Modeling Spotify Track Popularity

Nasrin Khansari & Max Tingle

Track Popularity

Duration

Single/Album

Key

Major/Minor

Speechiness

Acousticness

Instrumentalness

Liveness

Valence

Framework:

CRoss-Industry Standard Process for Data Mining (CRISP-DM)

Data:

637,265 Songs from Spotify API

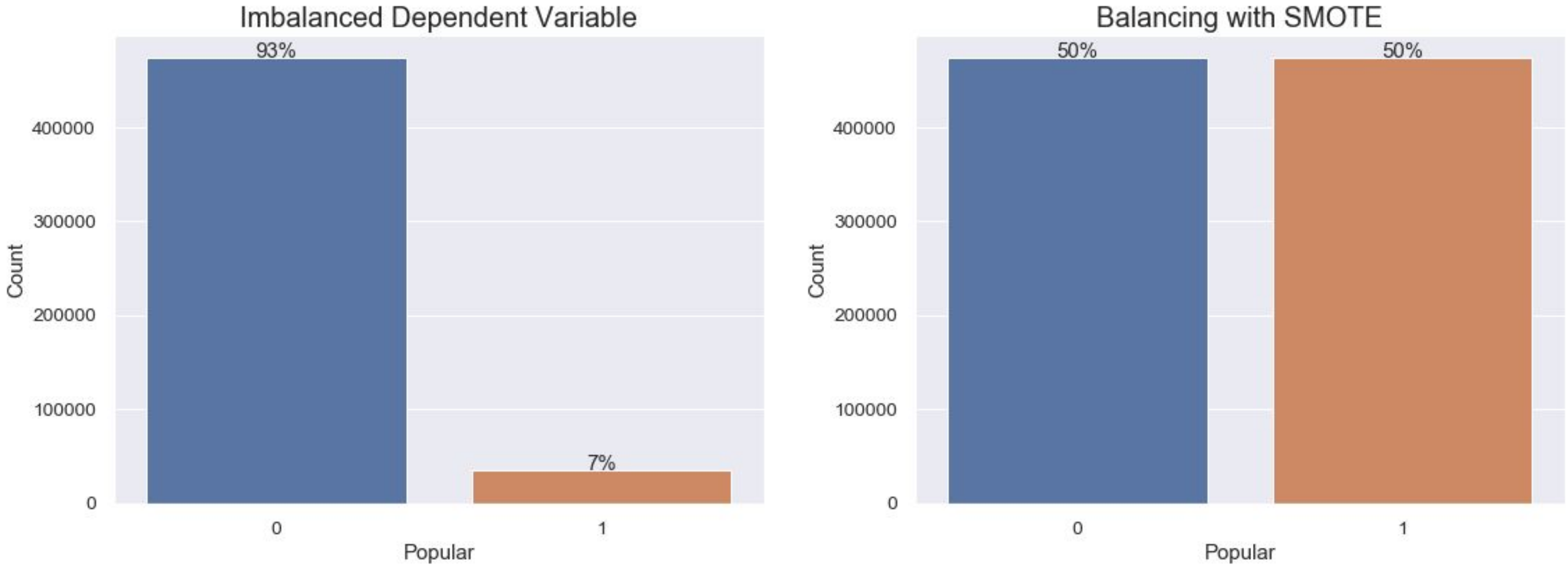
Research Question:

Do the **independent variables** predict the **dependent variable**?

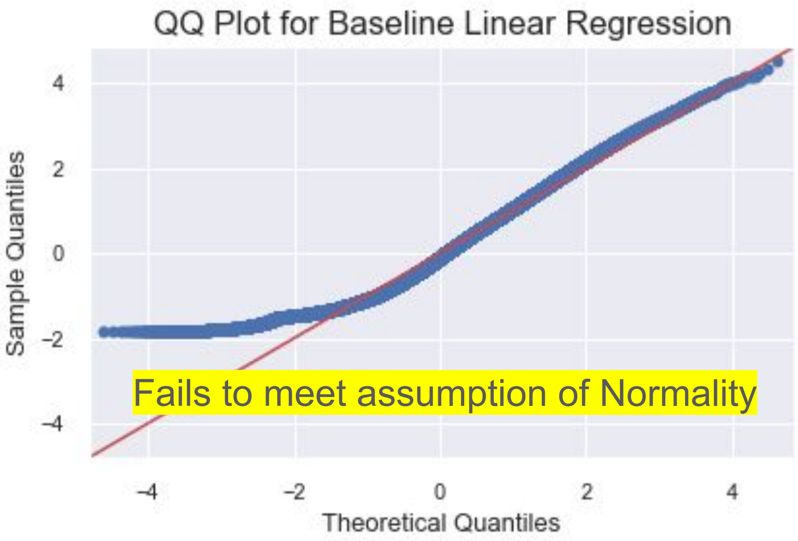
Methodology:

1. Linear Regression
2. Logistic Regression
3. Random Forest Classification

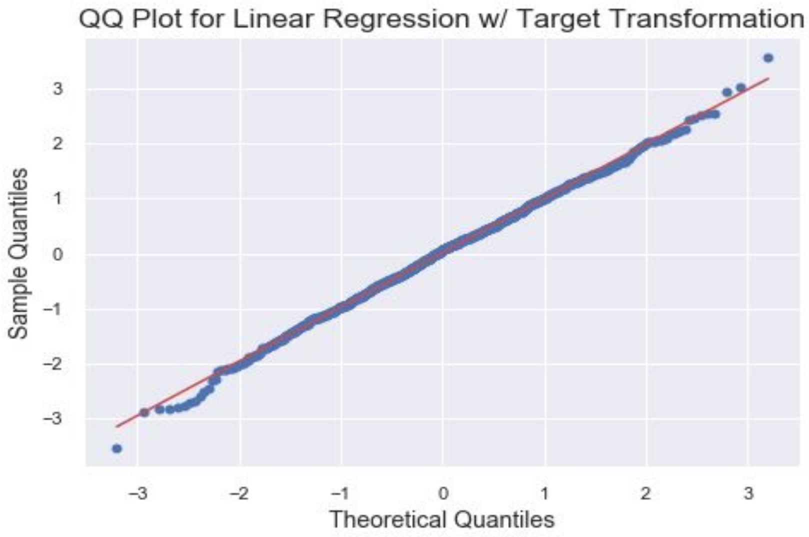
Imbalanced Data



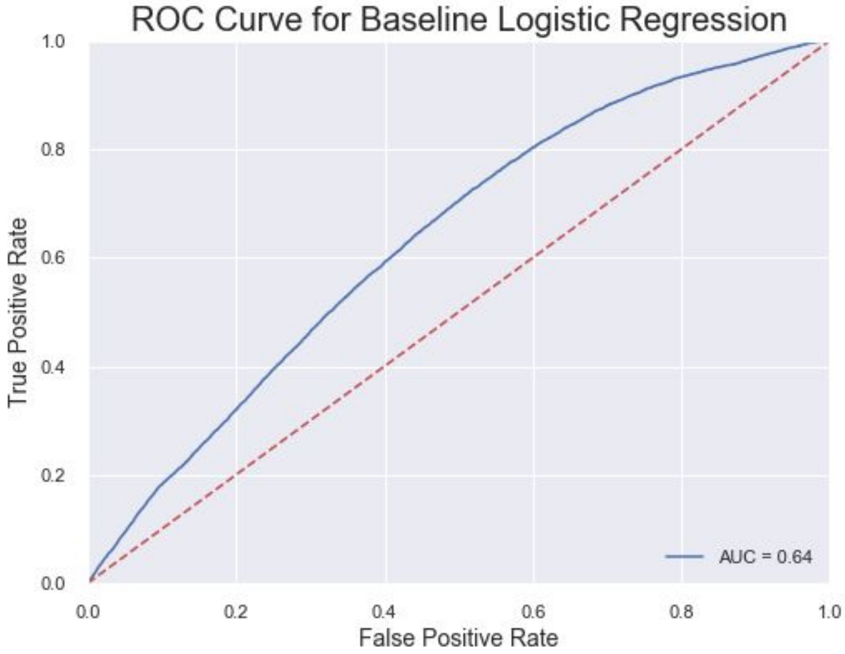
Linear Regression

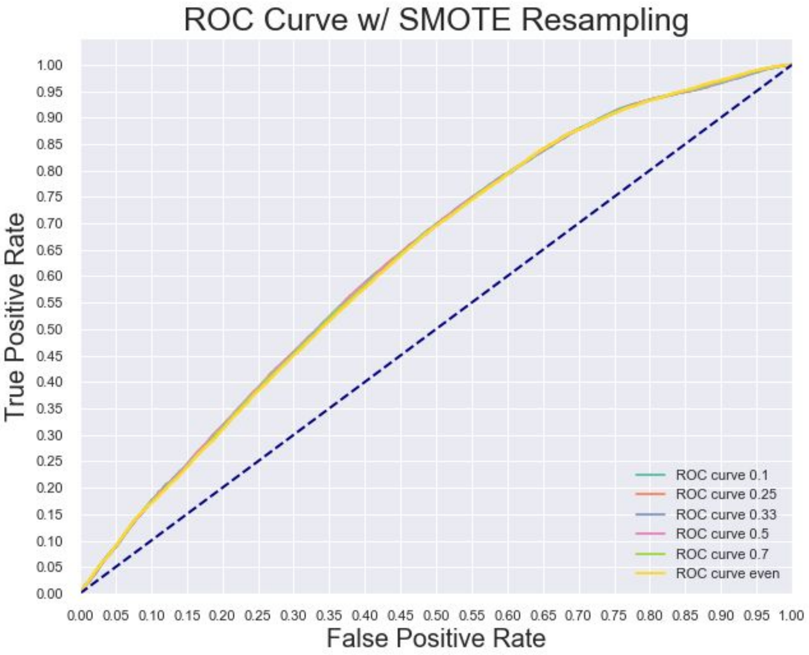


Fails to meet assumption of Normality

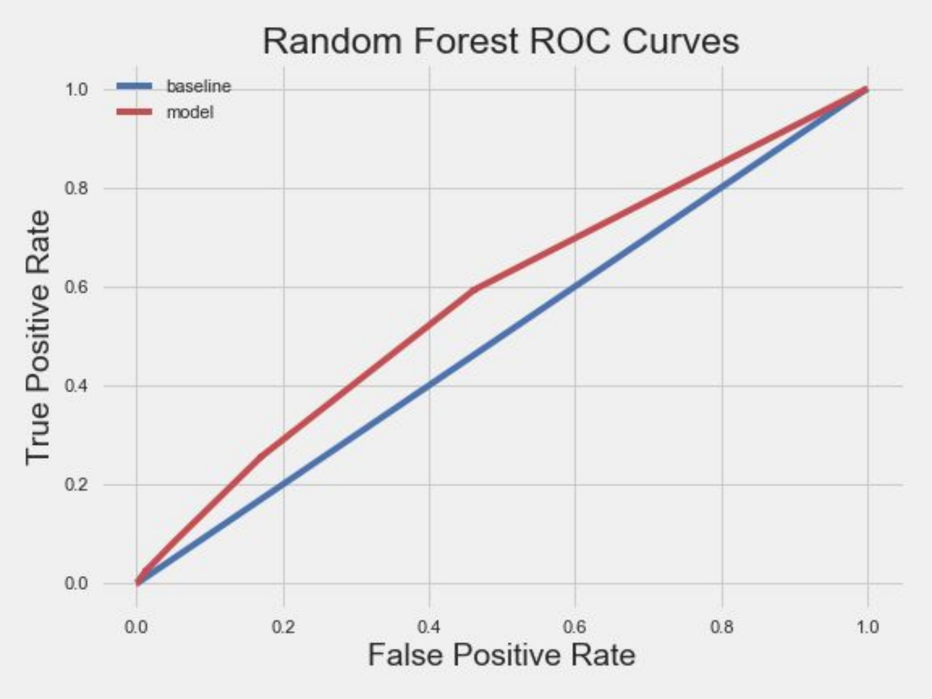


Logistic Regression

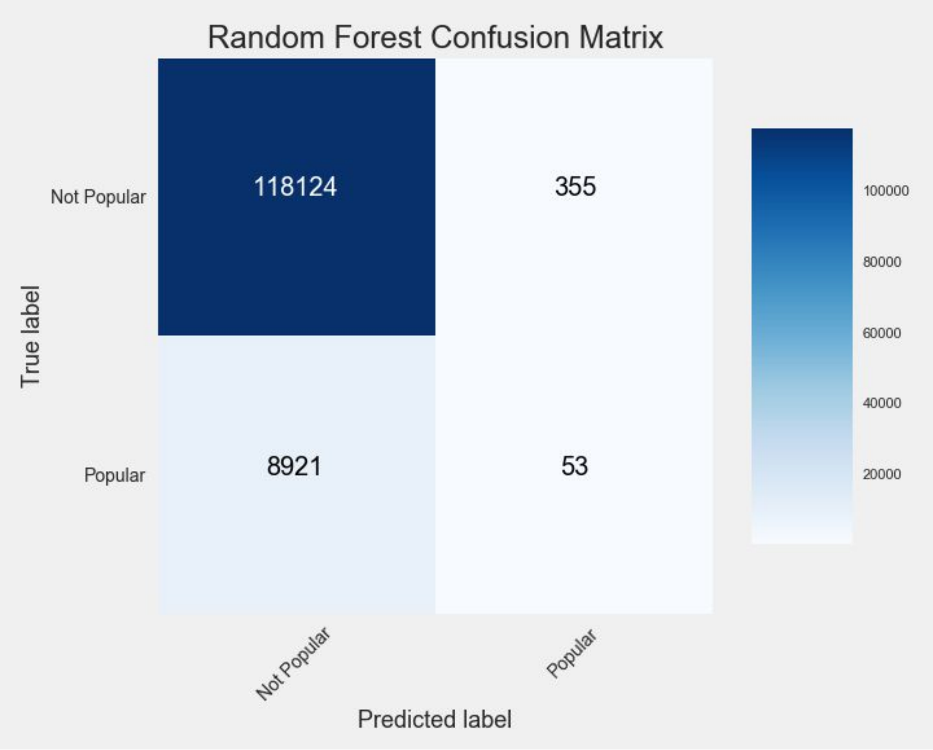




Random Forest

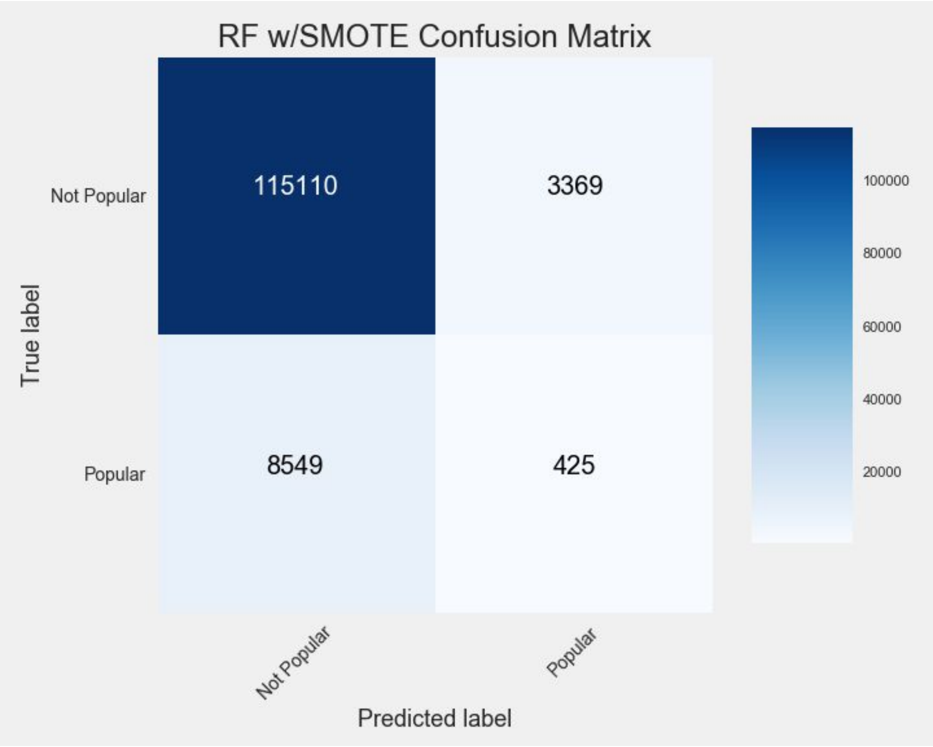


Random Forest



Random Forest Confusion Matrix

True label \ Predicted label	Not Popular	Popular
Not Popular	118124	355
Popular	8921	53



RF w/SMOTE Confusion Matrix

True label \ Predicted label	Not Popular	Popular
Not Popular	115110	3369
Popular	8549	425

Comparing Models

	Logistic Regression		Random Forest		
Scores	Baseline	SMOTE	Classifier	w/ Cross Val.	w/ SMOTE
Accuracy	0.9295	0.4860	0.9272	0.9288	0.9064
Precision	0.0000	0.0938	0.1299	0.1516	0.1120
F1	0.0000	0.1662	0.0112	0.0045	0.0665
Recall	0.0000	0.7276	0.0059	0.0022	0.0473

Track Popularity

- Duration
- Single/Album
- Key
- Major/Minor
- Speechiness
- Acousticness
- Instrumentalness
- Liveness
- Valence

Result:

Cannot say that **independent variables** predict the **dependent variable**.

Future Work

1. Random Forest with Random Search Cross Validation

2. Test Other Classification Algorithms

3. Isolate Collinear Terms

4. Recursive Feature Elimination