Song's energy predictor with multiple linear regression

Ana Dueñas Chávez – A01702080

Abstract – This document presents the implementation, explanation, and comparison between a linear regression model by hand and with sklearn framework to determine the energy of a song based on Spotify records.

I. INTRODUCTION

According to Gilbert Galindo music is very important in our lives. Each song transmits something that affects us in different ways. It can raise our mood, get us excited, make us fell calm, relax, and sometimes even nearly to emotions that we already experience.

This project focuses the energy of a song, that is associated a state form calm to happy or excited.

II. DATASET

The dataset used to develop this project was taken from Kaggle and made by Yamac Eren Ay, a self-taught data scientist and music enthusiast[1].

This compilation of data contains more than 175000 songs collected between 1921 to 2020 from the Spotify web API. Each instance includes audio and track features.

Detail Compact Column 10 of 19 columns

# acousticne...	artists	# danceability	# duration_ms	# energy	# expli
0.991	['Mamie Smith']	0.598	168333	0.22399999999999998	0
0.643	["Screamin' Jay Hawkins"]	0.852	150200	0.517	0
0.993	['Mamie Smith']	0.647	163827	0.18600000000000005	0
0.000173	['Oscar Velazquez']	0.73	422087	0.7979999999999999	0
0.295	['Mixe']	0.7040000000000001	165224	0.7070000000000001	1
0.996	['Mamie Smith & Her Jazz Hounds']	0.424	198627	0.245	0

Figure 1 Part of Kaggle dataset.

Since the dataset includes 19 fields for each song, the data was filtered to keep just the relevant features for the value to predict. To do this, a correlation between data was made to observe how each field was related to the y's. The chosen attributes were selected only if their correlation result was higher or equal

[1] https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

than 0.5, so the relationship between the data was high. The selected fields were:

- Loudness: Increasing audio levels in music, measured from -60 to 0.
- Popularity: refers to how attractive each song is, measured from 0 to 100.
- Valence: represent happiness, measured from 0 to 1.
- Year: time where the song was released from 1921 to 2020.

	acousticness	danceability	duration_ms	energy	explicit	instrumentalness
acousticness	1.000000	-0.263217	-0.089169	-0.750852	-0.208176	0.221956
danceability	-0.263217	1.000000	-0.100757	0.204838	0.200842	-0.215589
duration_ms	-0.089169	-0.100757	1.000000	0.060516	-0.033808	0.103621
energy	-0.750852	0.204838	0.060516	1.000000	0.102561	-0.177750
explicit	-0.208176	0.200842	-0.033808	0.102561	1.000000	-0.130609
instrumentalness	0.221956	-0.215589	0.103621	-0.177750	-0.130609	1.000000
key	-0.028028	0.026266	0.002020	0.035780	0.005282	-0.004619
liveness	-0.029654	-0.110033	0.028942	0.134815	0.037288	-0.047941
loudness	-0.546639	0.249541	0.019791	0.779267	0.106249	-0.317562
mode	0.064633	-0.048358	-0.046849	-0.056160	-0.062503	-0.056731
popularity	-0.396744	0.123746	0.024717	0.328939	0.152545	-0.300625

Figure 2 Correlation table between features and the value to predict.

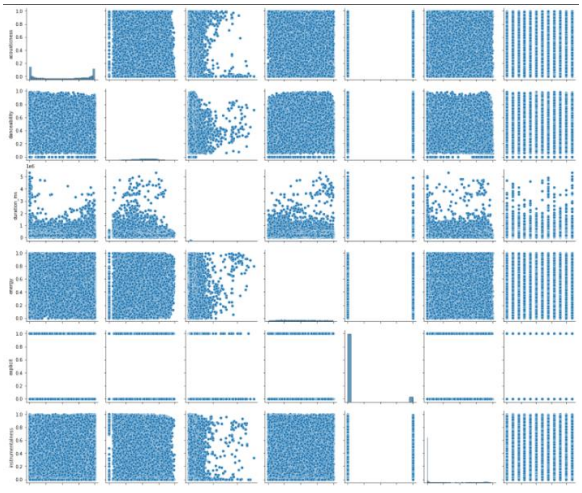

Figure 3 Portion of the correlation graph.

Furthermore, to get better predictions data was splinted in 66% to train the linear model and 33% to test its efficiency.

III. APPROACH

The first objective of the project is to determine how energic is a song based of the numerical features it has. This prediction can be performed by a multiple linear regression so that it can model a linear relationship between the features and the value to predict.

Additionally, the second objective of this project is to make a comparison of the results between the implementation by hand and the sklearn method.

Both linear regression implementation by hand and the framework are based on the Ordinary Least Squares. This statistical method estimates an unknown parameter in a linear regression model.

$$y = \beta X + \epsilon$$

Figure 4 Ordinary least squares method where x represents the features, beta a vector of parameter to estimate and epsilon the error.

The implementation by hand is based on the Multiple linear regression with NumPy article from Dario Radečić.

IV. RESULT

Variance was used to determine how far was each original instance form the average predicted values. In this project the variance for both implementations was 67%.

To evaluate the model skills to predict the energy of each song, this project used cross-validation score. For both implementations was 67%.

Finally, the mean squared error was measured to determine which prediction was better between the sklearn and the scratch model. The MSE of the framework implementation and the scratch one was 0.16.

V. CONCLUSION

As we can see, sklearn framework and by hand methods had pretty good and similar outcomes, 67% of variance, 67% cross validation score and the mean square error 0.16, since they were using the same statistical model. Based on the results, we can say that both implementations are as good as the other one. However, if we take into consideration the additional tools that the framework provides, it is better.

VI. REFERENCES

[1] Wikipedia, February 2021. [Online]. Available: https://en.wikipedia.org/wiki/Variance. [Accessed 24 February 2021].

[2] Wikipedia, Ferbuary 2021. [Online]. Available: https://en.wikipedia.org/wiki/Ordinary_least_squares [1]. [Accessed 24 February 2021].

[3] Scikit-learn, [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. [Accessed 24 February 2021].

[4] W. Kenton, "Investopedia," February 2021. [Online]. Available: https://www.investopedia.com/terms/m/mlr.asp. [Accessed 24 February 2021].

[5] Y. Eren, "Kaggle," January 2021. [Online]. Available: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-

tracks. [Accessed 24 February 2021].

[6] D. Radečić, "Towards data science," October 2019. [Online]. Available: https://towardsdatascience.com/multiple-linear-regression-from-scratch-in-numpy-36a3e8ac8014. [Accessed 24 February 2021].

[7] G. Galindo, November 2003. [Online]. Available: https://www.gilbertgalindo.com/importanceofmusic#:~:text=Music%20can%20raise%20someone's%20mood,we%20experience%20in%20our%20lives. [Accessed February 2021].