Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/38/20240602

Research on popularity of American pop singers' songs based on machine learning

Ao Feng
JKFZ Cambridge National School Nanchang, Jiangxi, 330000, China

2018903537@chd.edu.cn

Abstract. The U.S. market of music is the largest market in the globe with its huge influence spreading around, which enables it to be the dominant of the world music industry. The article is produced due to the prevailing music market in U.S. that has phenomenally influence around the world. Therefore, this article takes Taylor Swift as an example thanks to the significant influence power to calculate whether some factors such as acousticness that might directly affect the popularity of singer's songs for the purpose of formulating some market strategies by corresponding suggestions and advice. The research methods include 3 common mathematical model: linear regression, decision tree and random forest, which indicates that the year of release has the most contribution to the prediction with the value of importance of 0.549929. However other factors seem to have less relativity given the small values for the following 2 factors folklore and reputation with values of importance both below 0.2.

Keywords: Popular songs, linear regression, machine learning.

1. Introduction

Nowadays the U.S. market of music is the largest market in the globe with its huge influence spreading around, which enables it to be the dominant of the world music industry. With the advancement of technology, the means people can contact and appreciate music increases phenomenally. Therefore, the research of American music industry to formulate marketing strategies by projecting the degree of popularity of singers' songs becomes a focus.

Taylor Swift, serving as a representative singer among the whole American music circle, is the research subject owing to the astonishing number of fans and impact power. In accordance with a research conducted by a market research company named Morning Consult, there are 53% of adults living in the U.S. having fondness of the Midnights singer [1]. Besides, Caulfield stated that Taylor Swift reached the top level of the "Billboard" in 2023 after dominating the year on both "billboard" 200 albums and the "Billboard" Hot 100 [2]. MEAN PEERS illustrated that Taylor Swift owned 10 Grammys and 29 American Music Award [3]. It also declared that Taylor was the highest-paid female musician in the 2010s, during which there were a large number of gifted and popular musicians. If it were not enough to be a top female musician, the singer would have ranked the second on Forbes' highest-paid musicians in a decade, regardless of gender [4].

Plus, Dmitry Pastukhov announced that The Eras Tour, which is certainly Swift's biggest and most important tour to date and which is slowly shaping up to be one of the biggest music tours in the world, is gonging to last well into November 2024 beginning in March this year [5]. In addition, Marc Schneider

DOI: 10.54254/2753-8818/38/20240602

believed that in the 17 years since the debut, Taylor Swift's profound impact on the music industry extends beyond the unparalleled success in commercial aspect and the impact on everything, from artist rights to crushing traditional album release patterns to changing conversations about song rights and ownership. The singer is an advocate, a style icon, a marketing wiz, a prolific songwriter, a pusher of visual boundaries and a record-breaking road warrior [6]. Marc Schneider considered that people normally have scarce chance to get to the top music stardom like Swift, and it is even more scarce to have such a gigantic impact on the industry. Since the first album at age 16, Swift has presented supernatural gifts with the fans, inspiring dedication, leaving them waiting for every new song, album and merchandise to release and breathe, not to mention enthusiasm can collapse (and lead to Senate judicial hearings) [7]. Marc Schneider also claimed that in the course of Swift's reign, the singer used the numerous influence that seemed to change all aspects of the music industry- from helping to spark the vinyl revolution, to motivating record labels to change the way contracts were written, to changing the way concert tickets were bought and sold. When it comes to the music itself, the singer navigates the changing sound savvy and, to a large extent, elegant, retaining loyal young fans growing up with the and expanding to a new audience through the embrace from pop traps to folk hip hop [8].

To sum up, the research on popularity of American pop singers' songs has attracted countless scholars. This article will base on decision tree, random forest as well as linear regression to predict and analyse the factors that will affect degree of popularity of singers' songs and provide corresponding suggestions and advise to formulate marketing strategies.

2. Methodology

2.1. Data source and description

The dataset comes from Kaggle. The shown column chart is presented by giving the specific information of the every perimeter concerning names, instrumentalness and track numbers and acousticness, and so forth of every song instead of illustrating the property of acoustics alone, with the given interval of the number from 1 to 0 (Table 1).

Table 1. Basic information of dataset.

Variable	Meaning
acousticness	The acousticness measures of the audio, representing the acoustic / non-acoustic properties of the repertoire, ranges from 0 to 1, and the higher the value indicates a more pure acoustic repertoire
'danceability	Danceability, indicates the degree to which the song is suitable for dance, the value range from 0 to 1, the higher the value, the more suitable for dance
energy	Energy, indicates the energy and vitality of the repertoire, ranges from 0 to 1, with higher values indicating higher energy
instumentalness	Instrumentalness, indicates whether the track contains human voice, take the value range of 0 to 1, the higher the value, the more pure the instrument track
liveness	Liveness indicates the degree of live performance of the repertoire. The value range is 0 to 1. The higher the value, the more the sense of live performance
loudness	loudness indicates the overall volume of the track, in decibels (dB), the higher the value, the higher the volume
speechiness	speechiness indicates the degree to which the repertoire contains spoken English, taking the value range from 0 to 1. The higher the value, the better the oral English characteristics
tempo	The tempo of the track indicates the rhythm speed of the track (beats per minute)
valence	valence indicates the degree of positive emotion of the repertoire, ranges from 0 to 1, and a higher value indicates a more positive emotion
duration_ms	The duration of the repertoire, as measured in milliseconds

DOI: 10.54254/2753-8818/38/20240602

2.2. Method introduction

Linear regression model is widely used due to its simplicity, especially commonly in the measurement of relationships between two variables which depending on each other linearly. Decision tree is structure that resembles a tree similar to a flowchart, where each internal node represents the features, branches represent the rules, and leaf nodes represent the results of the algorithm (Figure 1). It is a general and supervised machine learning algorithm for both classification and regression problems. Root Node, Decision node, Leaf node and Splitting represents the staring point which indicates the complete dataset, a node that represents the choice concerning an input feature and connects to leaf nodes or other nodes, a node that represents a class or numerical value without any child notes and a process that splits a node into sub-notes by split criteria or selected features respectively [9, 10].

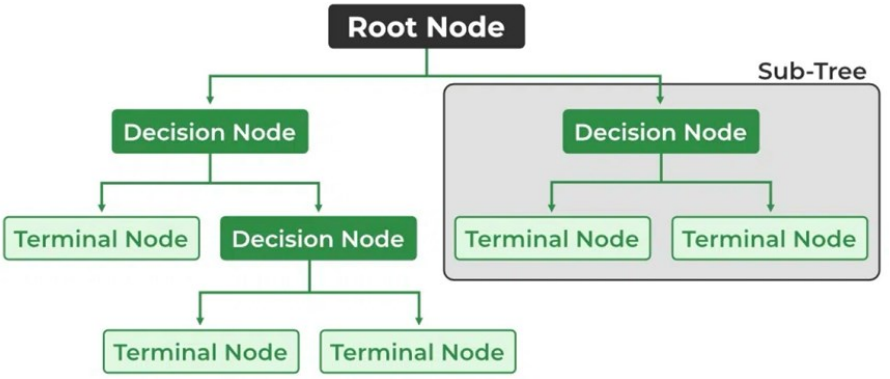

Figure 1. Machine learning process [8].

Entropy: This is a means of measuring the degree of randomness or uncertainty in the dataset and it is based on the distribution of class labels in the dataset in classifying cases. The entropy for a subset of the original dataset with k number of classes for the ith node can be defined as:

$$p(k) = \frac{1}{n}\sum I(y = k) \tag{1}$$

Information Gain: Information Gain means the reduction in entropy after the split of dataset. It is also named Entropy Reduction. Building a decision tree is all about discovering attributes that return the highest data gain (Figure 2).

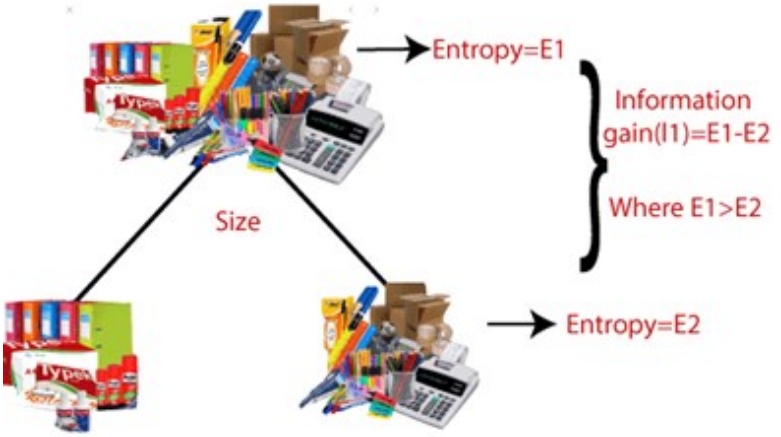

Figure 2. Entropy and Information Gain process [8].

DOI: 10.54254/2753-8818/38/20240602

The decision tree is useful due to its ability to fully analyse the consequence of one decision and to provide a frame to measure the outcome values as well as the probability to accomplish. Random Forest algorithm is a common means in Machine Learning. It works by creating a number of Decision Trees during the training phase. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. The algorithm aggregates the outcomes of all trees, either by voting or by averaging in the process of anticipating. This collaborative decision-making process provides an example stable and precise results owing primarily to multiple trees with their insights. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data and decrease overfitting, providing reliable projections in distinctive environments as well [10]. The final result of this system is drown by ordinary majority vote, the decision function is:

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \tag{2}$$

where $H(x)$ is combination of classification model, hi is a single decision tree model, Y is the output variable, $I(\cdot)$ is the indicator function. For a given input variable, each tree has right to vote to select the best classification result. In Random Forest, margin function is used to measure the average number of votes at $\boldsymbol{X}, Y$ for the right class exceeds that for the wrong class, margin function is defined as:

$$mg(X,Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \tag{3}$$

The larger the margin value, the higher accuracy of the classification prediction, and the more confidence in classification.

3. Results and Discussion

3.1. Descriptive analysis

The popularity interval of the first column, ranging from 33 to 37, corresponds 39 songs and it should be noted that intervals, from 37 to 43, from 55 to 60, from 60 to 66, from 82 to 88, from 88 to 94, from 94 to 98, are all below the initially mentioned first column. The remaining five groups without being mentioned in the range from 33 to 98 are all above 39 songs and the top 3 groups are fluctuating in the song number from 67 to 104 (Figure 3).

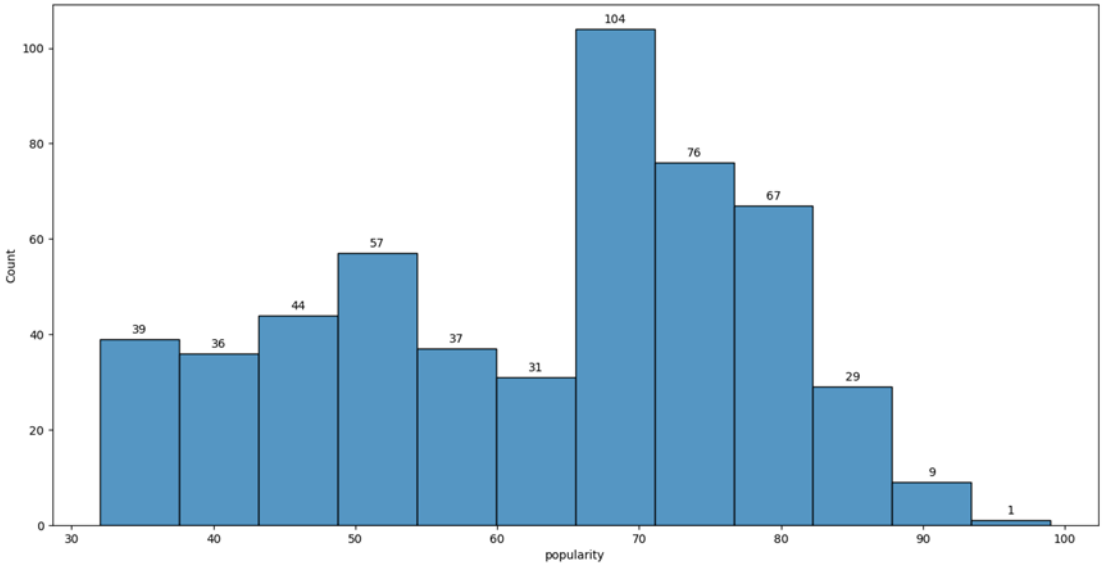

Figure 3. Distribution of popularity.

DOI: 10.54254/2753-8818/38/20240602

Shown are the three column charts, providing the details of the release dates of each song by setting the interval in days, months and years. By the figures shown in the first column group regarding day of songs released, it can be clearly seen that the first week witnesses the lowest quantity of no more than 40 with the next two days until 9th saw a nearly double time growth that reaches above 70. Before the days preceding 25, the entire song number are between 12 and 58, because the date of 25th experiences a summit of 82 and for the subsequent two days, 26th and 27th, the released songs all transcend those in the first week. But 28th just arrived at the lowest of no more than 10. Coming to the second chart concerning month, both October and November are in the highest range between 150 and 210, given that in the rest months all the songs released show a group of weak numbers below 30. As for the year release-related song numbers given in the third chart, 2023 goes for the pinnacle value of more than 80, which is closely followed by 2020 and 2021, since the years between 2006 and 2022, despite the mentioned two years of 2020 and 2021, are nearly half of the top value happening in 2023 (Figure 4).

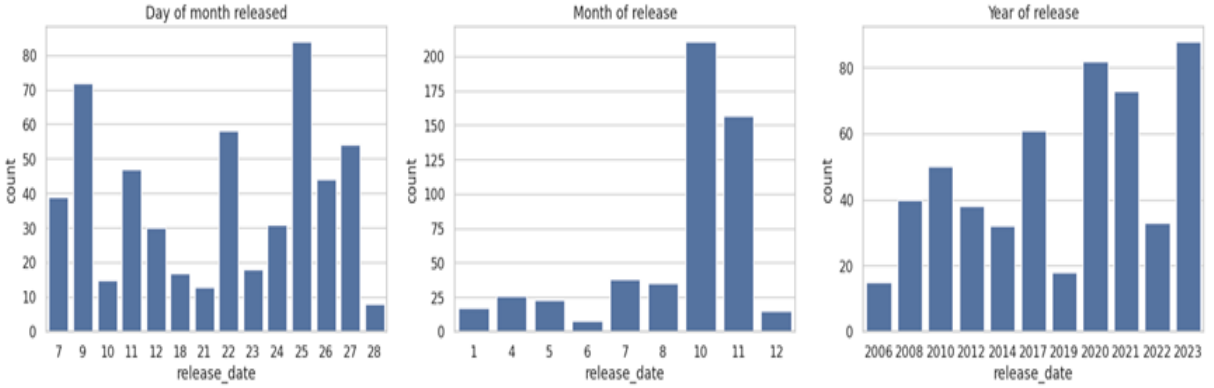

Figure 4. Distribution of other variables.

3.2. Correlation analysis

The heat map elaborately indicates the varying correlations between track number, acounsticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, popularity and duration_ms. According to the deepness of the colors belonging to each perimeter that is shown in the chart, the deeper lightness of the color suggests a higher relevance of the above-mentioned features, for instance, red color is associated with a positive correlation while blue color is bound with a negative one oppositely. For the further information that is well manifested in the very heat map showing in detail with different colors, loudness and energy belongs to a synergy of the highest strength in statistical relation but with regard to loudness as well as acousticness, both are in a negative correlation and it is, so to speak, the same relationship as it is indicated between energy and acousticness given in a negative (Figure 5).

DOI: 10.54254/2753-8818/38/20240602

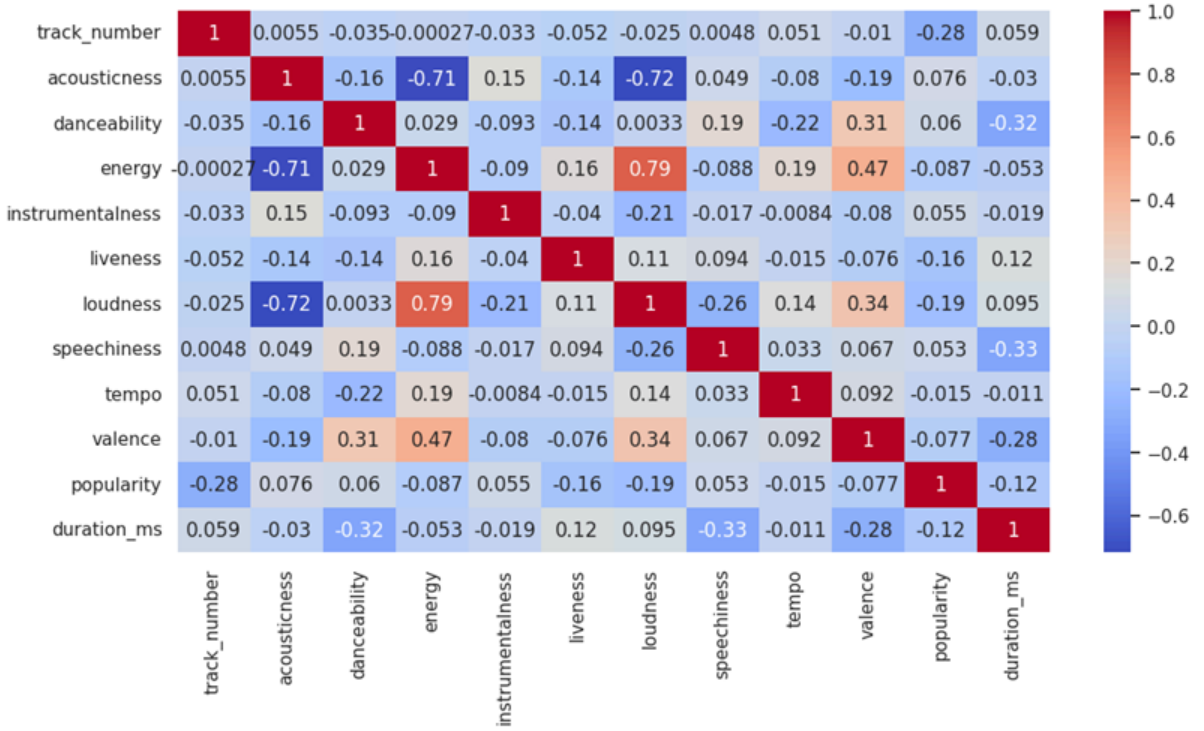

Figure 5. Correlation results.

3.3. Model results

The root mean square error (RMSE) has been widely used as a standard statistical metric of measuring model performance in meteorology, air quality, and climate research studies. The RMSE is calculated for the data set as: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$. The underlying assumption when presenting the RMSE is that the errors are unprejudiced and obey a normal distribution, which can offer a picture of the error distribution (Table 2).

Table 2. Model evaluation.

	linear regression	decision tree	random forest
MSE	72.275	87.057	71.500
RMSE	8.501	9.330	8.456
R^2	0.668	0.600	0.672

The table 2 shown is a comparison between 3 methods of 3 modals. It should be noticed that Random Forest should be the best way of measuring due to its phenomenal data. Linear Regression is acceptable and Decision tree has the worse outcomes (Figure 6).

DOI: 10.54254/2753-8818/38/20240602

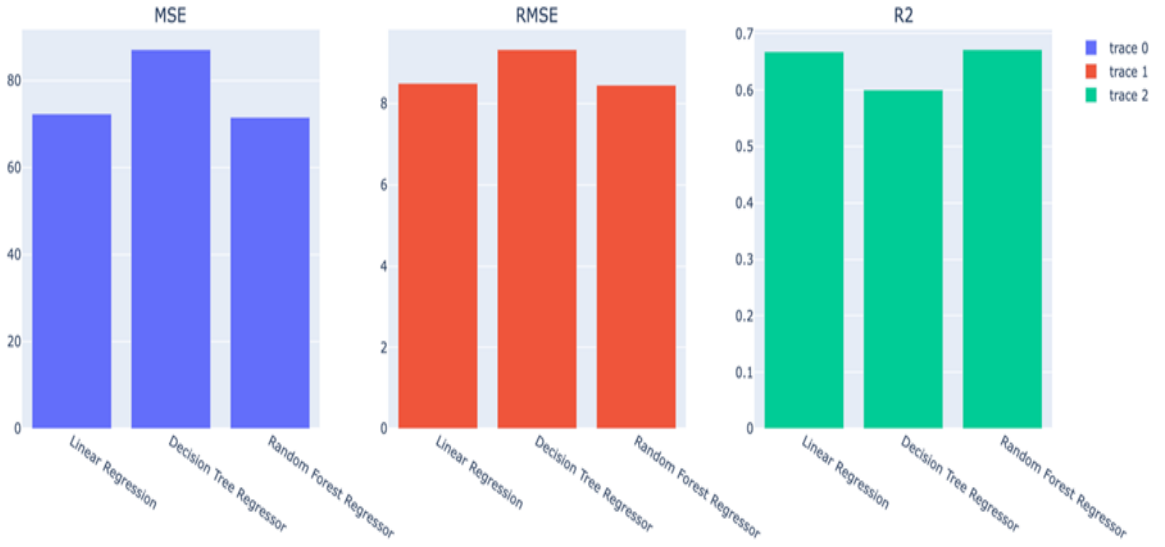

Figure 6. Model evaluation results.

4. Conclusion

The significance of year of release is the highest as a result of its value of importance of 0.549929, which indicates that the year of release contributes the most in the modal of prediction. Despite that reputation ranks the second at the importance list and which can suggest high relevance to some extent, the values are too low with only 0.122189. It is the same with folklore as well considering its lower relevance (with the values of importance of 0.084982) but the third ranking positions. Therefore, it is a shame that both reputation and folklore fail to refer due to the extremely low values.

References

[1] Xiong X B, Zhou G, Huang Y Z and Ma J 2012 Research on the Prediction Technology of Sina Weibo Topic Popularity. Journal of University of Information Engineering, 13(4), 7.
[2] Chen C Y, Zhang Y, Chang B and Lv J L 2016 Online TV drama popularity prediction based on ARMA model. Computer Science and Exploration, 8.
[3] Cao Q, Shen H W, Gao J H and Cheng X Q 2021 A review of research on popularity prediction based on deep learning. Chinese Journal of Information Technology.
[4] Zhu H L, Yun X C and Han Z S 2018 A Weibo popularity prediction method based on propagation acceleration. Computer Research and Development, 55(6), 12.
[5] Feng Y Q 2024 Research on personalized media recommendation algorithms based on popularity prediction. Doctoral dissertation, Ocean University of China.
[6] Hu J R, et al. 2023 Research on Weibo Popularity Prediction Algorithm Based on Propagation Characteristics. Computer and Digital Engineering, 51(4), 763-768.
[7] Xie X Q 2020 Research on Information Popularity Prediction for Social Networks. Doctor dissertation, Chongqing University of Posts and Telecommunications.
[8] Li X P 2016 Predicting the popularity of news client comments. Doctoral dispersion, University of Chinese Academy of Sciences.
[9] Shi L, et al. 2006 A ppm prediction model based on web object popularity. Small Micro Computer Systems, 7, 228-232.
[10] Wang X M, et al. 2019 Facebook message popularity prediction model based on connection strength. Journal of Communications, 40(10), 9.