Hit Predictor

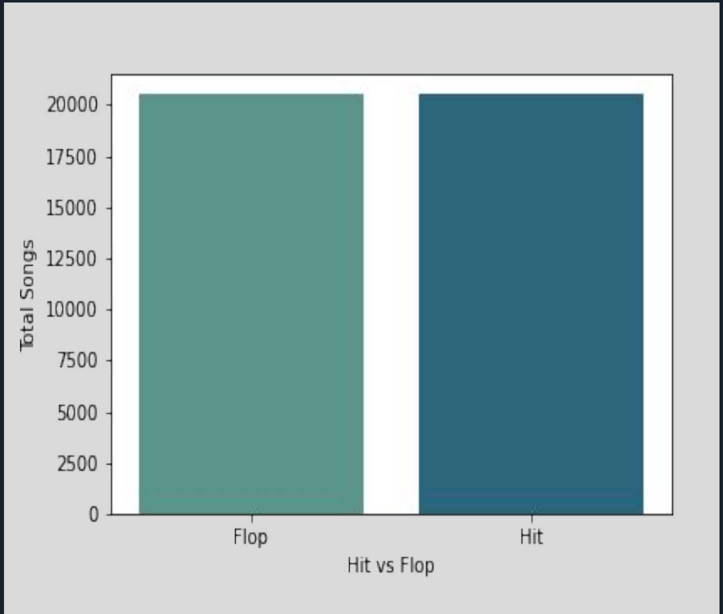
Sibongile Toure

Data

- Spotify Hit Predictor Kaggle Data Set (https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset)
- Song that were featured on Billboards hot 100 as the ‘hits’
- Songs that meet a certain set of criteria that the creator of the dataset created are deemed a flop
- Data contains songs from 1960-2019
    - For the sake of this project I am using data from 1980-2019

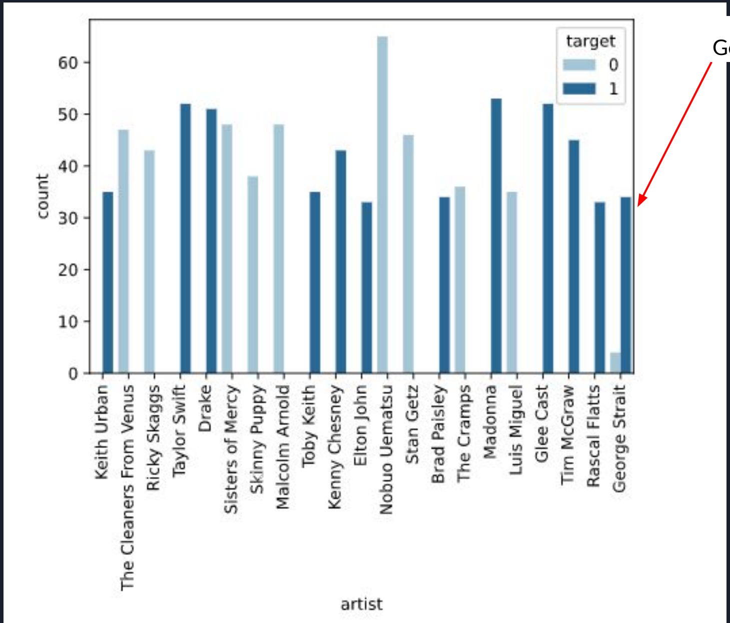
EDA

- Perfectly balanced classes
- Features
    - Categorical
        - Artist
        - Track
        - URI
    - Continuous
        - Danceability
        - Energy
        - Key
        - Loudness
        - Speechiness
        - Chorus_Hit



Feature Engineering

- 9453 Unique artist in dataframe
- Decided to encode top 100 artist that appears in the dataframe
- Adding the artist slightly improved the overall scores for all of the models

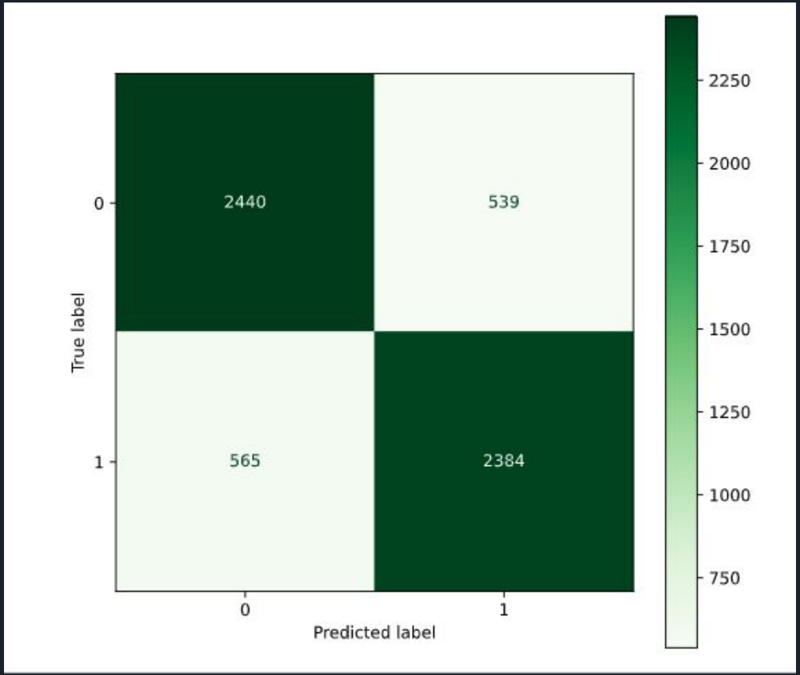


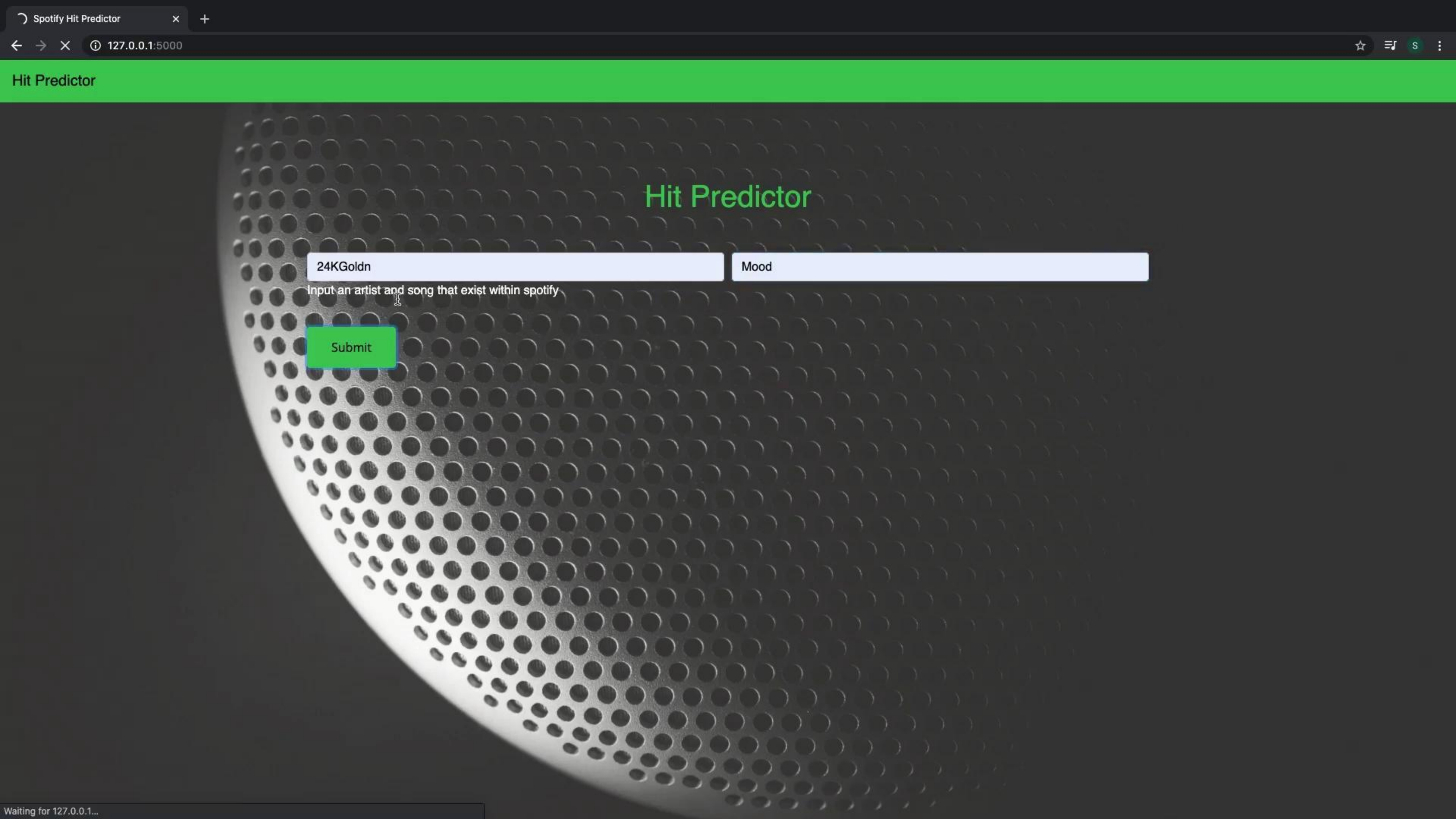
George Strait

Final Model

- Final Model chosen was an ensemble method Voting Classifier w/ ‘soft’ voting
- **VC val score: 0.813**
- **VC test score: 0.808**

	Precision	Recall	F1
Flop	0.81	0.82	0.81
Hit	0.82	0.82	0.81





Hit Predictor

24KGoldn

Mood

Input an artist and song that exist within spotify

Future Work

Model Improvements

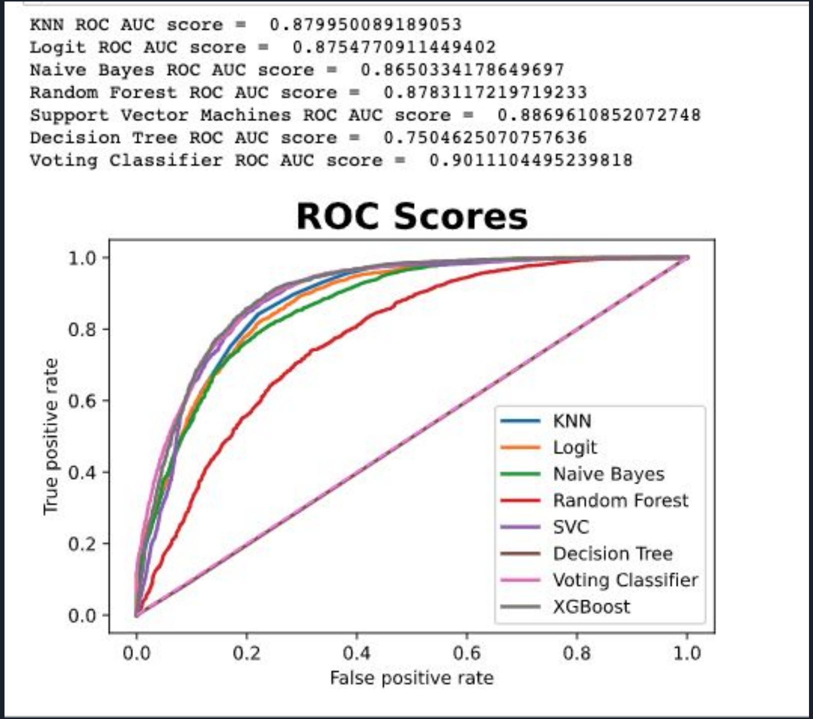
- Explore Genre
- Explore even more of the musical features

App Improvements

- Take in music not only from Spotify but from other smaller music sources such as Soundcloud ,Bandcamp,etc.

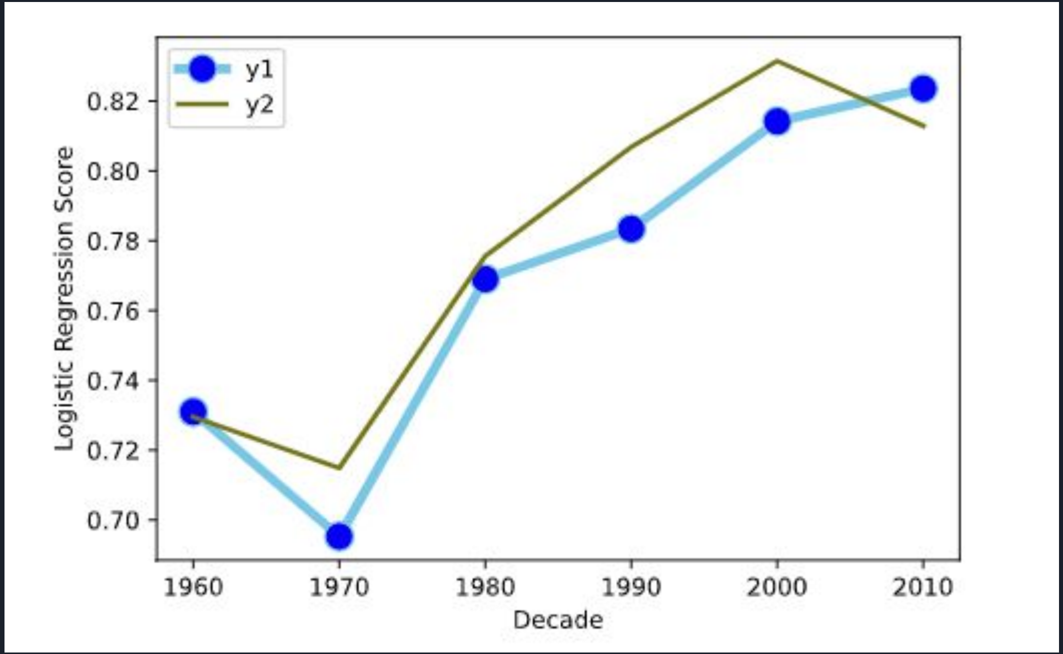
Questions

Appendix



```
KNN ROC AUC score =  0.879950089189053
Logit ROC AUC score =  0.8754770911449402
Naive Bayes ROC AUC score =  0.8650334178649697
Random Forest ROC AUC score =  0.8783117219719233
Support Vector Machines ROC AUC score =  0.8869610852072748
Decision Tree ROC AUC score =  0.7504625070757636
Voting Classifier ROC AUC score =  0.9011104495239818
```

Appendix


y1
y2
Logistic Regression Score
0.82
0.80
0.78
0.76
0.74
0.72
0.70
1960
1970
1980
1990
2000
2010
Decade