

Popularity Prediction of Spotify Music



By Nick Seguljic



Identify features and create
models to predict the
popularity of a song based on
the features provided by the
Spotify API



Overview



After model development the Linear Regression was best at predicting popularity with an MAE of 6.26 and RMSE of 8.65



Random Forest Model was best at making the binary prediction if a song would be popular not with an accuracy of 90% and MAE of 1.64 and RMSE of 0.32

	Model	Accuracy	MAE	MSE	RMSE
2	KNeighborsClassifier	0.000000	37.506617	1526.300102	39.067891
3	DecisionTreeClassifier	0.011266	19.802104	540.345029	23.245323
4	AdaBoostClassifier	0.036037	11.390024	236.521208	15.379246
1	RandomForestClassifier	1.000000	0.000000	0.000000	0.000000
0	LinearRegression	NaN	6.256225	74.748668	8.645731

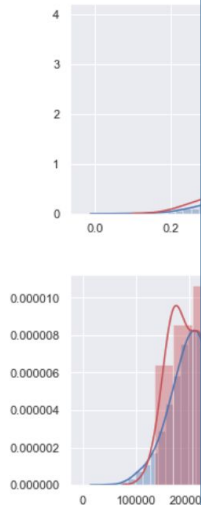


The Data and Feature Engineering

Feature	Data Type
genre	object
artist_name	object
track_name	object
track_id	object
popularity	int64
acousticness	float64
danceability	float64
duration_ms	int64
energy	float64
instrumentalness	float64
key	object
liveness	float64
loudness	float64
mode	object
speechiness	float64
tempo	float64
time_signature	object
valence	float64

Four features added:

1. Time buckets
2. Log transform of the time feature
3. A count of the number of songs that an artist has already made and put out on Spotify
4. Popularity (a binary designation if a song has a popularity greater than 80 or not).



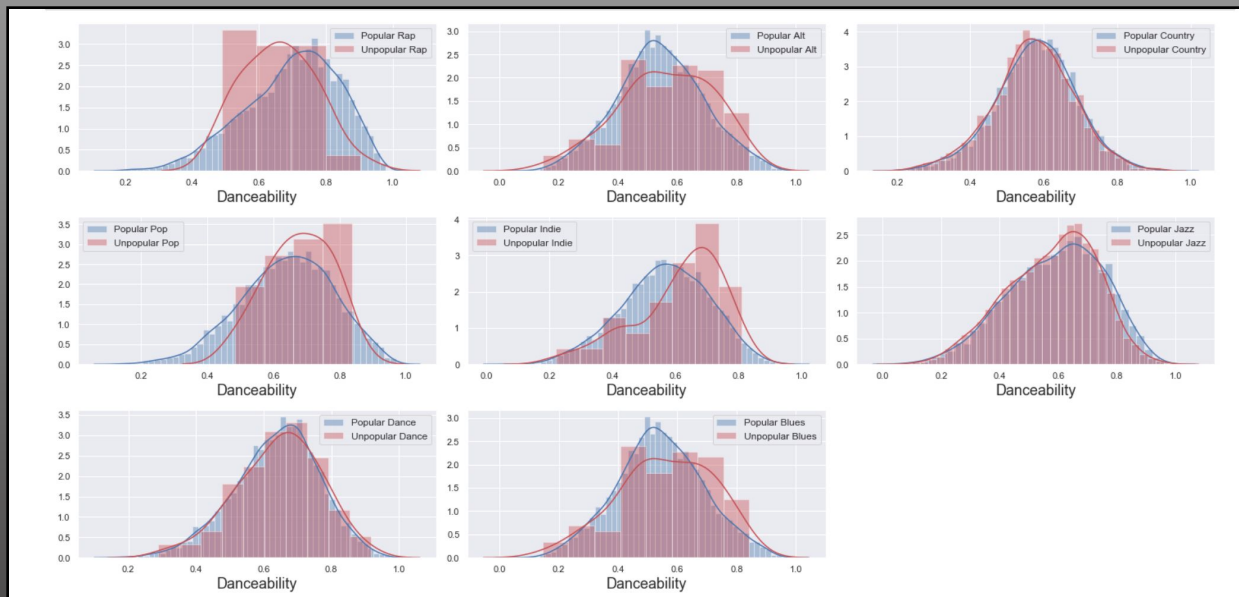
	Popularity Correlation
Unnamed: 0	0.797325
popularity	1.000000
acousticness	0.008367
danceability	0.059371
duration_ms	0.022568
energy	0.045087
instrumentalness	0.007961
liveness	0.055524
loudness	0.055885
speechiness	0.084576
tempo	0.017993
valence	0.003175
lognorm_duration	0.017973
Count	0.127838





Are popular and unpopular data sets statistically different?

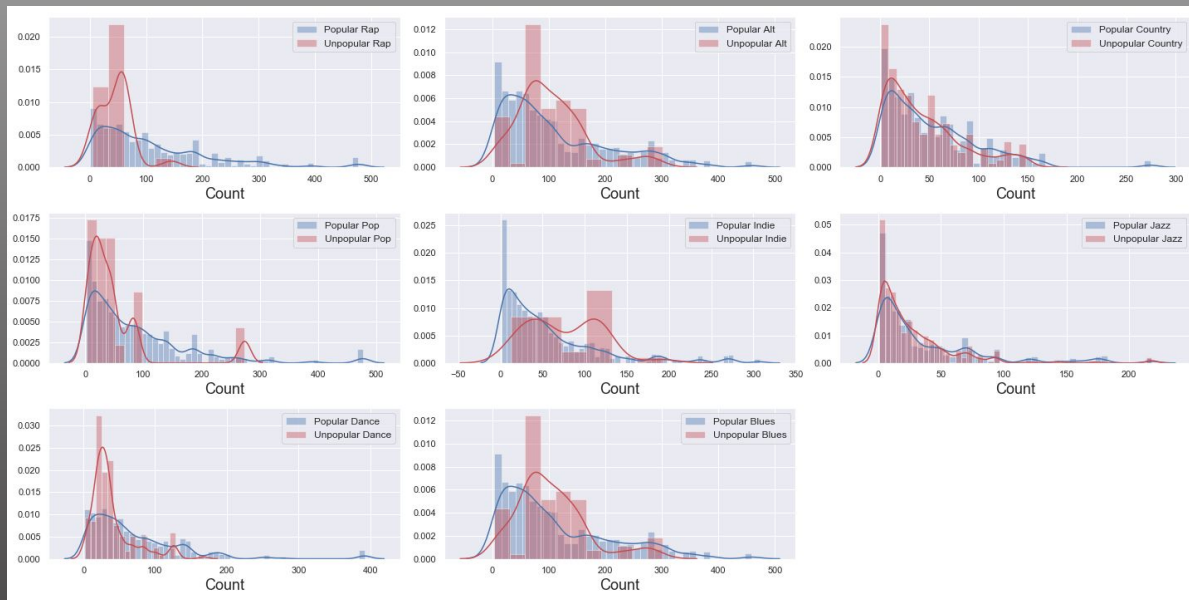
Genre	Statistic	p-value
Rap	-20.682	6.36E-93
Alternative	-74.657	0
Country	-74.657	0
Pop	-29.622	1.99E-184
Indie	-27.149	2.07E-156
Jazz	-100.249	0
Dance	-68.712	0
Blues	-41.185	0





Are popular and unpopular data sets statistically different?

Genre	Statistic	p-value
Rap	-3.305	4.57E-01
Alternative	-0.745	0.457
Country	-9.007	2.56E-19
Pop	-2.133	3.29E-02
Indie	2.496	1.26E-02
Jazz	-12.164	8.66E-34
Dance	-5.717	1.12E-08
Blues	-0.745	0.457





Models

Popularity Score Prediction

	Model	Accuracy	MAE	MSE	RMSE
4	AdaBoostClassifier	0.000068	44.764981	2149.432372	46.361971
2	KNeighborsClassifier	0.000271	36.673023	1462.290126	38.239902
3	DecisionTreeClassifier	0.016491	16.878181	409.673295	20.240388
1	RandomForestClassifier	1.000000	0.000000	0.000000	0.000000
0	LinearRegression	NaN	16.245047	423.599554	20.581534

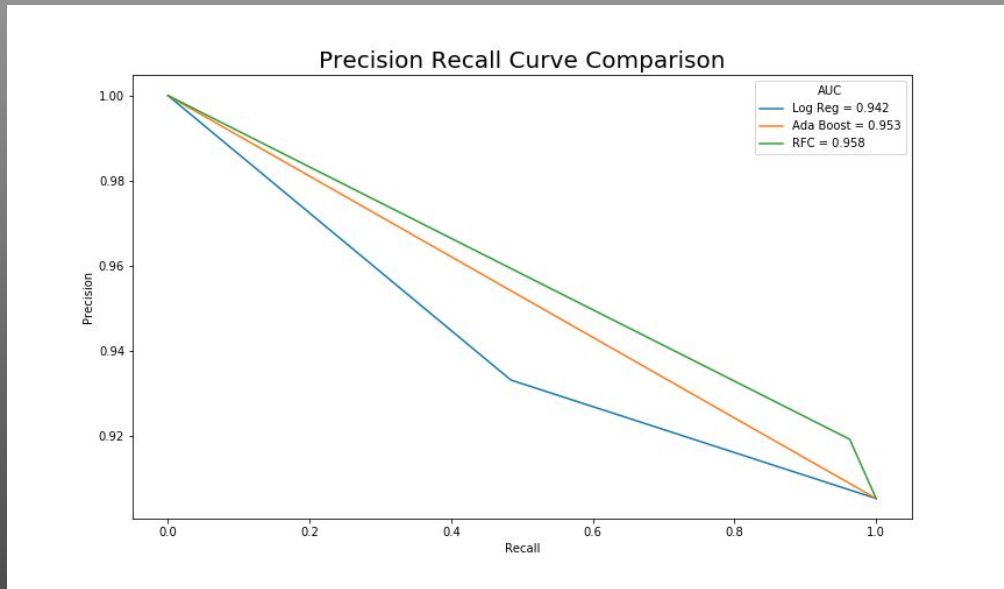
Binary Popularity Prediction

	Model	Accuracy	MAE	MSE	RMSE
1	RandomForestClassifier	0.899355	1.635222	0.101052	0.317887
0	LogisticRegression	0.489243	123.985612	0.510757	0.714672
4	AdaBoostClassifier	0.247302	0.097251	0.097251	0.311852
2	KNeighborsClassifier	0.097251	230.200882	0.902749	0.950131
3	DecisionTreeClassifier	0.097251	230.200882	0.902749	0.950131

Model Performance and Hyperparameters



Recall Curve for Binary Predictor



Model Performance and Hyperparameters



Popularity Score Prediction: Linear Regression

Parameters Tuned by:

```
LinearRegression(copy_X=True,  
fit_intercept=True,  
n_jobs=1,normalize=False)
```

Results:

Mean Absolute Error: 32.65353682668893

Mean Squared Error: 1279.4488412494802

Root Mean Squared Error: 35.76938413293525

Mean cross validation test score:
0.9594329283562244

Mean cross validation train score: 9614382

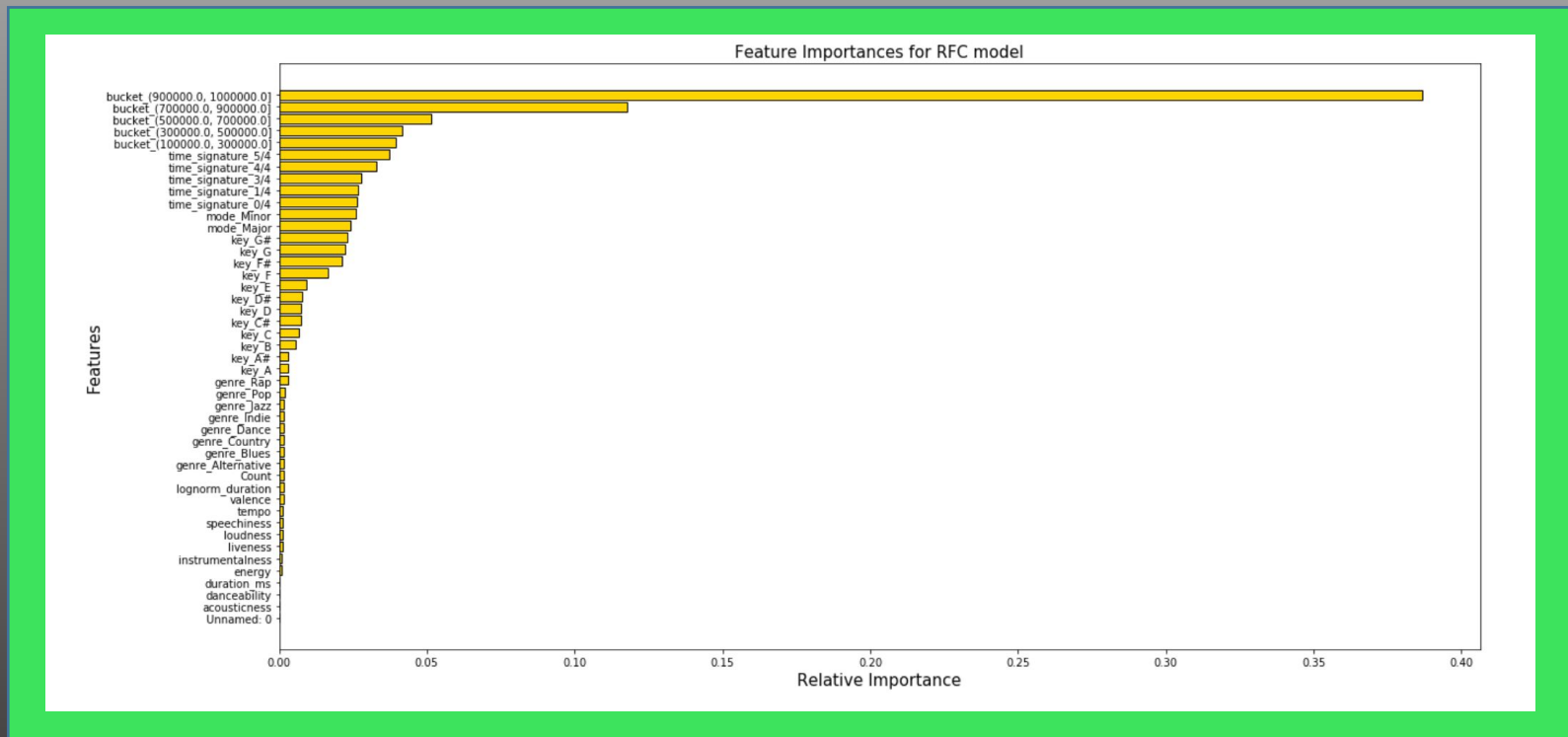
Binary Popularity Prediction: Random Forest

Parameters Tuned by:

```
bootstrap': True,  
'max_depth': 40,  
'max_features': 2,  
'min_samples_leaf': 4,  
'min_samples_split': 5,  
'n_estimators': 1000
```







Important Features





Key Takeaways

-  All genres (except for Blues) the more danceable a song is the more likely it is to be popular
-  Trying to predict popularity alone was best done by the Linear Regression model with MAE of 6.26 and RMSE of 8.65
-  Binary prediction was best done by Random Forest has a 90% chance of being correct and MAE of 1.64 and RMSE of 0.32
-  Time buckets contribute most to popularity and can be seen on the previous slide



- Look into different features and/or feature reduction to see if these features
- Designation of popularity at 80 could be arbitrary and changing to a different number could yield better results
- Training and testing data on models that look exclusively at one genre

Questions?