Method

The object of this project is to find the which factors influence song’s popularity. From the experience, several reasons including popularity of singer, tunes, type of lyric, length of the sang, can lead to a song be popular. To explore the most influential variables of the song’s popularity, this project use data from Spotify. We used regression model, logistic model, decision tree and Knn to describe and simulate the relationship between popularity and other variables including acousticness danceability duration\_ms energy instrumentalness loudness Liveness mode valence tempo and speechiness. Those variables describe a song in 13 aspects.

The first method we use is linear regression model. The linear regression model is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. And linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? In our project, we see popularity as response variable and see other 13 variables as explanatory.

(1)

This formular this the base model of linear regression model, and the result of this regression model is in table 1. Following the most important criterion (John, R. C. S. ,1983, p424)

(2)

Then we got which means the model is not following linear regression, then we delete variables, key and tempo of which the P-value is bigger than 0.05, to adjust the model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | coef | std err | t | P>|t| | 0.025 | 0.975 |
| Intercept | 34.7007 | 0.498 | 69.634 | 0.000 | 33.724 | 35.677 |
| key | -0.0162 | 0.014 | -1.139 | 0.255 | -0.044 | 0.012 |
| mode | -0.3397 | 0.106 | -3.211 | 0.001 | -0.547 | -0.132 |
| danceability | 5.9222 | 0.345 | 17.161 | 0.000 | 5.246 | 6.599 |
| duration\_ms | -1.545e-06 | 4.12e-07 | -3.747 | 0.000 | -2.35e-06 | -7.37e-07 |
| acousticness | 1.5730 | 0.217 | 7.251 | 0.000 | 1.148 | 1.998 |
| energy | -4.2910 | 0.372 | 11.520 | 0.000 | -5.021 | -3.561 |
| liveness | -2.1845 | 0.321 | -6.802 | 0.000 | -2.814 | -1.555 |
| loudness | 0.6537 | 0.014 | 45.711 | 0.000 | 0.626 | 0.682 |
| speechiness | -6.1727 | 0.435 | 14.190 | 0.000 | -7.025 | -5.320 |
| tempo | -0.0017 | 0.002 | -0.998 | 0.318 | -0.005 | 0.002 |
| valence | -4.8436 | 0.231 | 20.976 | 0.000 | -5.296 | -4.391 |
| instrumentalness | -6.1746 | 0.169 | 36.476 | 0.000 | -6.506 | -5.843 |

**Table 1**

Then of the modify model is same as before, which is , and all the P-value are less than 0.05. The both result show the linear regression is not a good model to predict the popularity of song.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | coef | std err | t | P>|t| | 0.025 | 0.975 |
| Intercept | 34.3889 | 0.439 | 78.301 | 0.000 | 33.528 | 35.250 |
| mode | -0.3198 | 0.104 | -3.073 | 0.002 | -0.524 | -0.116 |
| danceability | 5.9455 | 0.345 | 17.257 | 0.000 | 5.270 | 6.621 |
| duration\_ms | -1.548e-06 | 4.12e-07 | -3.753 | 0.000 | -2.36e-06 | -7.39e-07 |
| acousticness | 1.5871 | 0.216 | 7.339 | 0.000 | 1.163 | 2.011 |
| energy | -4.3124 | 0.372 | 11.588 | 0.000 | -5.042 | -3.583 |
| liveness | -2.1661 | 0.321 | -6.757 | 0.000 | -2.795 | -1.538 |
| loudness | 0.6528 | 0.014 | 45.772 | 0.000 | 0.625 | 0.681 |
| speechiness | -6.1878 | 0.435 | 14.240 | 0.000 | -7.039 | -5.336 |
| valence | -4.8645 | 0.230 | 21.105 | 0.000 | -5.316 | -4.413 |
| instrumentalness | -6.1752 | 0.169 | 36.498 | 0.000 | -6.507 | -5.844 |

**Table 2**

After that, the split the whole data into two parts and the parameter(train-size) = 0.08, which means the 80% of whole data is train data size and the 20%of whole data is test data size. The which means the result also not good, getting same conclusion.

Then we tried general linear model, logistic model, to simulate the data. Data are unbalanced on Y if y = 1 occurs relatively few times or if y = 0 occurs relatively few times. This limits the number of predictors for which effects can be estimated precisely. (Agresti, A. 2018, P 138), Before that, we have already changed the popularity column from scalar into catalogue variable. We settled the popularity as 0 if number of popularity smaller than 24 and settled the popularity as 1 if number of popularity bigger than 24. After that we got a balance data of y which can got meaningful result.

(3)

(4)

Our logistic model build bases the formular (3) and (4). To simulate the model, we split data as same as before. Then we got accuracy just 5%, that means the logistic model is not fit for the data, we can not use logistic model to predict song’s popularity.

Decision tree learning or induction of decision trees is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves) Then we use decision tree to simulate the data which used same way to split data. After this supervise learning, the accuracy of test data by using decision tree is just 51.29%, Then we considered feature importance of the model. ‘loudness’ is the most influential variable compared with the other variables on the popularity. Even though all the variables are not very influential.

图表, 条形图

描述已自动生成

**Figure 1**

Reference:

John, R. C. S. (1983). Applied linear regression models.

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.