Where is the Music Scene Headed?

Caleb Elgut - September 2020

Introduction

- Data Source: Spotify API
  - Consists of nearly 170,000 songs released between 1921 and 2020
  - Each song is broken down into a myriad of factors all measured by algorithms provided by Spotify.
    - These algorithms measure everything from reasonably-quantifiable factors such as duration, key, and loudness (measured by decibel) to factors one wouldn't initially think were quantifiable such as the level of energy in a song or its cheerfulness.
  - Python Libraries:
    - Pandas, Matplotlib, Seaborn, Numpy, sklearn, math, statsmodels, warnings, itertools, keras

GOAL

Guide music executives who are looking to invest in certain elements of music while shaping their artists into further popularity.

AND

Help independent artists understand which features are used the least and don't traditionally predict popularity--perhaps these artists can use the rarity of these features to their advantage in an attempt to stand out.

Part One: Classification

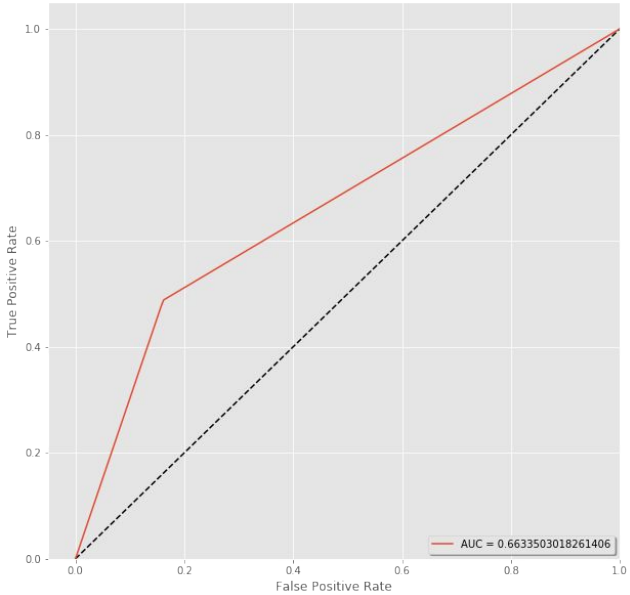
Classification Analysis

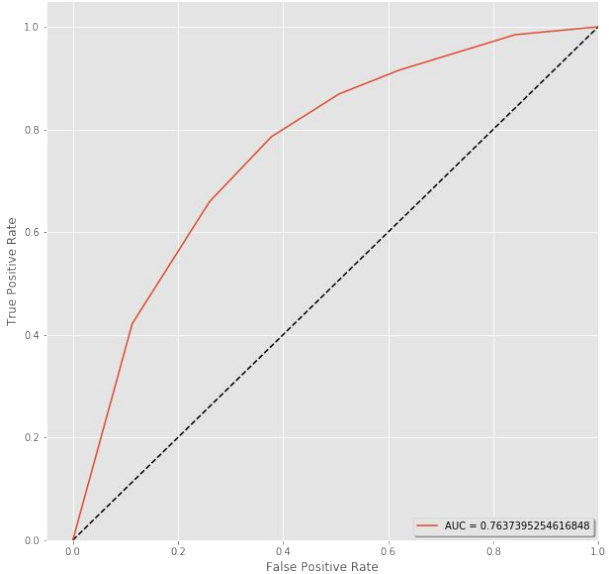
```
[[22115  4181]
 [ 3991  3695]]
              precision    recall  f1-score   support

           0       0.85      0.84      0.84     26296
           1       0.47      0.48      0.47      7686

    accuracy                           0.76     33982
   macro avg       0.66      0.66      0.66     33982
weighted avg       0.76      0.76      0.76     33982
```

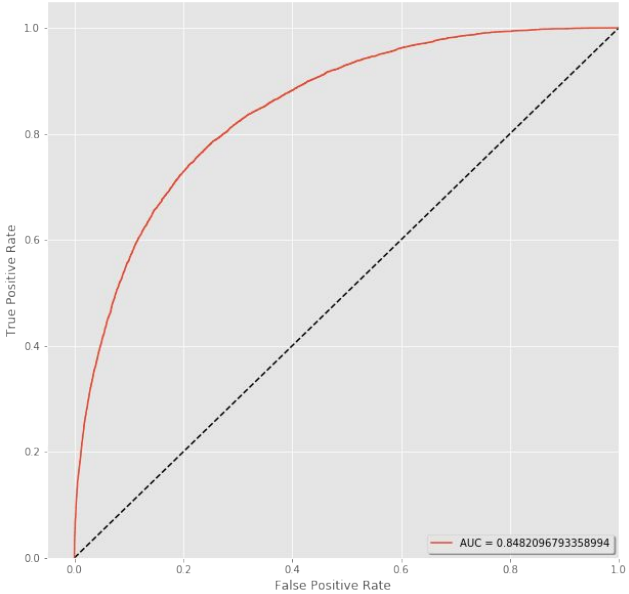
```
[[23320  2976]
 [ 4444  3242]]
              precision    recall  f1-score   support

           0       0.84      0.89      0.86     26296
           1       0.52      0.42      0.47      7686

    accuracy                           0.78     33982
   macro avg       0.68      0.65      0.66     33982
weighted avg       0.77      0.78      0.77     33982
```

```
[[24688  1608]
 [ 4226  3460]]
              precision    recall  f1-score   support

           0       0.85      0.94      0.89     26296
           1       0.68      0.45      0.54      7686

    accuracy                           0.83     33982
   macro avg       0.77      0.69      0.72     33982
weighted avg       0.82      0.83      0.81     33982
```

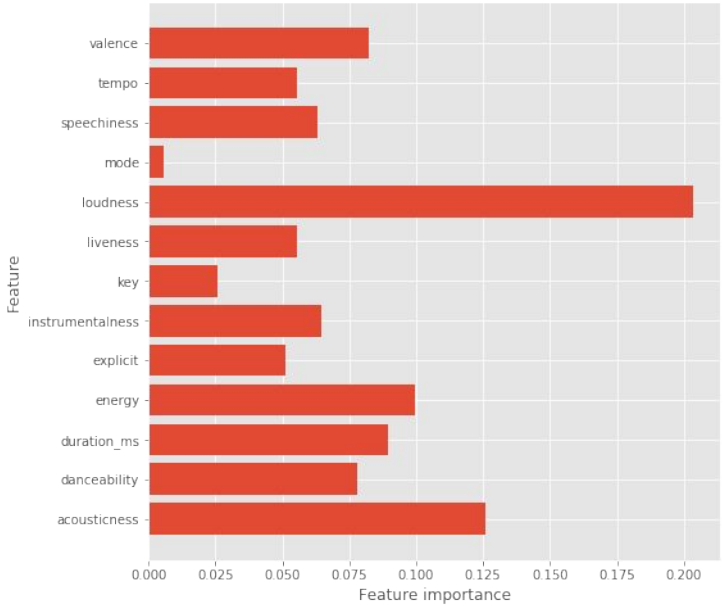
Classification Analysis: AUC Progression

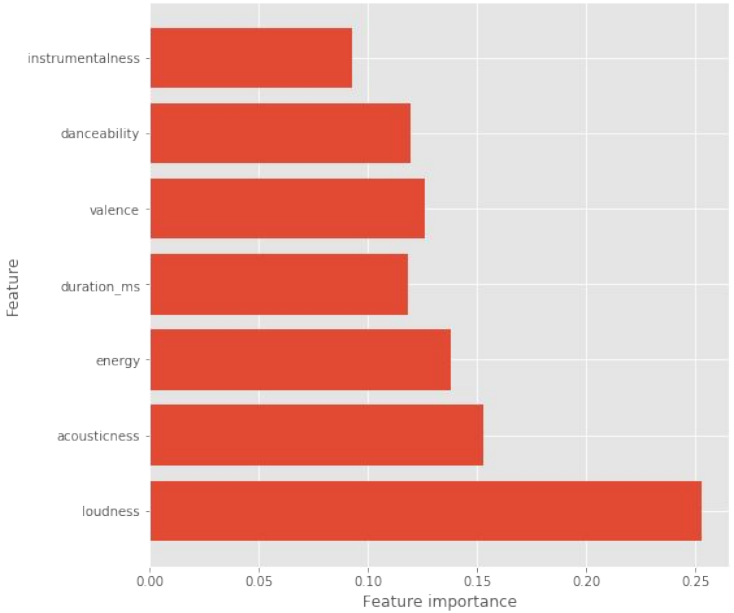






Classification Analysis



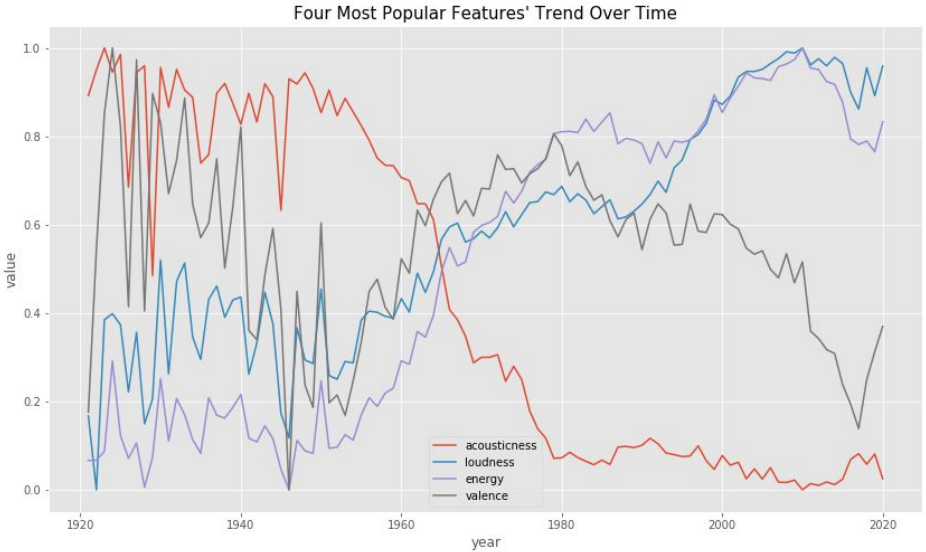


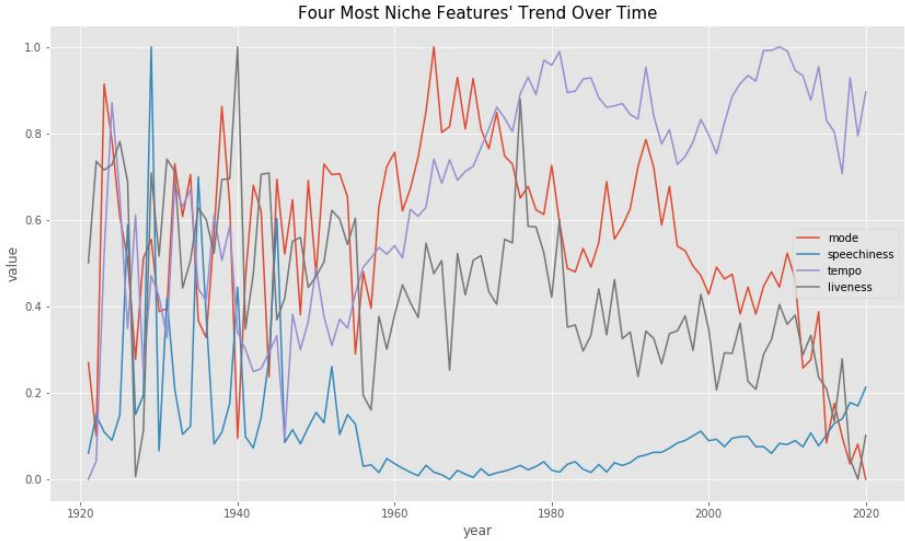
Classification Analysis

- Shows us which features will most likely predict popularity.
- Number of Models Run: Nearly 15
- Best Model:
  - Random Forest Grid-Searched on All Features & Top 7
- Best Features?
  - Popular: Acousticness, Energy, Loudness, Valence
  - Niche: Mode, Tempo, Liveness, Speechiness
- On to Time Series Analysis!

Time Series Analysis: ARIMA

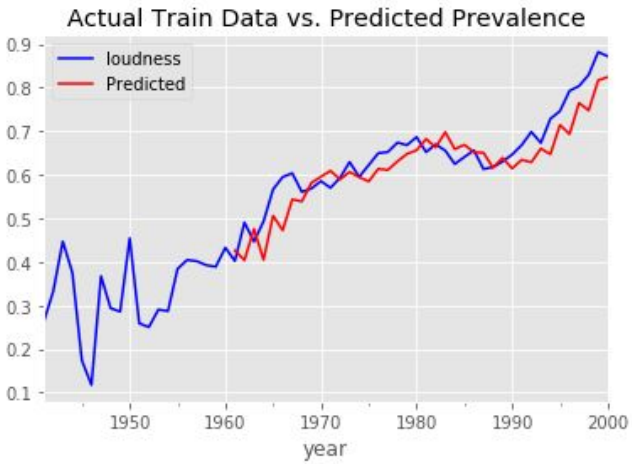
Time Series Analysis: ARIMA

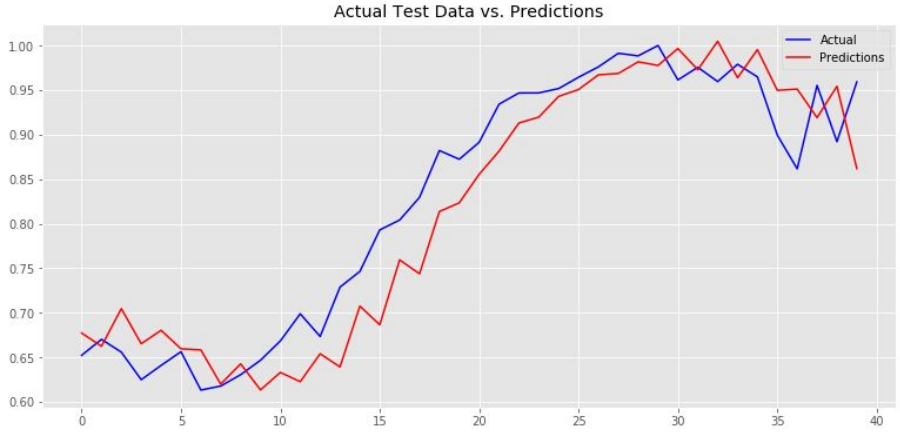




Which Model Performed Best?

- Loudness -- Both RMSEs: 0.048





Conclusions from ARIMA models?

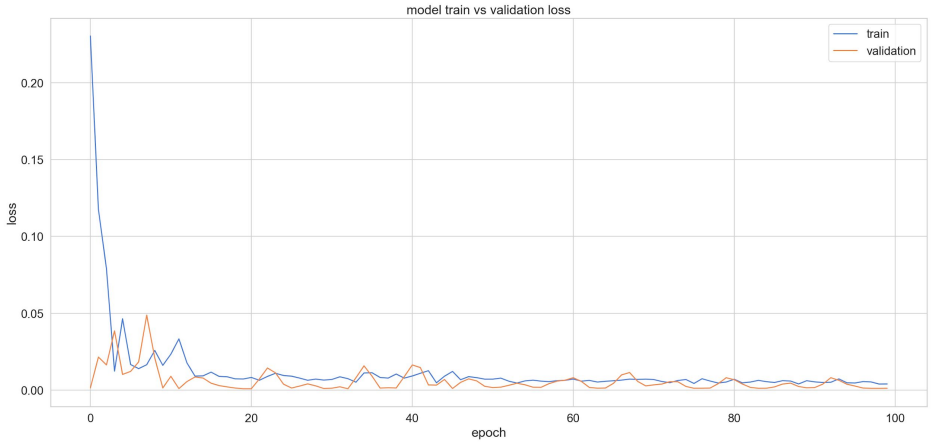
- **Conclusion for Popular Features**: The prevalence of Loudness & Energy will grow the most of the 4 features that most likely predict popularity.
- **Conclusion for Niche Features: Include Speechiness & a Major Scale to stand out**
    - **Both are not being used much right now and their growth in prevalence will be incredibly slow which will give you a chance to be unique when implementing them.**

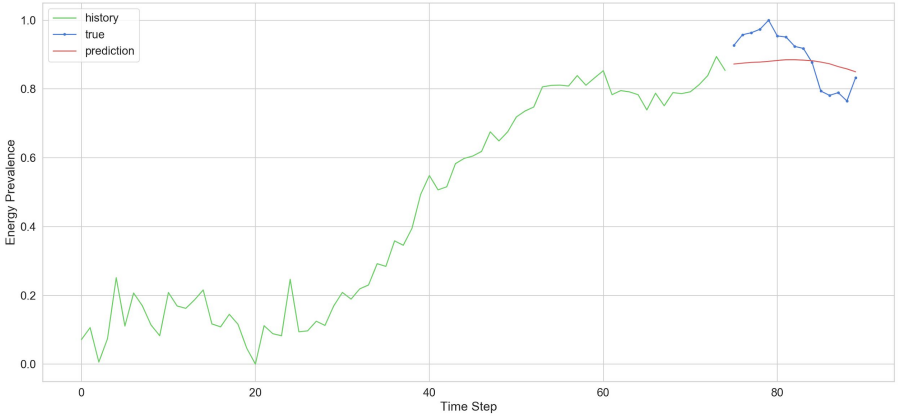
Time Series Analysis: LSTM

Time Series Analysis: LSTM

- Can be more accurate than an ARIMA model, we ran some tests to determine if this should be used in place of our ARIMA models.
- Results: Decent!

LSTM Visualized:





Future Research?

Include Monthly Data

Majority of Months (All but around 10,000) were missing so I relied on annual data.

Months would increase accuracy & reliability of models.

Conclusions:

My Pick: ARIMA Model

Music Executives, invest in energy & loudness!

Independent Artists, invest in speechiness & songs with a major scale!

THANK YOU!