Prediction of the Spotify Song's Valence

Denizcan Bozkurt
Middle East Technical University
Ankara, Turkey
denizcan.bozkurt@metu.edu.tr

Abstract*— This research offers a comprehensive examination of the variables affecting the valence of Spotify tracks and predictive models created through various machine learning methods. The connections between these features and the song's valence were investigated using a dataset of 1000 songs with 12 attributes, including audio features and metadata. To predict the valence of songs, the study uses machine learning techniques like Support Vector Machine, Neural Network, XGBoost, and Random Forest, and statistical models by using R Studio. Preprocessing procedures such as outlier analysis and data cleaning are carried out to improve the data quality. The predictive abilities of the models are assessed using R2, MAE, and RMSE. Regarding MAE, the General Linear Model (GLM) performs better than other models because of its less sensitivity to outliers. Results show that danceability and energy are the most important indicators of valence in a song.

Keywords—GLM, SVM, Artificial Neural Network, Random Forest, Xgboost, Spotify, Valence

I. INTRODUCTION

Music strongly influences human emotions, and knowing what influences a piece of music's emotional tone can be useful for various applications, such as psychological treatments or tailored recommendations. One of the most important characteristics that determines a song's mood is valence, which measures the musical positivity generated by the song. Broad preprocessing procedures are carried out to ensure data quality because of the complex nature of the data and the existence of outliers. To determine the best method for valence prediction, this study evaluates the predictive capabilities of multiple machine learning models, such as Support Vector Machine, Neural Network, XGBoost, and Random Forest. This study aims to develop and evaluate machine learning and statistical models for accurately predicting the valence of songs, which is an important measure of their mood. By achieving this, music firms can enhance their recommendation systems to better cater to users' emotional states. This study leverages the audio features and metadata associated with songs to build predictive models. Implementing sophisticated machine learning techniques is intended to assess how these models can predict the valence of songs, thus facilitating a more personalized music recommendation experience. Through this study, it is anticipated that significant insights will be gained into the predictive power of various audio features and the overall feasibility of mood-based song recommendations.

II. LITERATURE REVIEW

More complex models were used as machine learning techniques advanced. Because of their propensity to handle high-dimensional data and capture intricate relationships between features, support vector machines, or SVMs, gained popularity. SVM was used by Laurier and Herrera (2007) to predict the emotional content of music, showing how well it can differentiate between various emotional states [1]. Rachman et al. (2019) proposed a rule-based method for detecting song emotion using arousal and valence values. They used Random forests to categorize the song's emotion [2]. However, in this study, there was no classification of songs' moods because of a lack of information about the categorization of emotions in songs.

III. METHODOLOGY

A. Dataset

In this study, the dataset is taken from the Kaggle repository. The original dataset consists of 18454 observations with 25 different attributes. However, 1000 observation and 12 attributes were selected. Names of each variable used in this study are; "track_artist", "playlist_genre", "danceability", "energy", "key", "loudness", "speechiness", "acousticness", "liveness", "valence", " tempo ".

Because of lack of computational power and data complexity, some of the audio features were excluded. To make the data more manageable, the sample () function in R utilized to select a smaller portion of the data. In addition, outliers were not removed because there were too many of them. To remove them, could distort the statistical analysis and result in inaccurate conclusions or interpretations. The data descriptions of each variable used in this study are as follows.

1. track_artist ~ It represents song artist.
2. playlist_genre ~ It represents song genre.
3. danceability ~ It describes how suitable a track is for dancing.
4. energy ~ It represents a perceptual measure of intensity and activity.
5. key ~ The estimated overall key of the track. (E.g. 0 = C, 1 = C♯/D♭, 2 = D)
6. loudness ~ The overall loudness of a track in decibels (dB).
7. speechiness ~ It detects the presence of spoken words in a track.
8. acousticness ~ A confidence measure of whether the track is acoustic.
9. instrumentalness ~ Demonstrates whether a track contains no vocals.
10. liveness ~ Detects the presence of an audience in the recording.
11. valence ~ Describes the musical positiveness conveyed by a track.
12. tempo ~ The overall estimated tempo of a track

B. Descriptive Statistics

Descriptive statistics table are shown below. It has been obtained first, since it gives an insight into the data set at the beginning of the exploration. Besides, these values help to create research questions in the next steps of the analysis. A descriptive summary is attached in Table 1 for numerical attributes.

Table 1Descriptive Statistical Summary Some of Numerical Data

	liveness	speechiness	energy	danceability	valence
Min.	0.023	0.023	0.115	0.176	0.033
1st Q.	0.094	0.038	0.560	0.525	0.309
Median	0.127	0.056	0.700	0.634	0.487
Mean	0.191	0.103	0.683	0.628	0.487
3rd Q.	0.239	0.125	0.832	0.743	0.652
Max.	0.973	0.605	0.995	0.973	0.990
NA's	46	54	41	56	35

This table displays summary statistics for five numeric variables in the dataset. For "liveness," the values range from 0.02320 to 0.97300, with a mean of 0.19156 and 46 missing values. "speechiness" ranges from 0.0230 to 0.6050, with a mean of 0.1034 and 54 missing values. "Energy" varies from 0.1150 to 0.9950, with a mean of 0.6831 and 41 missing values. "Danceability" ranges from 0.1760 to 0.9730, with a mean of 0.6289 and 56 missing values. "Valence" varies from 0.0339 to 0.9900, with a mean of 0.4875 and 35 missing values. “valence” looks normal because of the mean and median. However, “liveness” and "speechiness” mean values are greater than the median. It suggests that the distribution is right-skewed. These statistics provide insights into the distribution and characteristics of each variable. Frequency for categorical Genre variable is attached in Figure 1.

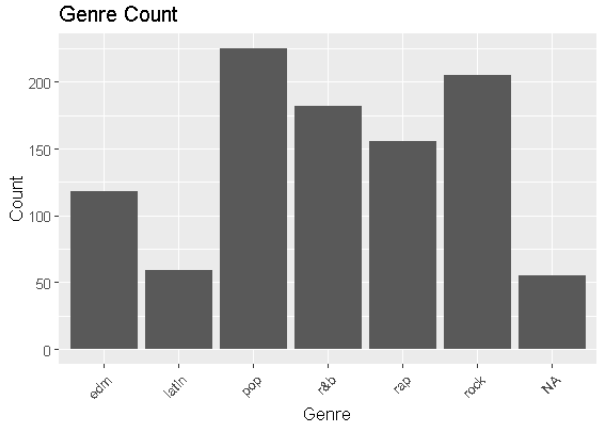

Figure 1Bar Chart for Genre’s frequency

Pop genre has the highest proportion and Latin genre has the least proportion. It can affect model because of different count of genre. Also, there are some NA values in genre.
The other categorical variable is “artist”. There are 753 unique artist and most frequent artist count is 6. Artist category may increase dimensions.

C. Exploratory/Confirmatory Data Analysis

The following part presents the answers to the four research questions, which were formulated using appropriate statistical techniques. These statistics provide a general understanding of the variables and allow us to advance the analysis.

C.1 What is the distribution of numeric variables?

The distribution of those variables within the dataset was analyzed to determine the numeric variables' shape, central tendency, and variability. It can be interpreted with limited descriptive statistics. A pair plot is attached in Figure 2 for numerical attributes to investigate distribution of variables.

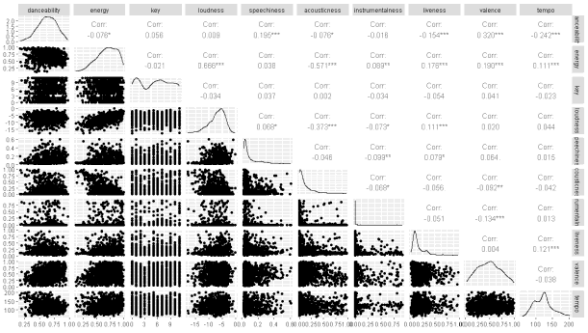

Figure 2 Pair Plot for numerical attributes

The plot demonstrates that the most relationships are not linear, indicating the complexity of interactions between these audio features. Features such as energy, loudness, and danceability tend to be positively correlated with valence, meaning that songs with higher values in these features are more likely to have a positive mood. valence variable seems not normal and zero skew. Other variables are right or left skewed.

We can verify the result shown by the previous plot using Shapiro Test. Its result confirm that valence value is not normally distributed. Shapiro test result was given below Table 2.

Table 2 Shapiro Test Result

valence	p-value
W = 0.98365	3.786e-09

Shapiro test shows that valence is not normal because p value is less than 0.05.

C.2 Do different genres have significantly different valance means?

If the means of valence values throughout genres differ, it implies that the valence variable represents the emotional tone that a piece of music conveys, and the genre of the music is related. This might suggest that some musical genres create more emotionally charged music than others, either positively or negatively. Valances mean distribution is attached in Figure 3.

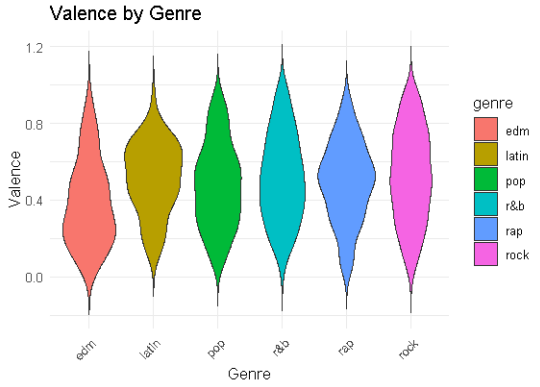

Figure 3 Violin Plot for Mean of Valence

The result shown by the previous plot using Kruskal Wallis can be verified. Kruskal Wallis was used because valence is not normally distributed ($p<0.05$). The Kruskal Wallis p-value was less than 0.05. Then, the post-hoc test was applied. There is a significant difference in valence mean between the EDM genre and other genres($p<0.05$). Others are not different from each other($p>0.05$).

C.3 Are there any outliers in dataset?

For accurate data analysis and modeling, outliers must be recognized and understood. Whether by transformation, removal, or robust modeling techniques, handling outliers correctly is essential to guarantee the validity and dependability of the outcomes. Box plot for outliers is attached to Figure 4.

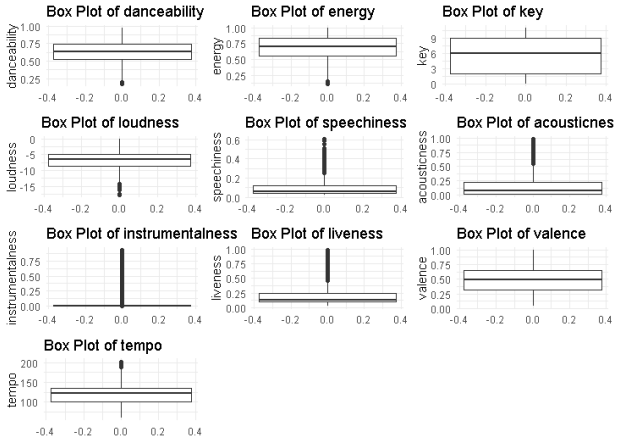

Figure 4 Box Plot for Outliers

The above plot displays, there are outliers in the data. It is checked by quartile range and 667 observations were outliers.

C.4 What is the relationship between numeric variables?

In the data set, especially valence value may be affected by other variables. Multicollinearity can lead to misleading interpretations of the relationships between predictors and the response variable. To check this, a correlation plot was used, and it is attached to Figure 5.

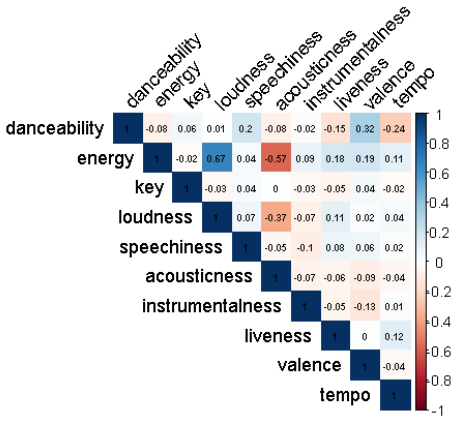

Figure 5 Corr Plot for Variables Relationships

The correlation plot shows that there seems to be multicollinearity between energy and loudness with 0.67.

To check multicollinearity, a linear model can be applied. However, energy and loudness were not normally distributed by the Shapiro Wilk test($p<0.05$). A robust linear model was applied. Both the intercept and the coefficient for energy and loudness are statistically significant, as indicated by their respective t-values, 86.93 and 29.25, respectively.

D. Imputation and Feature Engineering

In almost all significant statistical analyses, missing data occurs. It's critical to conclude on the mechanism of missingness when looking into missing data. This involves figuring out the cause of missing data. We have missing values in our data set as a result of the complete random missing mechanism. It denotes the likelihood that the values of the independent variables and all other variables in the data set have no bearing on the absence of any observations. In this study, mice packages were selected for imputation. Missingness percentages were given below chart Figure 6.

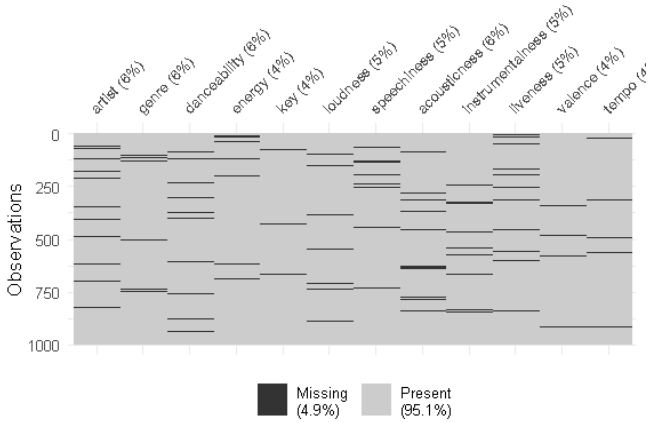

Figure 6 Missingness Chart

The mice package helps in adding the proper values to the data set's missing values. These numbers come from a distribution that was created specifically for each missing value in the set of data. As a result, there are numerous ways to complete the data gaps. Given the significance of the data's validity for analysis following imputation. We verify the answers to the questions in the EDA section to make sure our imputations are appropriate for further analysis. Numeric

variables are filled by mice package, and categorical variables are filled by their mode because it cannot filled by MICE package because there was two categorical variable and one of them had 753 unique observations. Kernel density was checked before continuing, and it seems there was no significant change in distributions is given in Figure 7.

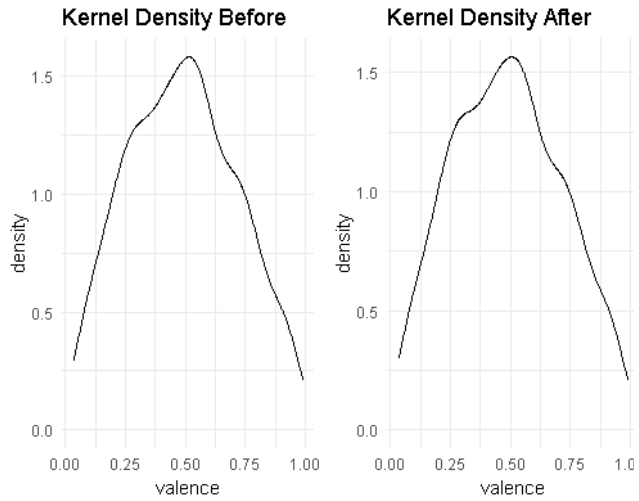

Figure 7 Density Plot for Valence

As mentioned above, the plots, the left density plot, were drawn before the imputation process, and the right bar plot was drawn after the imputation process. It can be clearly seen that both graphs have the same shape. Therefore, it can be said that our imputation of the missing values in the data set is appropriate for continuing.

Then, one-hot encoding was applied to convert categorical variables into binary indicator variables. However, this process significantly increased the data dimensions from 12 to 768 features. To manage this high-dimensional dataset, dimension reduction was performed using the Boruta algorithm. The Boruta method identified and retained the most relevant features, successfully reducing the number of features from 768 to 14. Most relevant features given below chart Figure 8.

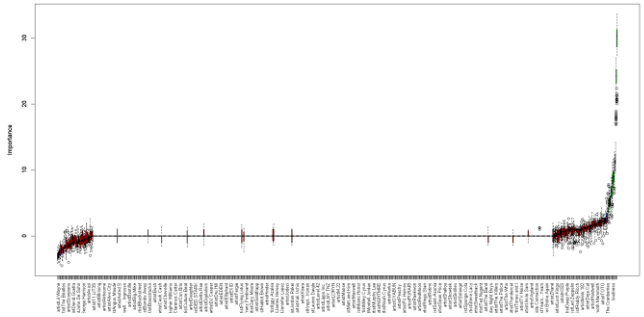

Figure 8 Dimension Reduction Chart

The most relevant features identified for the analysis are "artistMeghan.Trainor," "artistThe.Cranberries," "genreedm," "genrer.b," "genrerap," "genrerock," "danceability," "energy," "loudness," "speechiness," "acousticness," "instrumentalness," "liveness," and "tempo." The dependent variable in the study is "valence." These selected features will be utilized for building and evaluating the machine learning models to predict valence, thereby determining a song's mood. In our data preprocessing stage, efforts were made to normalize the data; however, achieving normalization cannot achieved. The distribution of the data and the presence of outliers within the dataset challenged the normalization process. Despite these challenges, alternative techniques were explored to address data normalization concerns and ensure robust modeling outcomes.

E. CV and Train/Test Split

For computational efficiency and robust evaluation, a choice was made for k-repeated cross-validation with 10 folds and 5 repeats, along with a train-test split ratio of 80% training data and 20% testing data. This approach thoroughly assesses the model's performance while balancing computational resources and reliability. This method facilitates comprehensive validation, allowing the model to generalize well to unseen data while leveraging the full potential of the available dataset for training.

F. Modelling

1. General Linear Model

Since normality cannot achieved, GLM was used instead of MLR. GLM is more general than the linear model since, in GLM, the response variable can have a nonnormal distribution.

In this study, some of the variables were eliminated because of their insignificance. The final output of the model is given below Table 3.

Table 3 Summary of Final GLM

Coefficients	Estimate	Std.Er	T val.	Pr(>|t|)
Intercept	-0.417	0.070	-5.893	<0.05
artistMeghan.Trainor	0.337	0.114	2.961	<0.05
genreedm	-0.109	0.024	-4.484	<0.05
genrer.b	0.043	0.023	2.102	<0.05
genrerap	-0.045	0.021	-2.104	<0.05
genrerock	0.097	0.021	4.599	<0.05
danceability	0.629	0.052	12.023	<0.05
energy	0.582	0.059	9.71	<0.05
loudness	-0.012	0.003	-3.691	<0.05
acousticness	0.116	0.034	3.084	<0.05
instrumentalness	-0.169	0.044	-3.837	<0.05
Nulldeviance:41.702 on 800 df Residualdeviance: 30.342 on 790 df	AIC:-324.8			

There is still some unexplained variability even though the model fits the data quite well according to the residuals and the deviance($R^2=0.273$). These predictors are significant in explaining the variability in valence, according to the significant p-values. Higher values in these features are linked to lower valence, as indicated by the negative coefficients for instrumentality and loudness. Songs with higher values in danceability, energy, and acousticness are

likely to have a higher valence, according to the positive coefficients for these characteristics. If danceability is equal to 1, then valence is increased 0.629.

2. *Support Vector Machine*

Support Vector Machine is a supervised learning algorithm that can be used for regression problems. By taking into consideration the following characteristics, we attempt to predict the response variable using all variables except valence. After determining the features, SVM is set as regression. A grid search with cross-validation was utilized to evaluate these parameters' combinations systematically. The cost and gamma parameters were fine-tuned from 0 to 1. Then, cost parameter was tuned as 0.25 and gamma parameter was tunes as 0.2. Regression performances is attached to Figure 9.

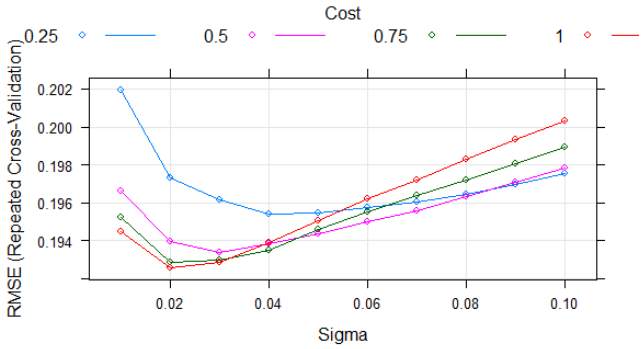

Figure 9 Regression Performance of SVM

3. *Artificial Neural Networks*

Artificial Neural Network is a supervised learning algorithm that can be used for regression problems. A grid search with cross-validation was utilized to evaluate these parameters' combinations systematically. Decay was tuned as 0.1. The model has five layers with one neuron. The plot of the model is given below in Figure 10.

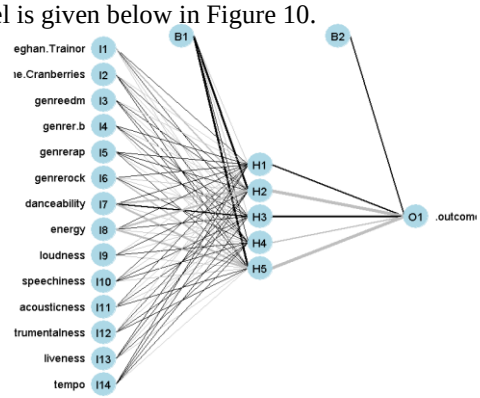

Figure 10 ANN Model

4. *Random Forests*

Random forests are a supervised learning based on tree algorithms which is applicable for regression problems. A cross-validation grid search was employed to assess these parameters' combinations.

Ntree, mtree, maxnodes numbers were tuned as 400,6 and 5 respectively. Its performance chart is attached to Figure 11.

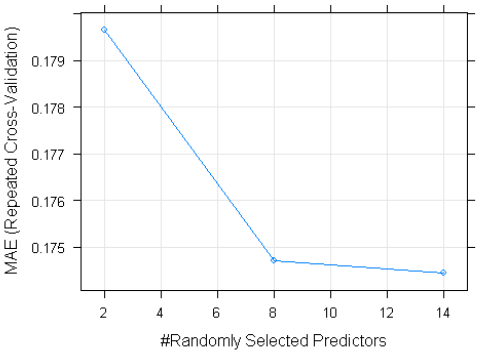

Figure 11 Regression Performance of Random Forest

5. *XgBoost*

Xgboost is a machine learning method based on tree algorithm and can be used for regression problems. A cross-validation grid search was employed to assess these parameters' combinations.
The eta, gamma colsample_bytree, min_child_weight, subsample, nrounds and max_depth parameters are tuned as 0.075, 0, 0.5, 2, 1, 100 and 1, respectively. Increasing max depth led to an overfitting model. Its chart is attached to Figure 12.

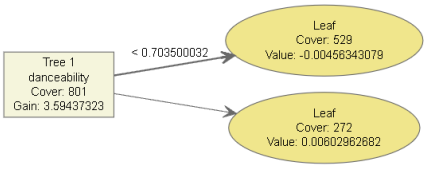

Figure 12 Regression Performance of XgBoost

G. *Performance Comparison on Train/Test Data*

In this work, RMSE, MAE and R^2 were used to identify the performance of the methods. Train performances of models is attached to Table 4. Test performances of models is attached to Table 5.

Table 4 Train Performances of Models

Model	RMSE	MAE	R^2
GLM	.1942563	.1593390	.2751846
SVM	.1952538	.1585464	.2837806
NN	.1946141	.160456	.2707477
RF	.2037813	.1677594	.2216469
Xgb	.1966070	.1628026	.3065461

Table 5 Test Performances of Models

Model	RMSE	MAE	R^2
GLM	.2032341	.1671868	.2172193
SVM	.2051929	.1685865	.2087317
NN	.2029333	.1681839	.2235591
RF	.2143903	.1750374	.1282365
Xgb	.2071888	.1678343	.2066503

IV. RESULTS

In this part, the results are expressed for following regression models.

i. General Linear Model
ii. Support Vector Machine
iii. Artificial Neural Network
iv. Random Forests
v. XgBoost

In Table 5, RMSE, MAE and R^2 of each method is compared. According to the table, best prediction for the validation set is provided by GLM with value of .1671868. The General Linear Model performs better than others based on MAE. MAE is selected because there are lots of outliers in the dataset. MAE is less sensitive to outliers. According to the GLM, feature importance was given below the bar chart Figure 12. The most important feature is danceability and the least important feature is liveness.

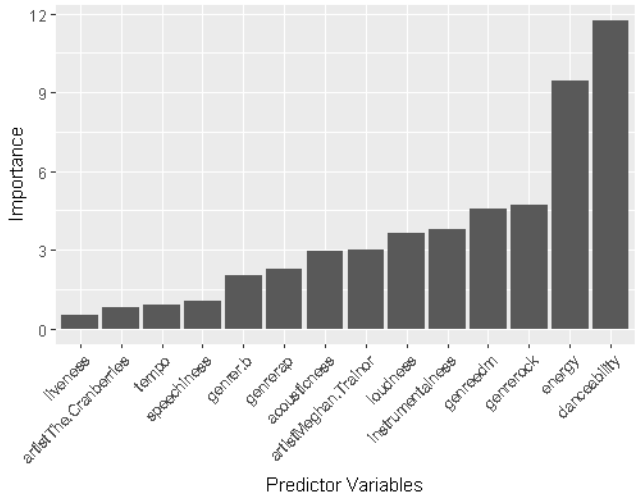

Figure 12 Feature Importance of GLM

V. CONCLUSION

In this work, exploratory data analysis, such as graphical techniques and descriptive statistics, is first conducted. Then, the outlier analysis, data cleaning, and imputing missing values are applied to improve the data quality. Finally, the valence value is tried to predict by using several methods. Their results are shown in the previous section. According to the best model, it is observed that danceability and energy are the most efficient factors in the valence value of songs.

VI. REFERENCES

[1] Laurier C, Herrera P (2007). Audio music mood classification using support vector machine. Music Information Retrieval Evaluation eXchange (MIREX) extended abstract.

[2] F. H. Rachman, R. Samo and C. Fatichah, "Song Emotion Detection Based on Arousal-Valence from Audio and Lyrics Using Rule Based Method," 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2019, pp. 1-5, doi: 10.1109/ICICoS48119.2019.8982519.