Predicting Song Popularity on Spotify Using Supervised Learning

Author: Susmita Sanyal

Problem Definition

- Predict weather a song will be hit or not
  - 1 = Hit
  - 0 = Non-Hit
- Github: https://github.com/susmitaSanyal/Supervised-Learning-Hit-Or-Not

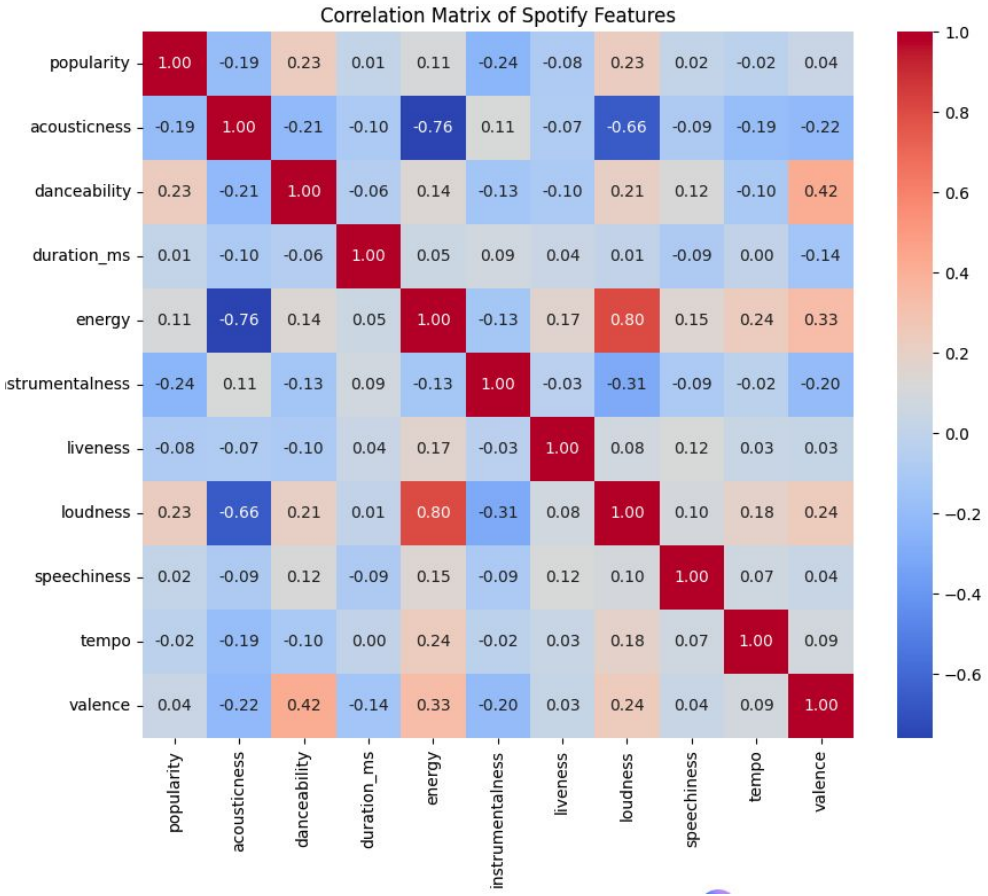
Data Collection

Dataset: Ultimate Spotify Tracks Dataset

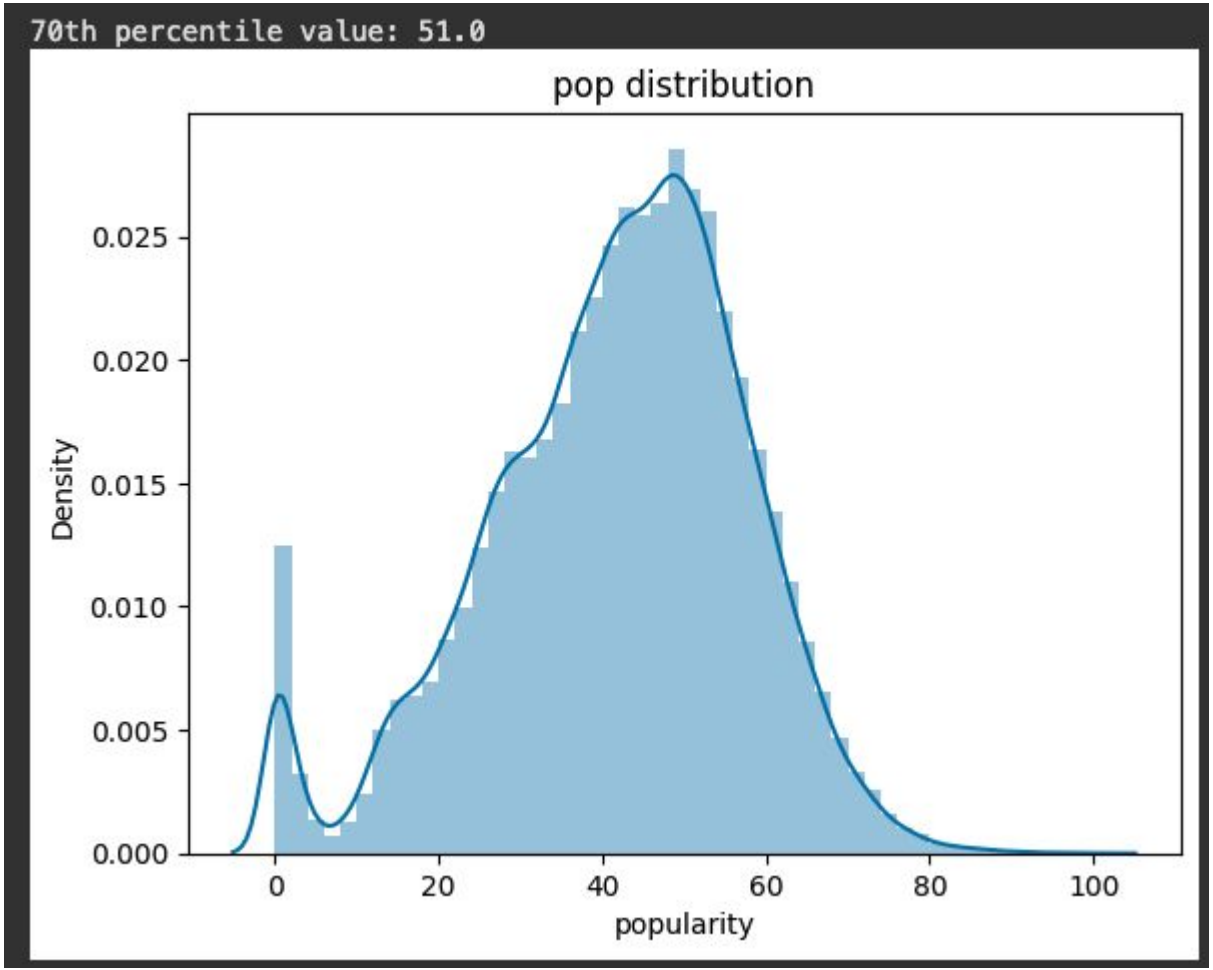
Repository: Kaggle Dataset – Ultimate Spotify Tracks DB

Author: Zaheen Hamidani (Kaggle Contributor)

Features


Correlation Matrix of Spotify Features
popularity 1.00 -0.19 0.23 0.01 0.11 -0.24 -0.08 0.23 0.02 -0.02 0.04
acousticness -0.19 1.00 -0.21 -0.10 -0.76 0.11 -0.07 -0.66 -0.09 -0.19 -0.22
danceability 0.23 -0.21 1.00 -0.06 0.14 -0.13 -0.10 0.21 0.12 -0.10 0.42
duration_ms 0.01 -0.10 -0.06 1.00 0.05 0.09 0.04 0.01 -0.09 0.00 -0.14
energy 0.11 -0.76 0.14 0.05 1.00 -0.13 0.17 0.80 0.15 0.24 0.33
strumentalness -0.24 0.11 -0.13 0.09 -0.13 1.00 -0.03 -0.31 -0.09 -0.02 -0.20
liveness -0.08 -0.07 -0.10 0.04 0.17 -0.03 1.00 0.08 0.12 0.03 0.03
loudness 0.23 -0.66 0.21 0.01 0.80 -0.31 0.08 1.00 0.10 0.18 0.24
speechiness 0.02 -0.09 0.12 -0.09 0.15 -0.09 0.12 0.10 1.00 0.07 0.04
tempo -0.02 -0.19 -0.10 0.00 0.24 -0.02 0.03 0.18 0.07 1.00 0.09
valence 0.04 -0.22 0.42 -0.14 0.33 -0.20 0.03 0.24 0.04 0.09 1.00
popularity
acousticness
danceability
duration_ms
energy
instrumentalness
liveness
loudness
speechiness
tempo
valence
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

```
70th percentile value: 51.0
```



- Logistic Regression: Highly sensitive; collinearity can inflate standard errors and make coefficient interpretation unreliable
- Naive Bayes: Assumes feature independence; correlated features violate this assumption
- Random Forest: Relatively robust; can handle correlated features
- XGBoost: Relatively robust; can handle correlated features
- KNN: Not directly affected by collinearity

Logistic Regression Summary

```
[LogisticRegression] Test Accuracy: 0.7031
[LogisticRegression] Test PR-AUC:   0.4595
[LogisticRegression] Test ROC AUC:  0.6779

Classification Report (thr=0.5):
              precision    recall  f1-score   support

         0.0     0.7126    0.9604    0.8182      9105
         1.0     0.5598    0.1152    0.1911      3985

    accuracy                         0.7031     13090
   macro avg     0.6362    0.5378    0.5046     13090
weighted avg     0.6661    0.7031    0.6272     13090
```

KNN Summary

```
[KNN] Test Accuracy: 0.7475
[KNN] Test PR-AUC:   0.6045
[KNN] Test ROC AUC:  0.7422

Classification Report (thr=0.5):
              precision    recall  f1-score   support

         0.0     0.7811    0.8851    0.8298      9105
         1.0     0.6227    0.4331    0.5109      3985

    accuracy                         0.7475     13090
   macro avg     0.7019    0.6591    0.6704     13090
weighted avg     0.7328    0.7475    0.7327     13090
```

Naive Bayes Summary

```
[NaiveBayes] Test Accuracy: 0.5758
[NaiveBayes] Test PR-AUC:   0.4430
[NaiveBayes] Test ROC AUC:  0.6603

Classification Report (thr=0.5):
              precision    recall  f1-score   support

         0.0     0.7909    0.5304    0.6349      9105
         1.0     0.3877    0.6795    0.4938      3985

    accuracy                         0.5758     13090
   macro avg     0.5893    0.6050    0.5643     13090
weighted avg     0.6681    0.5758    0.5920     13090
```

Random Forest Summary

```
[RandomForest] Test Accuracy: 0.7676
[RandomForest] Test PR-AUC:   0.6217
[RandomForest] Test ROC AUC:  0.7708

Classification Report (thr=0.5):
              precision    recall  f1-score   support

         0.0     0.7908    0.9053    0.8442      9105
         1.0     0.6768    0.4529    0.5427      3985

    accuracy                         0.7676     13090
   macro avg     0.7338    0.6791    0.6935     13090
weighted avg     0.7561    0.7676    0.7524     13090
```

XGBoost Summary

```
[XGBoost] Test Accuracy: 0.6544
[XGBoost] Test PR-AUC:   0.5696
[XGBoost] Test ROC AUC:  0.7512

Classification Report (thr=0.5):
              precision    recall  f1-score   support

         0.0     0.8428    0.6185    0.7134      9105
         1.0     0.4579    0.7365    0.5647      3985

    accuracy                         0.6544     13090
   macro avg     0.6504    0.6775    0.6391     13090
weighted avg     0.7257    0.6544    0.6682     13090
```

The Winner is……

Random Forest 🎉

- (AUC=0.7708)

Test Data

track_name	artist_name	acousticness	danceability	energy	instrumentalness	liveness	mode	speechiness	tempo	time_signature	valence
Blinding Lights	The Weeknd	0.05	0.8	0.95	0.0	0.1	1	0.05	171.0	4	0.85
Uptown Funk	Mark Ronson ft. Bruno Mars	0.1	0.87	0.92	0.0	0.12	1	0.08	115.0	4	0.9
Shake It Off	Taylor Swift	0.15	0.82	0.88	0.0	0.09	1	0.06	160.0	4	0.95
Happy	Pharrell Williams	0.12	0.9	0.89	0.0	0.11	1	0.05	160.0	4	0.92
Motionless	Indie Artist	0.65	0.45	0.4	0.02	0.2	0	0.04	85.0	3	0.35
Dreams in Grey	Lo-Fi Producer	0.7	0.5	0.35	0.05	0.18	1	0.03	90.0	4	0.4
Dark Matter	Experimental Band	0.55	0.35	0.5	0.03	0.22	0	0.07	110.0	5	0.3
Acoustic Soul	Folk Singer	0.85	0.4	0.25	0.06	0.15	1	0.04	98.0	4	0.5
Rainfall	Chillhop Collective	0.75	0.48	0.45	0.04	0.16	0	0.05	92.0	4	0.42
Low Tide	Jazz Ensemble	0.68	0.38	0.55	0.02	0.19	1	0.06	100.0	4	0.37