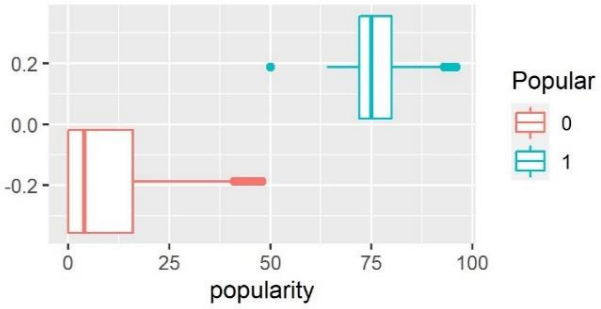
Spotify Tracks Popularity Analysis

Gokcen Buyukbas, Adnan Cihan Cakar S632 – Final Project 05-06-2021

We analyzed the Spotify tracks data [1] consisting of 174389 songs with 19 features.

Research Questions:

1. What song features play a significant role in a song's popularity?
2. Do the features affecting the popularity of tracks change in years?



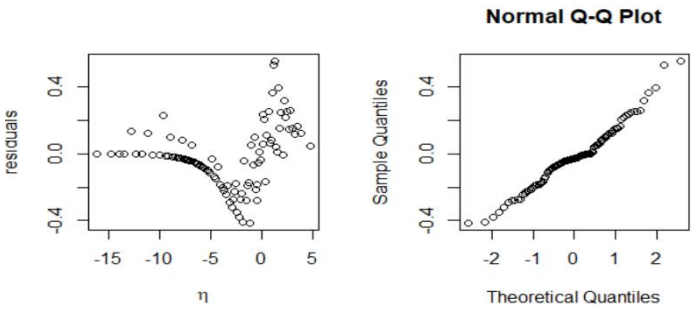
To answer these questions, we first prepared the data for the analysis:
- removed the features unique to songs such as *id*, *name* of the song, etc,
- added a new variable, *season,* based on the release date,
- created a binary response for popularity, *binaryPop*, 0 if popularity< 50, 1 otherwise.

Methods:

Logistic regression: We did forward and backward propagation for variable selection. The model recommended by both were the same:

binaryPop ~ album_type + acousticness + danceability + duration_ms + energy + instrumentalness + liveness + loudness + mode + valence + season.

The *qqplot* seemed to be almost normal, however, the residuals are not uniformly distributed. Furthermore, we did a Hosmer-Lemeshow test and rejected this model because the p-value was practically zero.



Binomial regression: We fitted a binomial regression model with forward and backward selection. Again, the diagnostics and Pearson-chi square test indicated that this model was not a good fit either.
We fitted other models too such as quasibinomial, beta regression, Poisson, zero-inflated model etc. None of them produced a decent model that passed the deviance or chi-square test.

Random Effects: We investigated random effects of the features: artist, album, and album type, considering these as nested effects. First, we selected variables using Kenward-Roger test for fixed effects, and exactRLRT for the random effects of features. We concluded that the album type within albums is not significant, yet the albums within artists and the artist itself are significant for the model.

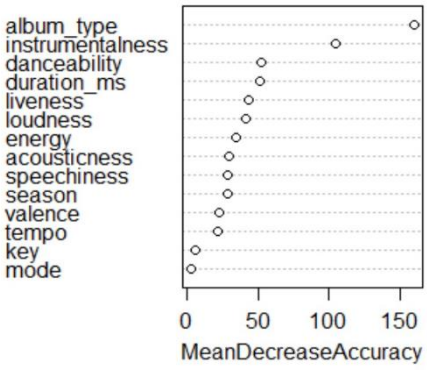
```
Random effects:
 Groups       Name        Variance Std.Dev.
 artist:album (Intercept)  74.49    8.631
 artist       (Intercept) 569.92   23.873
 Residual                  10.04    3.169
Number of obs: 3666, groups:  artist:album, 2077; artist, 1502
```

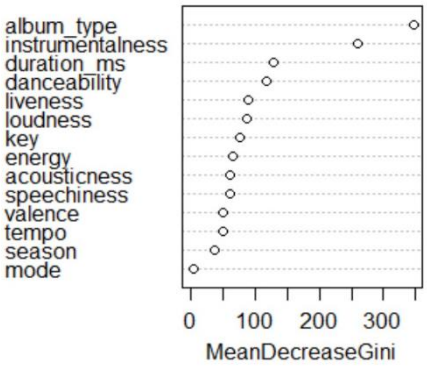
Random forest:

We investigated the importance of features of tracks for the popularity. We split the data into training and test sets, with 20% allocated for testing.

	Accuracy	Sensitivity	Specificity	F1
RandomGuess	0.49	0.30	0.67	0.57
RandomForest	0.79	0.77	0.79	0.86

We first trained a random forest algorithm with the training data using popularity as a binary response and compared the performance to random guessing. Random forest performed much better with an F1 score of 0.86. The most important feature to predict the popularity was *season followed by duration, instrumentalness,* and *loudness*.

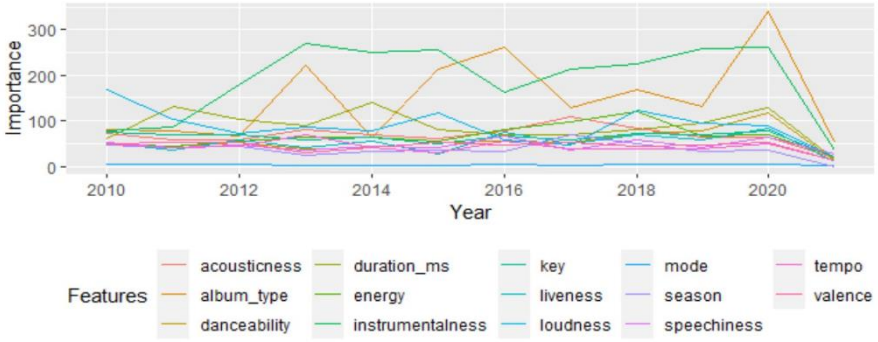




We observed that the higher the number of end nodes is, the better the F1 score is. However, increasing the number of variables used in the trees did not improve the accuracy much.

We continued our analysis with regression tree, using continuous *popularity* variable with random forest, and we reached at similar results for feature importance in slightly different orders. In the regression tree model, *album type* and *instrumentalness* were two most important features for *popularity*, followed by *duration_ms, and season*.

Finally, we trained a random forest model for each year between 2010 to 2021 and observed that even though the topmost important features have alternated with each other, there wasn't a drastic change in the list of top 5 most important features over the years.



Conclusion:

1. The song features *album type, duration, instrumentalness,* and *liveness* were the most important features to predict popularity.
2. *Instrumentalness* and *album type* were the two most important features even though their order alternated over the years. The list of top five most important features has not changed much.

References:

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

[2] Extending the Linear Model with R, Julian Faraway, Second edition, Boca Raton : CRC Press, Taylor & Francis Group, [2016] and ©2016

[3] https://developer.spotify.com/documentation/web-api