PREDICTING SONG POPULARITY ON Spotify

CS 6375 Final Project

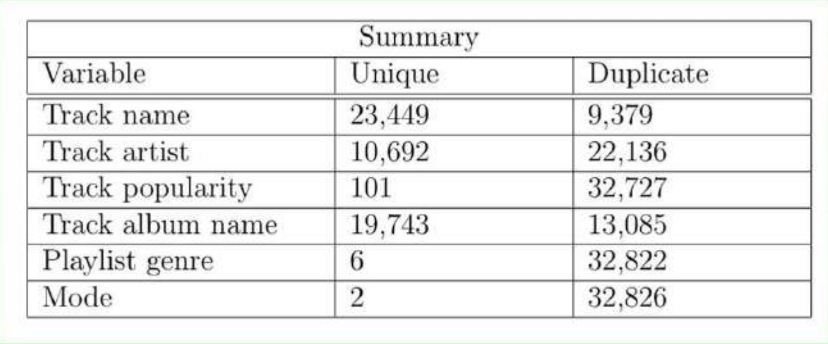
by Lakshmipriya Narayanan

ABOUT THE DATA

- Objective is to predict song popularity using various machine learning algorithms
- Dataset source: Kaggle 30000 Spotify Songs
- Contains over 32,828 songs from the 60s to 2020s. (1957 - 2020)
- There are 23 song attributes like energy, danceability, loudness, key, speechiness, etc.
- Response variable is 'track_popularity' measured on a scale of 0 – 100.
- Features we will be using to predict response are danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveliness, valence and, tempo.
- Future applications

DATA PRE-PROCESSING

- Removal of irrelevant attributes like ID's of songs and albums, mode (categorical) and duration of song.
- Dropped missing values. 5 missing values in all.
- Analyzed unique and duplicate value counts to avoid inaccuracy in analysis and model performances.
- Centering and scaling to remedy presence of multicollinearity.
- Data was mostly cleaned in terms of massive rectification of observations.

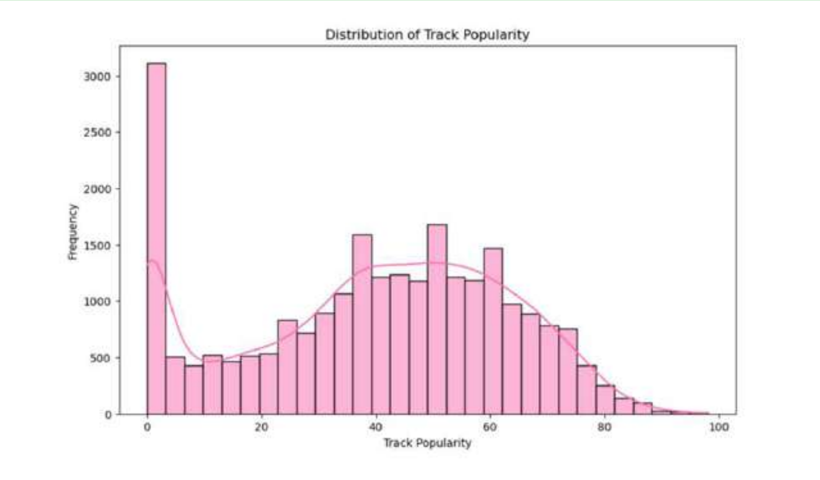


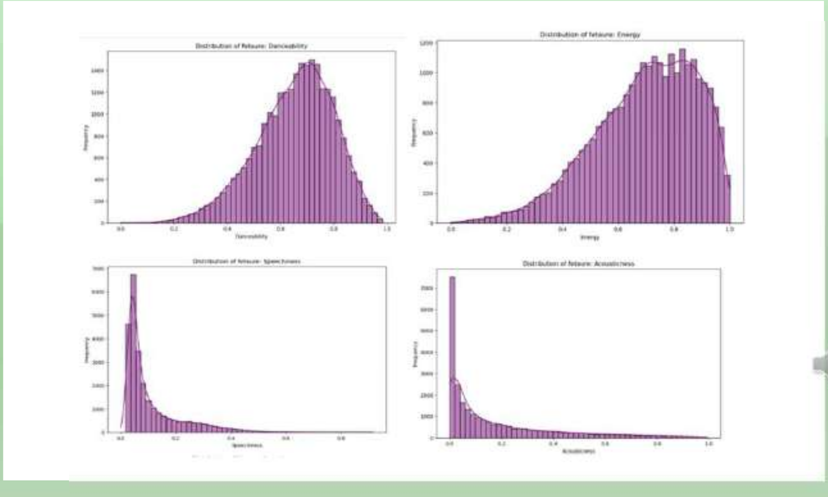
Summary		
Variable	Unique	Duplicate
Track name	23,449	9,379
Track artist	10,692	22,136
Track popularity	101	32,727
Track album name	19,743	13,085
Playlist genre	6	32,822
Mode	2	32,826

EXPLORATORY DATA ANALYSIS

1) Analysis of feature distributions:

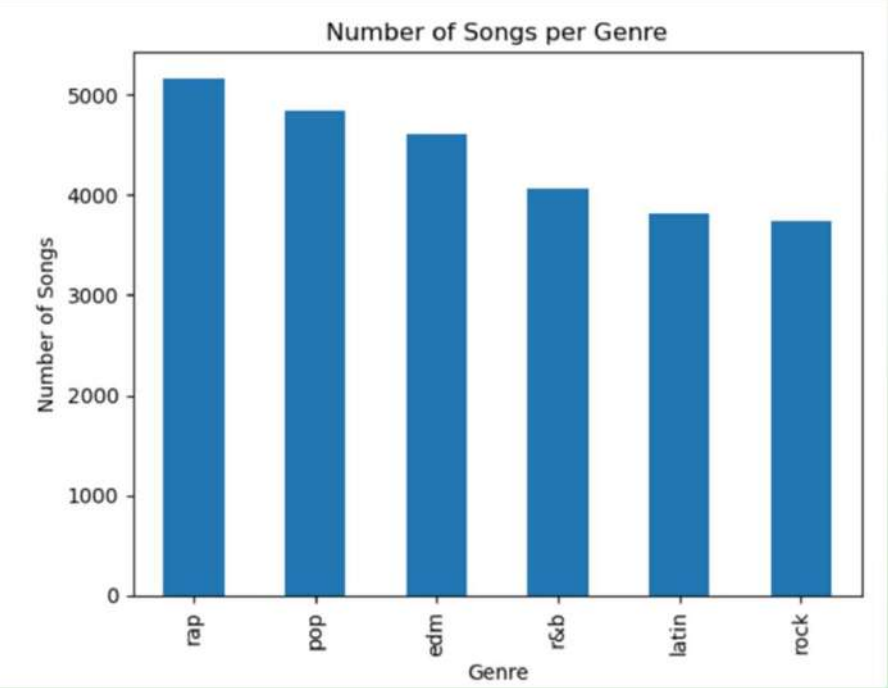
- Response follows mostly normal distribution whereas other numerical features are skewed and some similar to normality.

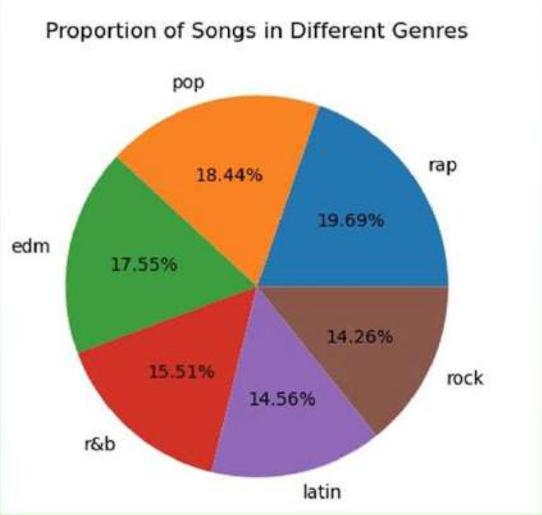


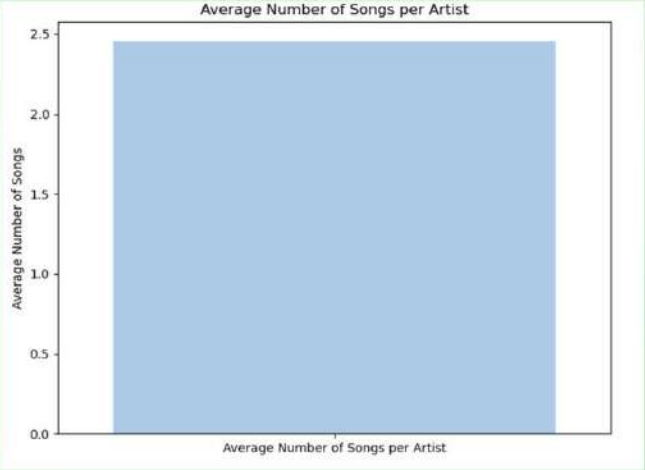


2. Popularity

- With respect to song genre

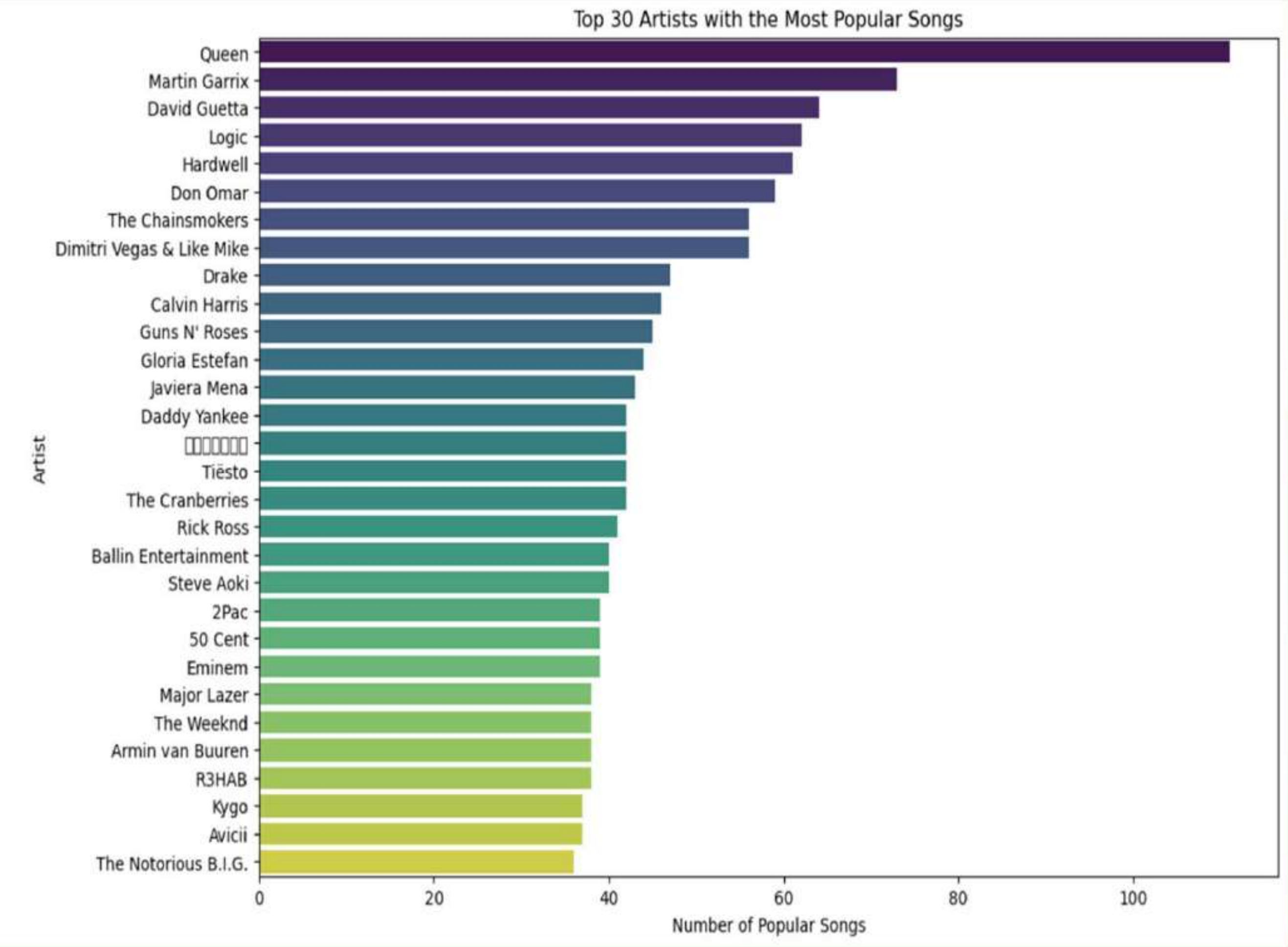






Most popular genre is rap with 19.69 % of songs and the least popular genre is rock with 14.26 %

- With respect to artists

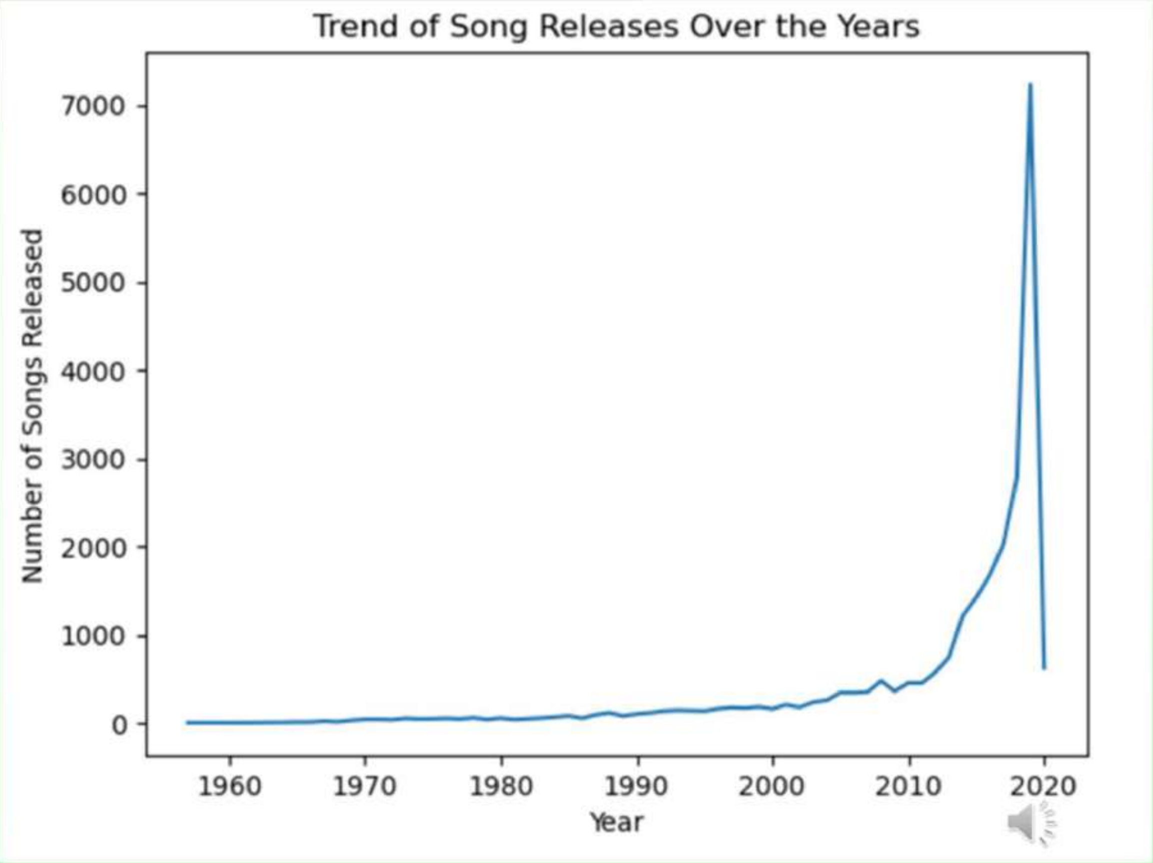


Most Popular

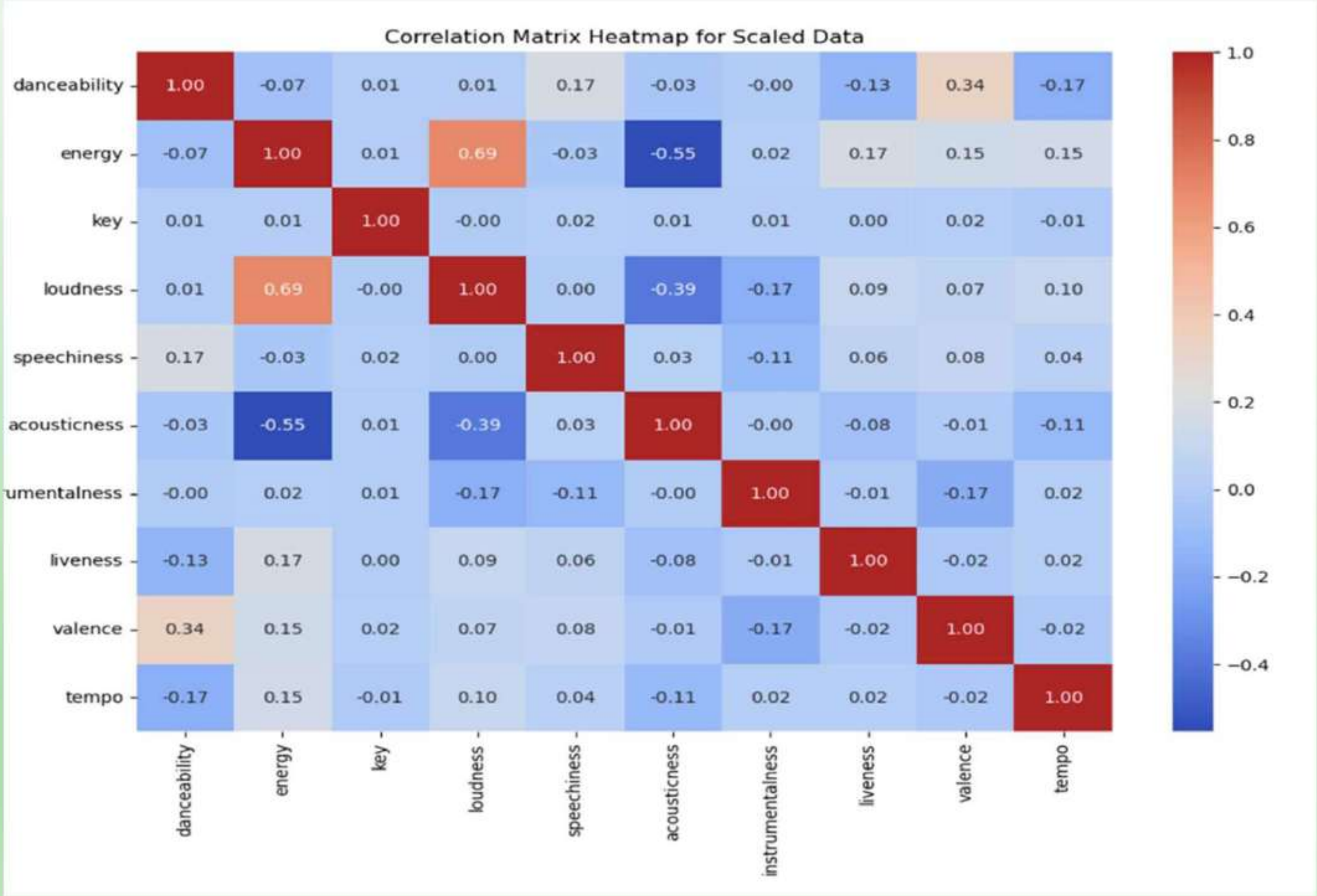
Least popular

This Photo by Unknown Author is licensed under CC BY-ND

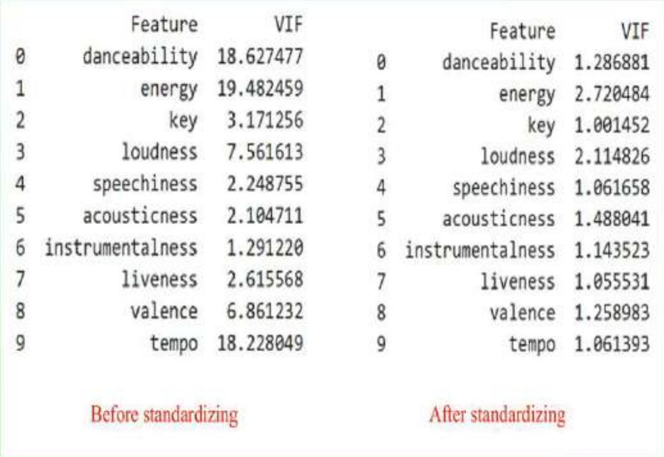
An interesting observation



3. Multicollinearity and correlation between feature variables



	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo
danceability	1.00	-0.07	0.01	0.01	0.17	-0.03	-0.00	-0.13	0.34	-0.17
energy	-0.07	1.00	0.01	0.69	-0.03	-0.55	0.02	0.17	0.15	0.15
key	0.01	0.01	1.00	-0.00	0.02	0.01	0.01	0.00	0.02	-0.01
loudness	0.01	0.69	-0.00	1.00	0.00	-0.39	-0.17	0.09	0.07	0.10
speechiness	0.17	-0.03	0.02	0.00	1.00	0.03	-0.11	0.06	0.08	0.04
acousticness	-0.03	-0.55	0.01	-0.39	0.03	1.00	-0.00	-0.08	-0.01	-0.11
umentalness	-0.00	0.02	0.01	-0.17	-0.11	-0.00	1.00	-0.01	-0.17	0.02
liveness	-0.13	0.17	0.00	0.09	0.06	-0.08	-0.01	1.00	-0.02	0.02
valence	0.34	0.15	0.02	0.07	0.08	-0.01	-0.17	-0.02	1.00	-0.02
tempo	-0.17	0.15	-0.01	0.10	0.04	-0.11	0.02	0.02	-0.02	1.00



```
             Feature        VIF
0       danceability  18.627477
1             energy  19.482459
2                key   3.171256
3           loudness   7.561613
4        speechiness   2.248755
5       acousticness   2.104711
6   instrumentalness   1.291220
7           liveness   2.615568
8            valence   6.861232
9              tempo  18.228049
```

Before standardizing

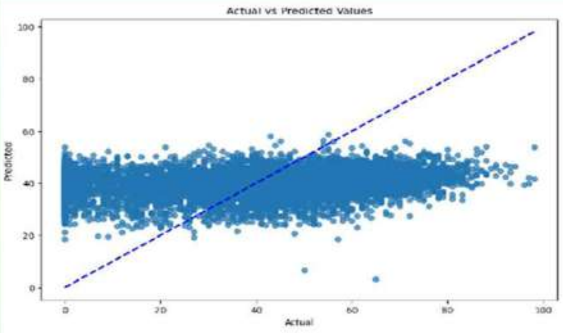
```
             Feature       VIF
0       danceability  1.286881
1             energy  2.720484
2                key  1.001452
3           loudness  2.114826
4        speechiness  1.061658
5       acousticness  1.488041
6   instrumentalness  1.143523
7           liveness  1.055531
8            valence  1.258983
9              tempo  1.061393
```

After standardizing

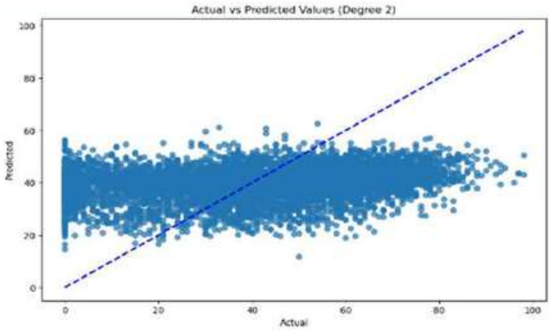
MODEL FITTING WITH MULTIPLE LINEAR REGRESSION

Without polynomial features

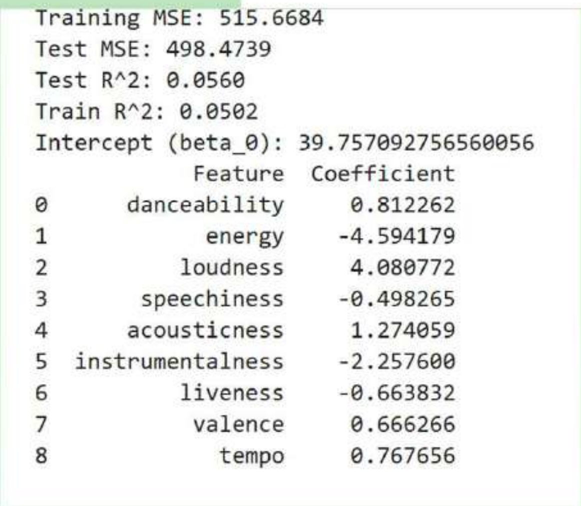
With polynomial features of degree 2



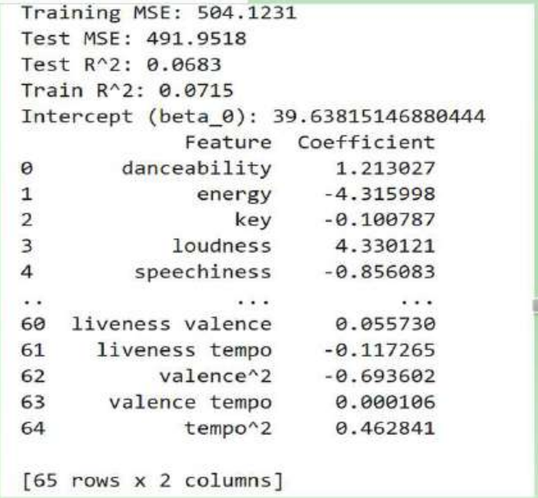
Multiple linear regression



Linear regression with polynomial features

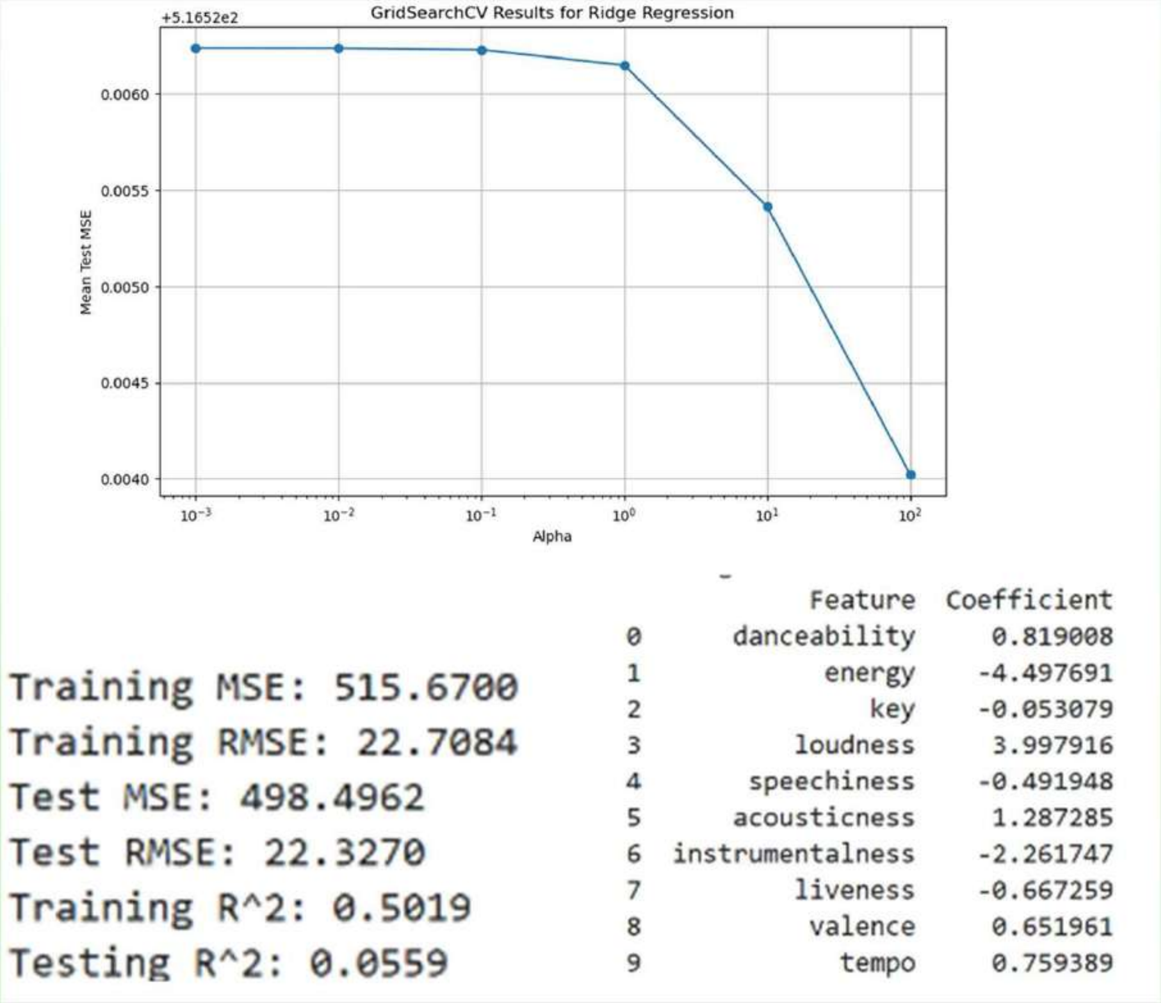


```
Training MSE: 515.6684
Test MSE: 498.4739
Test R^2: 0.0560
Train R^2: 0.0502
Intercept (beta_0): 39.757092756560056
            Feature  Coefficient
0      danceability     0.812262
1            energy    -4.594179
2          loudness     4.080772
3       speechiness    -0.498265
4      acousticness     1.274059
5  instrumentalness    -2.257600
6          liveness    -0.663832
7           valence     0.666266
8             tempo     0.767656
```



```
Training MSE: 504.1231
Test MSE: 491.9518
Test R^2: 0.0683
Train R^2: 0.0715
Intercept (beta_0): 39.63815146880444
            Feature  Coefficient
0      danceability     1.213027
1            energy    -4.315998
2               key    -0.100787
3          loudness     4.330121
4       speechiness    -0.856083
..              ...          ...
60  liveness valence     0.055730
61    liveness tempo    -0.117265
62         valence^2    -0.693602
63     valence tempo     0.000106
64           tempo^2     0.462841

[65 rows x 2 columns]
```

MODEL FITTING WITH RIDGE REGRESSION



```
Training MSE: 515.6700
Training RMSE: 22.7084
Test MSE: 498.4962
Test RMSE: 22.3270
Training R^2: 0.5019
Testing R^2: 0.0559
```

```
            Feature  Coefficient
0      danceability     0.819008
1            energy    -4.497691
2               key    -0.053079
3          loudness     3.997916
4       speechiness    -0.491948
5      acousticness     1.287285
6  instrumentalness    -2.261747
7          liveness    -0.667259
8           valence     0.651961
9             tempo     0.759389
```

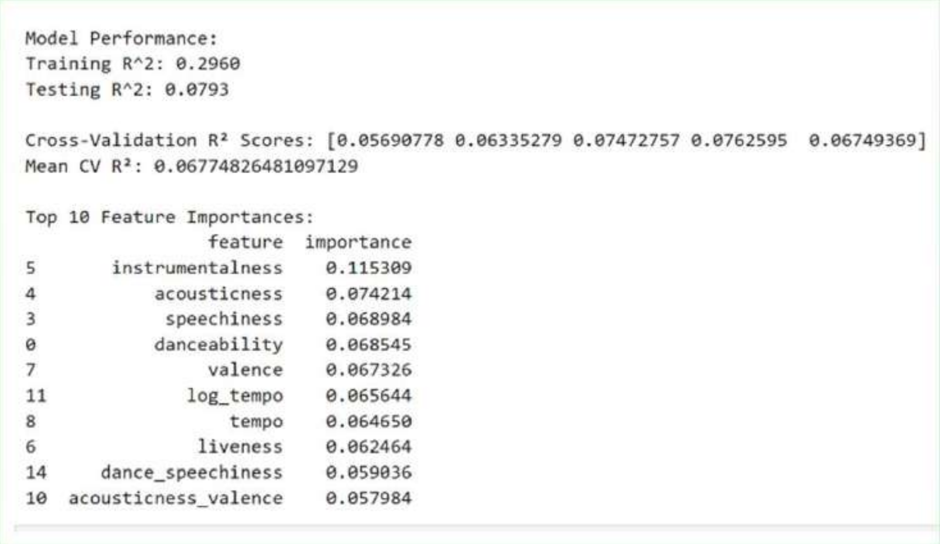
MODEL FITTING WITH RANDOM FOREST

- Took into consideration interaction terms and transformation.
- Hyperparameters used to obtain a Test R^2 of 0.076 and training R^2 of 0.75 are as follows:
  - ❑Number of trees = 150
  - ❑Maximum depth of each tree in the forest = 20
  - ❑Minimum number of samples required to split an internal node = 5
  - ❑Minimum number of leaf nodes of each subtree = 2 (binary tree)
- Severely overfit
- Hyperparameter tuning not done because of technical issues.

MODEL FITTING WITH GRADIENT BOOSTING

WITHOUT HYPERPARAMETER TUNING

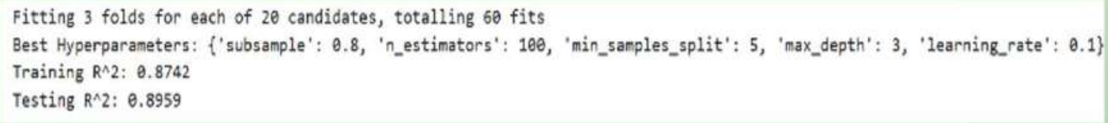
- Did some feature engineering to obtain various interaction terms and more transformations.



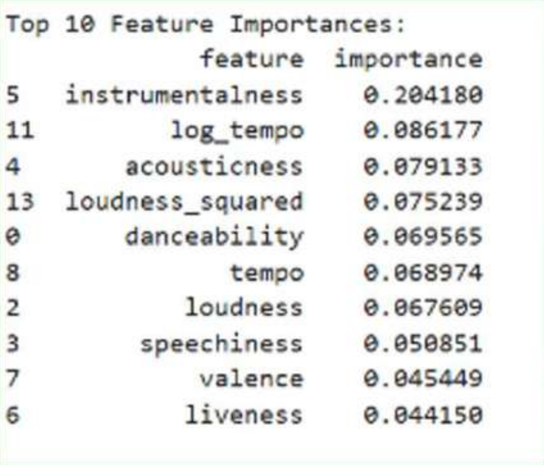
```
Model Performance:
Training R^2: 0.2960
Testing R^2: 0.0793

Cross-Validation R² Scores: [0.05690778 0.06335279 0.07472757 0.0762595  0.06749369]
Mean CV R²: 0.06774826481097129

Top 10 Feature Importances:
                 feature  importance
5       instrumentalness    0.115309
4           acousticness    0.074214
3            speechiness    0.068984
0           danceability    0.068545
7                valence    0.067326
11             log_tempo    0.065644
8                  tempo    0.064650
6               liveness    0.062464
14     dance_speechiness    0.059036
10  acousticness_valence    0.057984
```

WITH HYPERPARAMETER TUNING

- Used the same feature engineering used in model with hyperparameter tuning.

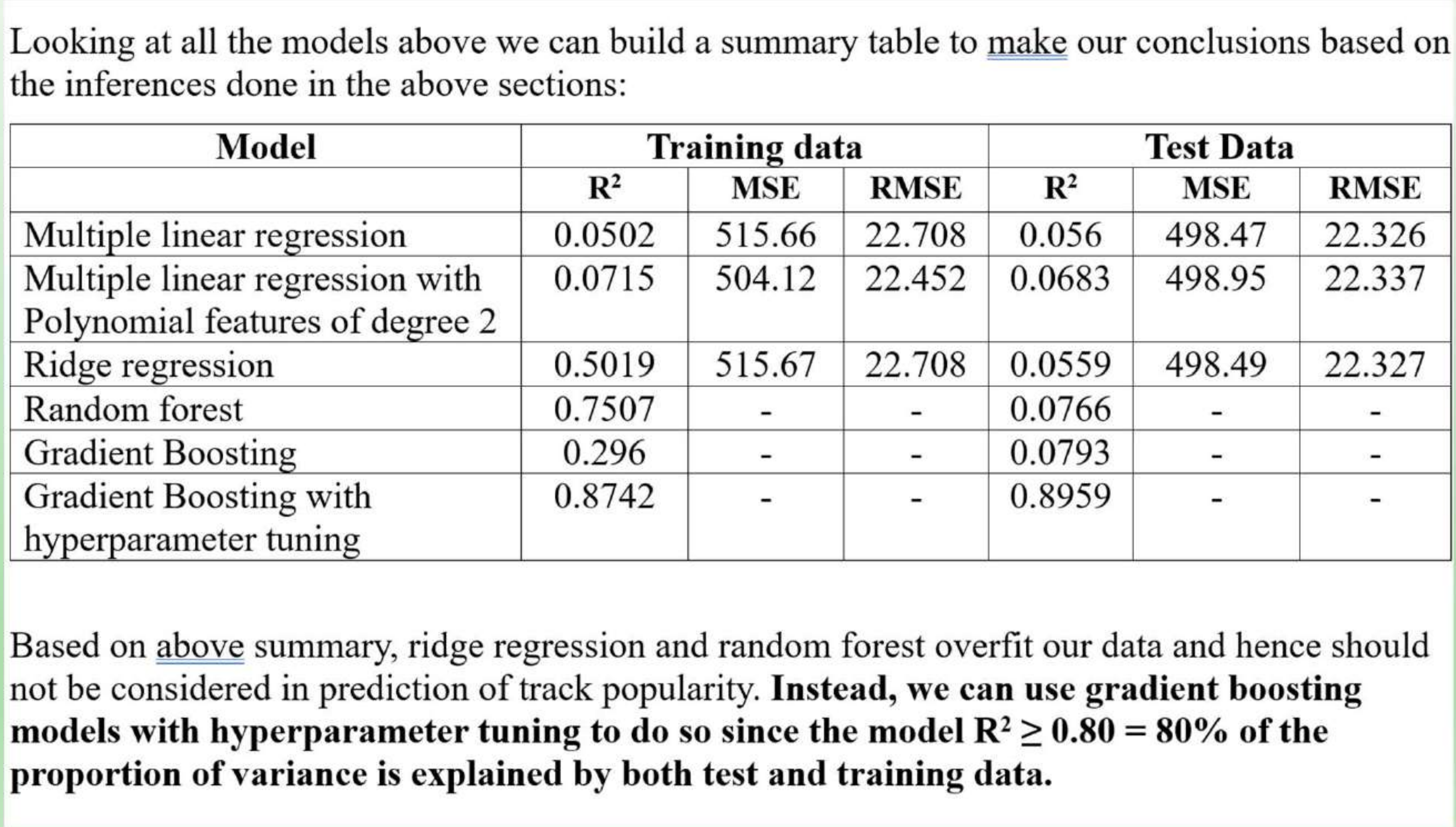


```
Fitting 3 folds for each of 20 candidates, totalling 60 fits
Best Hyperparameters: {'subsample': 0.8, 'n_estimators': 100, 'min_samples_split': 5, 'max_depth': 3, 'learning_rate': 0.1}
Training R^2: 0.8742
Testing R^2: 0.8959
```



```
Top 10 Feature Importances:
             feature  importance
5   instrumentalness    0.204180
11         log_tempo    0.086177
4       acousticness    0.079133
13   loudness_squared    0.075239
0       danceability    0.069565
8              tempo    0.068974
2           loudness    0.067609
3        speechiness    0.050851
7            valence    0.045449
6           liveness    0.044150
```

CONCLUSION



Looking at all the models above we can build a summary table to make our conclusions based on the inferences done in the above sections:

Model	Training data			Test Data		
	R^2	MSE	RMSE	R^2	MSE	RMSE
Multiple linear regression	0.0502	515.66	22.708	0.056	498.47	22.326
Multiple linear regression with Polynomial features of degree 2	0.0715	504.12	22.452	0.0683	498.95	22.337
Ridge regression	0.5019	515.67	22.708	0.0559	498.49	22.327
Random forest	0.7507	-	-	0.0766	-	-
Gradient Boosting	0.296	-	-	0.0793	-	-
Gradient Boosting with hyperparameter tuning	0.8742	-	-	0.8959	-	-

Based on above summary, ridge regression and random forest overfit our data and hence should not be considered in prediction of track popularity. **Instead, we can use gradient boosting models with hyperparameter tuning to do so since the model $R^2 \geq 0.80$ = 80% of the proportion of variance is explained by both test and training data.**