Predicting Release Year of Songs Based on Timbre Features Using Regression

Danielle Angelini

deangelini@wpi.edu

December 5, 2022

Table of Contents

***Abstract* – The evolution of sound quality and recording techniques has contributed greatly to the way music is recorded today. Sound features of songs have been documented from the 1920s to the 2010s, allowing data scientists to process, analyze, and predict the release year of these songs. For this project, a dataset of 515,345 songs was analyzed to see if timbre features can predict the release year of a song. Various machine learning algorithms were compared based on their test mean squared error and *R*-squared error. The significance of these features was also analyzed to see if any one feature had more of an impact on predicting a song's release year. It was concluded that the linear regression method had the best performance when including an interaction term between the two most correlated features.**

INTRODUCTION

Over the decades, sound technology has evolved from using acoustical recording methods in the 1920s to using the digital recording methods known today [4]. During this advancement in technology, researchers have been able to extract certain timbre features of each song. These timbre features have been collected and added to a well-known dataset called the Million Song Dataset.

The Million Song Dataset is a freely available collection of audio features and metadata for a million contemporary popular music tracks [7]. The information contained in this dataset has allowed many data scientists to conduct their own research and machine learning to perform data analysis, song and lyric recognition, artist and song recognition, mood classification, and song release year prediction [7]. This paper mainly focuses on using regression methods to predict the release year of a song given its audio features. A subset of the Million Song Dataset, the YearPredictionMSD, will be the main dataset used in this research.

This paper will explore the significance of these audio features in predicting the release year of a song through feature engineering. Various regression models will be used to train on this dataset, and their performance will be evaluated against each other using performance metrics. The main goal of this research is to gather whether there is a strong correlation between the audio features of a song and its release year through exploring these models. This research will also be focused on discovering whether certain audio features have more significance in predicting song release year than others.

BACKGROUND

The Million Song Dataset was first released on February 8, 2011, and its creation was supported by the National Science Foundation [7]. Much of its data has been contributed by The Echo Nest, a music intelligence and data platform [8], where they collected timbre data from a plethora of songs. In music, timbre refers to the character or quality of a musical sound; it is the quality of a sound by its overtone [6]. The researchers at The Echo Nest divided songs into segments to begin extracting the timbre data. There are 12 main timbre qualities that were extracted, some of which include the loudness, brightness, flatness, and attack of the song's sound. The values of these 12 qualities make up a 12-dimensional timbre vector to describe each segment of the song. Averages of these vectors are taken over all the song's segments, resulting in 12 timbre averages to help summarize the different qualities of the entire

song. Timbre covariances are also calculated by finding the covariance between timbre features over all the segments of the song [7].

APPROACH

The dataset used for this research is a subset of the Million Song Dataset called the YearPredictionMSD dataset. This dataset is a collection of audio features for 515,345 songs along with the target variable, the song's release year ranging from 1922 to 2011. Since songs can be released at different points throughout a year, regression methods are best suited given the continuity of time. Therefore, predicting release year based on audio data is a regression problem.

The following regression models will be used:

A. Linear and Multivariate Regression

A linear regression model will be used as a baseline model. This model will mainly be used to gather variables' predictive significance and information on multicollinearity between predictor variables. Three other variations of this model will be a part of the final model performance comparison.

B. Lasso Regression

Given the large number of predictor variables, it will be beneficial to see whether all variables are meaningful in predicting the release year. Lasso regression is useful in feature selection and regularization, and its formula is as follows:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

The first term of the formula is responsible for minimizing the reduced sum of squares (RSS) of the model, while the second term is the shrinkage penalty. The shrinkage penalty has the effect of forcing some of the coefficient estimates to be exactly zero when the tuning parameter, λ, is sufficiently large [3]. This allows for variable selection to best minimize the quantity of the formula above. Therefore, this method is beneficial to explore in this research.

C. Ridge Regression

Like Lasso regression, Ridge regression is also useful in controlling the values of coefficients with the use of its tuning parameter, λ. One main distinction between the two is that Ridge regression can only shrink coefficients towards zero without being exactly zero. Therefore, Ridge regression cannot achieve variable selection like Lasso regression. This is due to the change in the shrinkage penalty seen below:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

In the case that most features prove significant in predicting the target variable, Ridge regression is a worthy method to explore.

D. Principal Component Regression

Principal component regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA). PCA is used to generate the principal components from the predictor variables [5]. This will be decided based on which variables explain most of the variability in predicting the target variable. Like Lasso, PCR selects a subset of the variables to optimize model performance. Both PCA and PCR will be done separately to better understand the output when using PCR.

MATERIALS AND METHODS

Through the course of this research, the main programming tool used to train these

regression models on this dataset is R Studio. The following data exploration was also performed in R Studio.

A. Train and test set split

This dataset had a defined train-test split where the first 463,715 samples were dedicated to the training set, and the last 51,630 samples were dedicated to the test set.

B. Target and predictor variables

There were 91 features in this dataset - the first feature in the dataset was the target variable, the release year, and the remaining 90 were the predictor variables. Twelve of these predictor variables represented the average of each timbre quality taken over segments of a song. The remaining 78 variables represented the timbre covariances.

C. Data exploration

The dataset consisted of no missing data and no sample repeats, which simplified the data processing stage. Since methods like Lasso and Ridge regression are sensitive to the scale of the data, the data was standardized in R Studio.

Using a simple linear regression, we gathered an analytic summary of the predictor variables and their *p*-values to detect predictive significance (see Table A1). With a *p*-value of 0.05 or less, at least 64 variables had some predictive significance, and at least 59 had great predictive significance. This preliminary analysis indicated that most of the predictor variables could be considered important for use in the regression models.

When dealing with multivariate regression problems, there is a chance for multicollinearity to exist between predictors. Diagnosing this problem was done using a correlation matrix and measuring the variance inflation factor (VIF) of predictors. Variable pairs with high correlation values included the following:

Table 1: Predictor variables with high correlation values

Variable 1	Variable 2	Correlation
timbre_avg_06	timbre_cov_06	0.64793
timbre_avg_09	timbre_cov_31	0.57678
timbre_avg_06	timbre_cov_10	0.57193
timbre_avg_02	timbre_avg_01	0.56111

Variables considered to have a high potential for multicollinearity had a VIF value greater than 5. The following VIF information was gathered:

Figure 1: Variance inflation factor values for multicollinearity

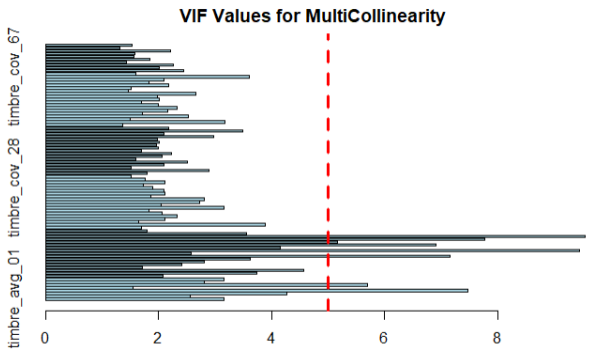

Table 2: Predictor variables indicated for high potential of multicollinearity

Variable	VIF
timbre_avg_04	7.46594
timbre_avg_06	5.70063
timbre_cov_04	7.15860
timbre_cov_06	9.44550
timbre_cov_08	6.90184
timbre_cov_09	5.16826
timbre_cov_10	7.77564
timbre_cov_11	9.55249

From looking at Table 1 and Table 2, we saw that some variables had high correlation and high potential for high multicollinearity; these variables included timbre_avg_06 and timbre_cov_06. These findings will be discussed further given their impact on choosing model scenarios.

D. Model scenarios

As we conducted model training and testing on the different algorithms discussed previously, four different scenarios were used based on the results founded during data exploration. The four scenarios are described as follows:

Table 3: Model Scenarios

Scenario Number	Scenario Description
1	Includes all features in the model
2	Includes all features except the 8 detected for high multicollinearity
3	Includes all features except the 2 with the highest correlation
4	Includes all features and an interaction term between the two variables with the highest correlation

The first scenario acted as the base scenario for comparison between the other three scenarios. The second scenario removed the eight predictor variables detected for high potential of multicollinearity (see Table 2); these were removed as one of the possible remedies for multicollinearity and to see if there was an improvement in results. The third scenario removed variables, timbre_avg_06 and timbre_cov_06, from the training and test sets. These two variables had the highest correlation score and were detected as high potential for multicollinearity, so removing these variables could help improve results. On the contrast, the last scenario created an interaction term between these two variables, timbre_avg_06 and timbre_cov_06. Combining variables acted as another potential remedy for multicollinearity, so this scenario was added to complement the third scenario.

E. Cross validation and hyperparameter tuning

For all Lasso, Ridge, and PCR regression models, cross-validation techniques were used to minimize the cross-validation error. Cross-validation chose the optimal value of λ for Lasso and Ridge regression models, and it was selected to minimize the cross-validation error. For PCR, cross-validation selected the number of components in the model based on the cross-validation error.

F. Testing the models

After training and cross-validation, the models were tested on the test dataset to measure their performance. This was measured based on the following metrics:

- Mean Squared Error (MSE): Purpose was to measure how close a regression line is in representing the dataset. This was done by measuring the mean squared difference between observed and predicted values.
- R-Squared Value: Purpose was to show how well the data fits the regression model. This was done by determining the proportion of variance in the target variable that could be explained by the predictor variables [1].

RESULTS

Following are the results of testing the four regression algorithms with each of the scenarios:

Table 4: Comparison Between Regression Models Under Four Scenarios

Models	MSE	*R*-Squared
Linear 1	0.7682353	0.2317499
Linear 2	0.7841534	0.2158314
Linear 3	0.7786483	0.2213366
Linear 4	0.7675689	0.2324162
Ridge 1	0.7685234	0.2314617
Ridge 2	0.7844898	0.2154950
Ridge 3	0.7788325	0.2211524
Ridge 4	0.7678458	0.2321394
Lasso 1	0.7792073	0.2207776
Lasso 2	0.7944189	0.2055657
Lasso 3	0.7901709	0.2098138
Lasso 4	0.7783950	0.2215899
PCR 1	0.7812382	0.2187466
PCR 2	0.7901583	0.2098264
PCR 3	0.7893180	0.2106667
PCR 4	0.7811047	0.2188802

The average performance for each model is summarized in the following table:

Table 5: Average Performance by Each Regression Model

Models	MSE	*R*-Squared
Linear	0.7746515	0.2253335
Ridge	0.7749229	0.2250621
Lasso	0.7855480	0.2144368
PCR	0.7854548	0.2145299

The performances among all the models were similar in their mean squared error and *R*-squared values, with a slight improvement when using the linear regression model. Meanwhile, the Lasso regression model performed the worst on average.

The following results summarize the best performances given the model and its optimal scenario:

Table 6: Best Performances by Each Regression Model

Models	MSE	*R*-Squared
Linear 4	0.7675689	0.2324162
Ridge 4	0.7678458	0.2321394
Lasso 4	0.7783950	0.2215899
PCR 4	0.7811047	0.2188802

For all model runs, the fourth scenario seemed to be the most optimal based on performance results. The linear regression model also performed the best compared to the other models given this scenario, while the PCR model performed the worst.

As for predictor variable significance, the linear regression model testing provided a summary, which can be referenced in the Appendix as Table A1. This table shows significance for many of the predictors. More specifically, the variables timbre_avg_01 and timbre_avg_02 had the highest absolute *t*-values, indicating a greater confidence in their coefficient values.

DISCUSSION

The main notable findings from this research included that the linear regression model produced the best results, the Lasso regression model produced the worst results on average, all models under the fourth scenario performed the most optimally, and most of the predictor variables proved to be of at least some significance in predicting the target variable.

The comparison between the linear regression, Lasso regression, and PCR models highlights that this dataset does not benefit from using a subset of the predictor variables. This is also

supported by Table A1 given that many of the predictor variables had at least some predictive significance. This table summary also shows two variables with the most potential in predicting the target variable. With the highest *t*-values, timbre_avg_01, the average loudness, and timbre_avg_02, the average brightness, could benefit from further research in their relationship with release year.

As for the overall performance of these regression models, the average mean squared errors and *R*-squared values indicated only some correlation between timbre features and song release year. This could be a result of the limited number of scenarios that these models were tested under. Model performance could also have been limited through mainly exploring algorithms that focused on variable selection.

CONCLUSION AND FUTURE WORK

The research summarized in this paper goes over the performance of various regression models to model song release year based on its timbre features. A series of steps were taken under data exploration to best reflect the characteristics of the dataset for these model runs. Cross-validation was used to tune model hyperparameters where applicable and to produce a model that minimized the cross-validation error. It was concluded that the linear regression model gave the most optimal performance, and it performed best when interacting the most highly correlated variables. Findings also included that song loudness and song brightness were not only predictively significant, but they also carried the most confidence in their coefficients.

It is worth noting that model performance could benefit from further exploration in model scenarios. Some examples include involving song loudness and brightness more heavily and focusing more on algorithms that use the entire dataset in modeling. Other avenues of research include exploring more complex algorithms, such as using random forest methods to address this problem.

REFERENCES

[1] CFI Team. (2022, November 24). *R-squared*. Corporate Finance Institute. Retrieved December 5, 2022, from https://corporatefinanceinstitute.com/resources/data-science/r-squared/#:~:text=R%2DSquared%20(R%C2%B2%20or%20the,(the%20goodness%20of%20fit).

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning: with applications in R. New York: Springer.

[4] Kelley, A. (2022, October 13). *Evolution of sound – audio technology past, present, and future*. Evolution of Sound | Audio Technology Past, Present, and Future | Audio Visual. Retrieved December 5, 2022, from https://insights.ges.com/us-blog/evolution-of-sound-audio-technology-past-present-and-future.

[5] Leung, K. (2022, September 14). *Principal component regression - clearly explained and implemented*. Medium. Retrieved December 5, 2022, from https://towardsdatascience.com/principal-component-regression-clearly-explained-and-implemented-608471530a2f.

[6] Merriam-Webster. (n.d.). *Timbre definition & meaning*. Merriam-Webster. Retrieved December 5, 2022, from https://www.merriam-webster.com/dictionary/timbre

[7] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Societyfor Music Information Retrieval Conference (ISMIR 2011), 2011.

[8] Wikimedia Foundation. (2022, October 7). *The Echo Nest*. Wikipedia. Retrieved December 5, 2022, from https://en.wikipedia.org/wiki/The_Echo_Nest#cite_note-Spotify-4.

APPENDIX

Table A1: Linear Regression Model Summary of Predictor Variables

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.9338 -0.3074  0.1591  0.5321  6.1205

Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)    1.427e-16  1.282e-03    0.000 1.000000
timbre_avg_01  4.854e-01  2.280e-03  212.892  < 2e-16 ***
timbre_avg_02 -2.633e-01  2.053e-03 -128.290  < 2e-16 ***
timbre_avg_03 -1.401e-01  2.652e-03  -52.815  < 2e-16 ***
timbre_avg_04  5.245e-03  3.504e-03    1.497 0.134469
timbre_avg_05 -3.104e-02  1.592e-03  -19.492  < 2e-16 ***
timbre_avg_06 -2.599e-01  3.062e-03  -84.884  < 2e-16 ***
timbre_avg_07 -8.340e-03  2.151e-03   -3.877 0.000106 ***
timbre_avg_08 -7.409e-02  2.277e-03  -32.537  < 2e-16 ***
timbre_avg_09 -6.549e-02  1.846e-03  -35.473  < 2e-16 ***
timbre_avg_10  1.357e-02  2.478e-03    5.474 4.39e-08 ***
timbre_avg_11 -6.566e-02  2.741e-03  -23.951  < 2e-16 ***
timbre_avg_12 -1.520e-03  1.676e-03   -0.907 0.364207
timbre_cov_01  9.580e-02  1.989e-03   48.172  < 2e-16 ***
timbre_cov_02  5.712e-02  2.149e-03   26.586  < 2e-16 ***
timbre_cov_03 -5.009e-02  2.438e-03  -20.543  < 2e-16 ***
timbre_cov_04  6.118e-02  3.431e-03   17.832  < 2e-16 ***
timbre_cov_05  1.896e-02  2.057e-03    9.217  < 2e-16 ***
timbre_cov_06  7.642e-02  3.941e-03   19.388  < 2e-16 ***
timbre_cov_07  5.730e-02  2.611e-03   21.943  < 2e-16 ***
timbre_cov_08  6.240e-02  3.369e-03   18.521  < 2e-16 ***
timbre_cov_09  1.427e-02  2.916e-03    4.895 9.83e-07 ***
timbre_cov_10 -8.324e-03  3.576e-03   -2.328 0.019924 *
timbre_cov_11  1.297e-01  3.964e-03   32.732  < 2e-16 ***
timbre_cov_12  3.841e-02  2.417e-03   15.896  < 2e-16 ***
timbre_cov_13 -3.987e-02  1.717e-03  -23.219  < 2e-16 ***
timbre_cov_14  2.463e-03  1.673e-03    1.472 0.141029
timbre_cov_15  7.957e-02  2.529e-03   31.460  < 2e-16 ***
timbre_cov_16  9.481e-03  1.643e-03    5.771 7.90e-09 ***
timbre_cov_17  1.367e-02  1.861e-03    7.346 2.04e-13 ***
timbre_cov_18 -2.938e-03  1.955e-03   -1.503 0.132778
timbre_cov_19 -1.126e-02  1.840e-03   -6.117 9.56e-10 ***
timbre_cov_20 -9.236e-03  1.735e-03   -5.324 1.01e-07 ***
timbre_cov_21 -3.748e-02  2.276e-03  -16.464  < 2e-16 ***
timbre_cov_22  1.258e-02  1.834e-03    6.858 7.01e-12 ***
timbre_cov_23  6.311e-03  2.118e-03    2.980 0.002881 **
timbre_cov_24 -5.137e-02  2.148e-03  -23.915  < 2e-16 ***
timbre_cov_25 -9.395e-03  1.752e-03   -5.362 8.22e-08 ***
timbre_cov_26  2.751e-02  1.866e-03   14.738  < 2e-16 ***
timbre_cov_27  3.351e-02  1.857e-03   18.047  < 2e-16 ***
timbre_cov_28 -3.241e-02  1.766e-03  -18.349  < 2e-16 ***
timbre_cov_29 -2.288e-02  1.686e-03  -13.567  < 2e-16 ***
timbre_cov_30 -8.874e-03  1.860e-03   -4.772 1.83e-06 ***
timbre_cov_31 -1.018e-02  1.698e-03   -5.997 2.01e-09 ***
timbre_cov_32 -8.492e-03  1.580e-03   -5.375 7.66e-08 ***
timbre_cov_33 -1.230e-02  1.715e-03   -7.173 7.32e-13 ***
timbre_cov_34  3.348e-02  2.178e-03   15.370  < 2e-16 ***
timbre_cov_35  2.052e-02  1.573e-03   13.045  < 2e-16 ***
timbre_cov_36 -4.893e-02  1.858e-03  -26.337  < 2e-16 ***
timbre_cov_37  5.361e-03  2.031e-03    2.639 0.008317 **
timbre_cov_38  2.296e-02  1.617e-03   14.200  < 2e-16 ***
timbre_cov_39  8.966e-05  1.841e-03    0.049 0.961146
timbre_cov_40 -1.203e-02  1.911e-03   -6.291 3.15e-10 ***
timbre_cov_41  1.192e-02  1.670e-03    7.138 9.48e-13 ***
timbre_cov_42  8.752e-04  1.812e-03    0.483 0.629011
timbre_cov_43 -1.631e-03  1.793e-03   -0.910 0.362915
timbre_cov_44  4.068e-03  1.816e-03    2.240 0.025067 *
timbre_cov_45 -5.350e-02  1.806e-03  -29.617  < 2e-16 ***
timbre_cov_46  4.572e-02  2.211e-03   20.681  < 2e-16 ***
timbre_cov_47 -1.917e-02  1.855e-03  -10.331  < 2e-16 ***
timbre_cov_48  4.038e-03  2.397e-03    1.685 0.092013 .
timbre_cov_49 -1.514e-02  1.893e-03   -7.998 1.26e-15 ***
timbre_cov_50 -7.815e-03  1.501e-03   -5.209 1.90e-07 ***
timbre_cov_51 -2.720e-02  2.283e-03  -11.918  < 2e-16 ***
timbre_cov_52  3.280e-02  1.568e-03   20.921  < 2e-16 ***
timbre_cov_53 -4.161e-02  2.037e-03  -20.428  < 2e-16 ***
timbre_cov_54  8.785e-03  1.678e-03    5.236 1.64e-07 ***
timbre_cov_55 -1.310e-03  1.882e-03   -0.696 0.486362
timbre_cov_56 -4.869e-03  1.956e-03   -2.489 0.012813 *
timbre_cov_57 -4.104e-02  1.811e-03  -22.657  < 2e-16 ***
timbre_cov_58 -1.748e-02  1.667e-03  -10.488  < 2e-16 ***
timbre_cov_59 -2.371e-02  1.815e-03  -13.062  < 2e-16 ***
timbre_cov_60  6.641e-03  1.807e-03    3.676 0.000237 ***
timbre_cov_61  1.141e-02  2.091e-03    5.456 4.86e-08 ***
timbre_cov_62  2.465e-02  1.551e-03   15.899  < 2e-16 ***
timbre_cov_63  1.769e-02  1.573e-03   11.243  < 2e-16 ***
timbre_cov_64  3.629e-02  1.895e-03   19.152  < 2e-16 ***
timbre_cov_65  5.140e-03  1.736e-03    2.961 0.003063 **
timbre_cov_66 -5.933e-02  1.858e-03  -31.933  < 2e-16 ***
timbre_cov_67 -7.489e-05  2.435e-03   -0.031 0.975462
timbre_cov_68 -1.737e-03  1.622e-03   -1.071 0.284090
timbre_cov_69 -2.136e-03  2.002e-03   -1.067 0.286114
timbre_cov_70 -1.015e-02  1.815e-03   -5.590 2.27e-08 ***
timbre_cov_71  1.498e-02  1.926e-03    7.777 7.42e-15 ***
timbre_cov_72  8.628e-03  1.534e-03    5.625 1.85e-08 ***
timbre_cov_73  4.065e-02  1.742e-03   23.331  < 2e-16 ***
timbre_cov_74  7.911e-04  1.606e-03    0.493 0.622336
timbre_cov_75  1.876e-02  1.609e-03   11.658  < 2e-16 ***
timbre_cov_76 -3.789e-02  1.907e-03  -19.863  < 2e-16 ***
timbre_cov_77 -2.268e-02  1.465e-03  -15.482  < 2e-16 ***
timbre_cov_78 -2.437e-03  1.588e-03   -1.535 0.124767
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8733 on 463624 degrees of freedom
Multiple R-squared:  0.2375,    Adjusted R-squared:  0.2373
F-statistic:  1604 on 90 and 463624 DF,  p-value: < 2.2e-16
```

Project_Danielle Angelini_v2.RMD