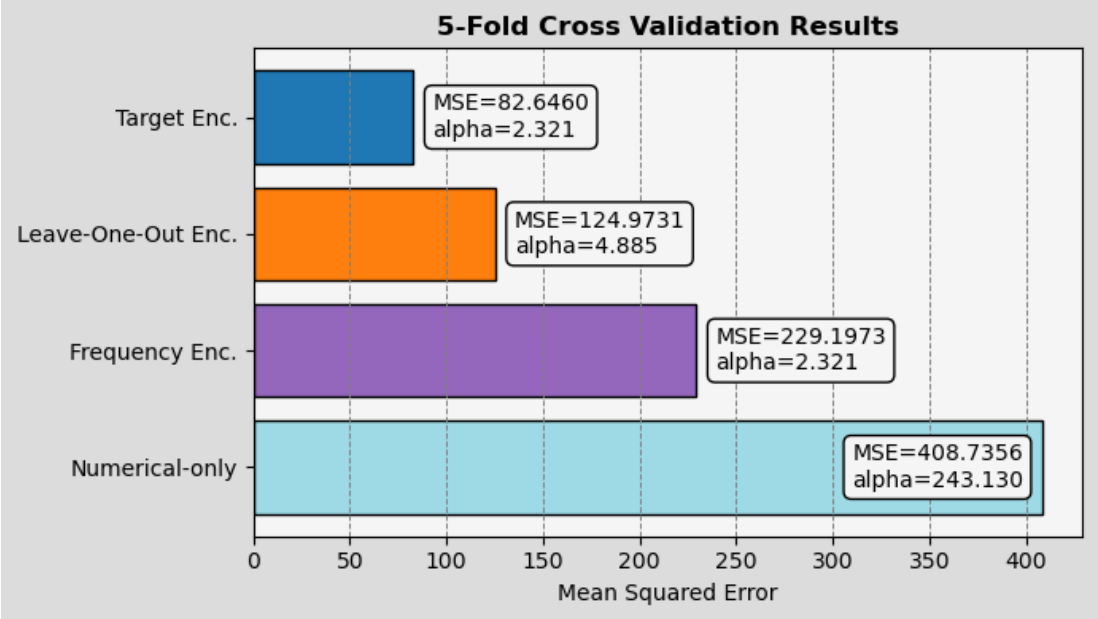

Figure 8 - Best Tuned Models

Conclusion

In this project, the Spotify track dataset with numerical and categorical features was used to train a ridge regression model defined from scratch without using any ready libraries. Data preprocessing was handled with controlling for missing data, skewness, and duplicates. Four different approaches were selected for subsetting and encoding the features for modelling preparation. All versions were modelled and fine tuned with a 5-fold cross validation.

The results showed that target encoded method outperformed all other approaches and the dataset with only the numerical features was the weakest model among the others. This implies that encoding methods for categorical features in multi-dimensional datasets plays a crucial role in models' final performance. Selecting the right encoding method requires together a comprehensive knowledge of the encoding methods and the nature of the data.

Declaration

I declare that this material, which I now submit for assessment, is entirely my 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 work. I 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 or any other person for assessment on this or any other course of study.