Using Data to Predict the Next Hit Album

A Springboard Capstone Project
By Jessica Wulzen

Problem Statement

- Music critics play a key role in promoting new releases from up and coming artists
- Understanding what kinds of music critics respond well to will help aspiring musicians craft albums, and record labels to make informed decisions about what acts have the best chance of being successful
- The data prepared to answer this question can also be used to develop a classification model to predict what genre an album is based on its musical features

Data

- The most widely used music streaming platform
- Spotify's API contains a tremendous amount of data about every song, such as: key, time signature, energy, loudness, acousticness, speechiness, and much more

- #1 Website for music reviews
- Over 17,000 reviews from 1996 to 2017
- Each review contains a numeric score from zero to ten

Data Cleaning

- The data needs to be prepared so that the albums in the Pitchfork dataset can be searched for in the Spotify API
- Many albums have differences in how they are listed in the two data sources, such as:
  - Fleet Foxes - Sun Giant EP vs Fleet Foxes - Sun Giant
  - Massive Attack - Mezzanine vs Massive Attack - Mezzanine [Remastered]
- To fix this, regular expressions are used to remove troublesome phrases such as ‘EP’ and other common sources of discrepancies

Using the Spotify API

- The python package Spotipy allows for easy use of the Spotify API
- Spotify limits results to a certain number, making it difficult to find obscure musicians and albums
- To find as many matches as possible, we'll look for matches in two ways:

Sun Giant by Fleet Foxes

1. Search for the album 'Sun Giant'
2. Check to see if the musician is 'Fleet Foxes'

1. Search for the musician 'Fleet Foxes'
2. Retrieve the list of albums released by this musician
3. Search for a match on the album name 'Sun Giant'

Using this two-part algorithm, we are able to find 12k albums in Spotify out of 17k reviews

Exploratory Data Analysis

To understand the data, we'll look for how certain variables are related with each other and how they may have changed over time, such as:

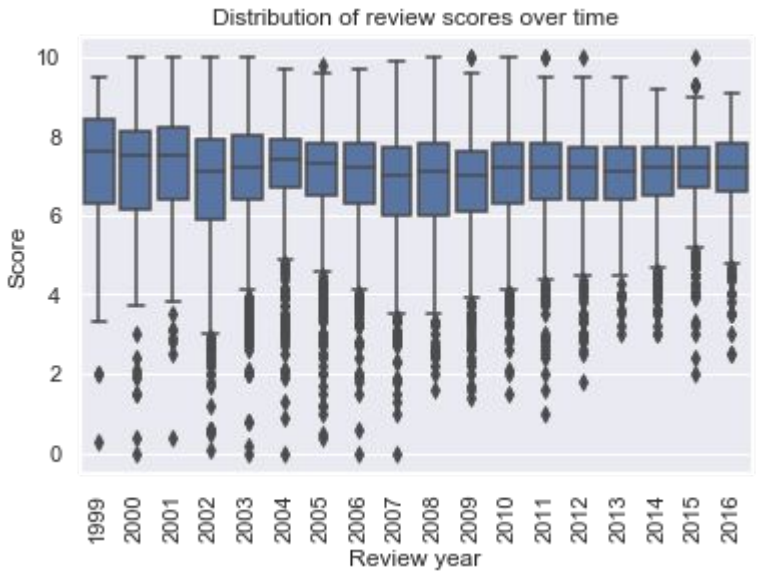
- How have review scores changed over time?
- How has music itself changed over time?
- Does the key music is composed in impact the review score?
- Are certain genres of music better-received than others?

Review Scores over Time

Over time, we observe a change in how Pitchfork has issued scores for reviews

While the average score has remained consistent at around a 7, the distribution has consistently gotten more tightly focused around the mean

In other words, Pitchfork has given less extreme scores over time

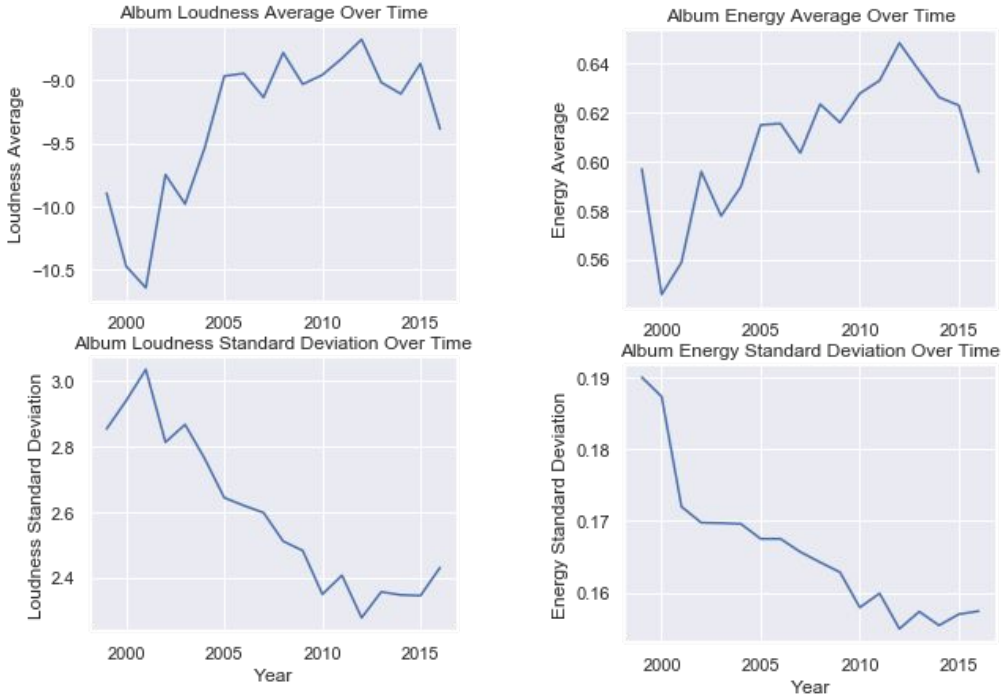


How music has changed over time

Using Spotify's musical attribute data, we can look at trends in music over time

From these, it is clear that music has gotten louder and more energetic over time

As well, the album energy and loudness standard deviation has dropped over this time, meaning albums have varied less and less over time - music is louder, more energetic, and songs on an album tend to be more similar to each other than they used to be

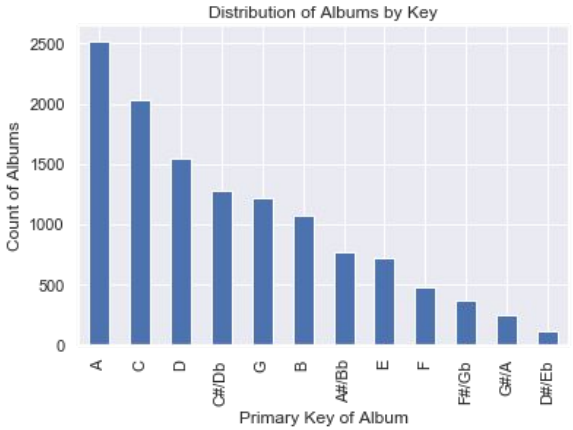


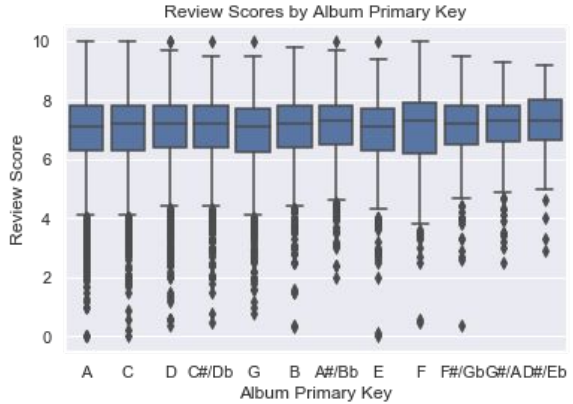
Music by key

One way of looking at these albums is examining the primary key they are composed in

Some keys are vastly more common than others

The obscure keys tend to do better than the popular ones, and do not typically get extremely low scores





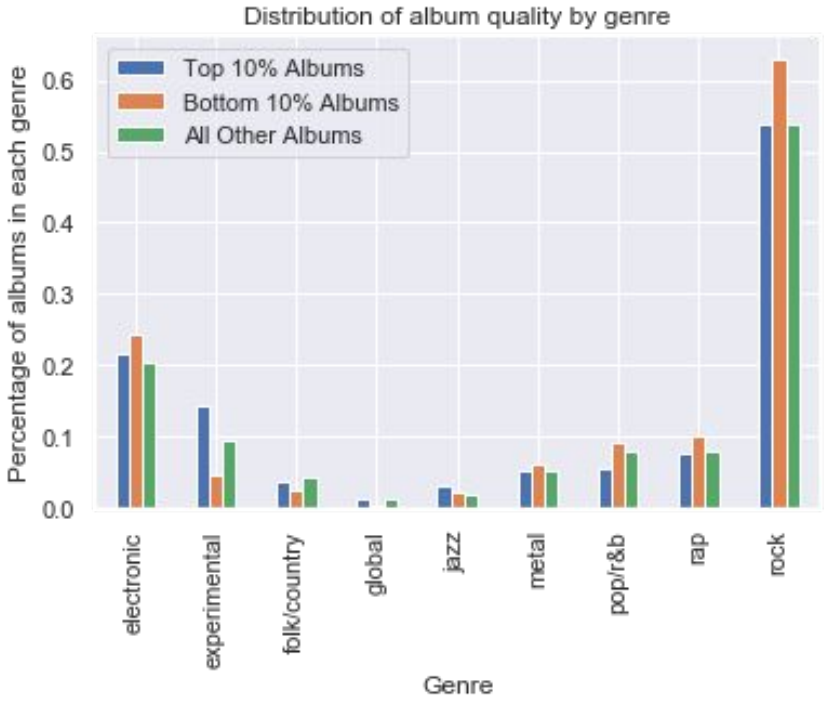
Music by Genre

Breaking albums up into quality categories by review score and looking at how genre impacts this reveals that:

Rock, Electronic, Pop/R&B and Rap are overrepresented in the bottom 10%

Experimental music is massively overrepresented in the top 10%

Other genres have even distributions



Applying Linear Regression

To predict review scores, linear regression is a good starting place to get a baseline of the predictive power of our data and uncover which variables are the most significant

So that each genre can be learned from independently, linear regression models will be fit to each genre's data separately

These starting models have some variables with significant predictive power, but very small adjusted R-squared values - the largest is 0.11

Thus, these models are not very powerful, but lessons about what attributes are important for each genre can still be learned

Lessons from linear regression

Genre	Positive Attributes	Negative Attributes
Rock	● Acousticness ● Instrumentalness ● Valence	● Danceability ● Minor Key
Metal	● Instrumentalness	● Danceability
Jazz	● Speechiness	● Composed in the key E
Folk/Country	● Acousticness ● Minor Key	
Experimental	● Longer duration ● Albums with variety in acousticness	
Electronic	● Instrumentalness ● Albums with variety in danceability	

Random Forest Regression

To improve on the predictive power gotten from linear regression, more sophisticated machine learning algorithms can be used

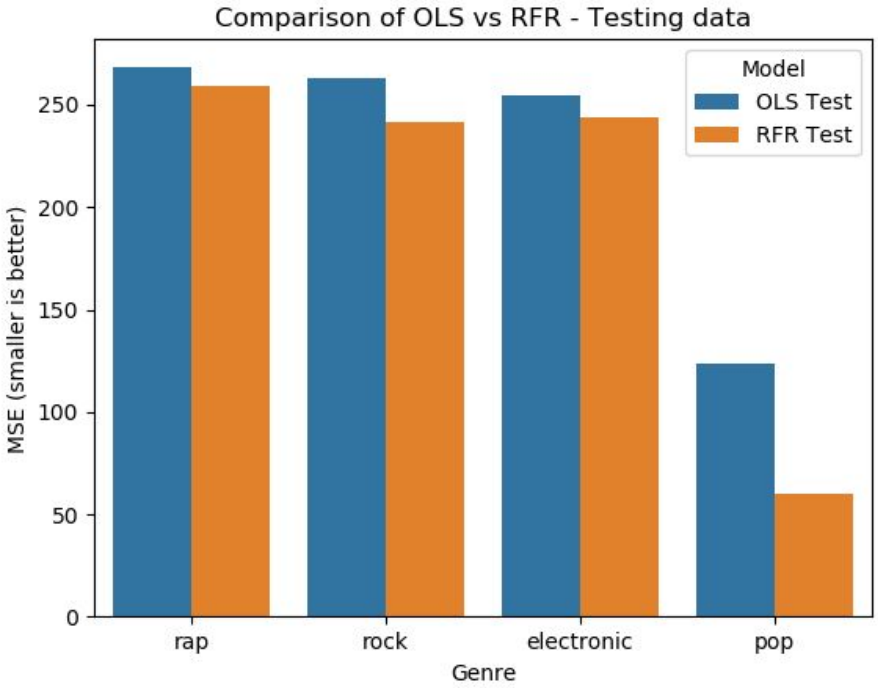
We’ll use Random Forest Regression, trained by testing a large number of different values of hyperparameters to set up the model optimally

Linear Regression vs Random Forest Regression

Mean Squared Error, calculated on test data our models were not trained on, can compare the performance of linear regression vs random forest regression

Lower values of MSE indicate a model with more predictive power

Random forest outperforms linear regression on the four biggest genres, and thus is the best choice for predictions



Data preparation for genre classification

Our next task is to use each album’s audio features to predict what genre of music the album is

To do this, we need to set up the data to have an even distribution of genres - if our data is predominately all the most common genre (rock), it will lead to an unbalanced model

Thus, we will look at the top four genres (rock, rap, electronic, pop), and sample from the most represented genres so that our training data has an even distribution

Calibrating classification models

There are three different classification albums to try, each with their own sets of hyperparameters to find the best values for our data:

Support Vector Classification: Value of C, type of kernel to use, value of gamma

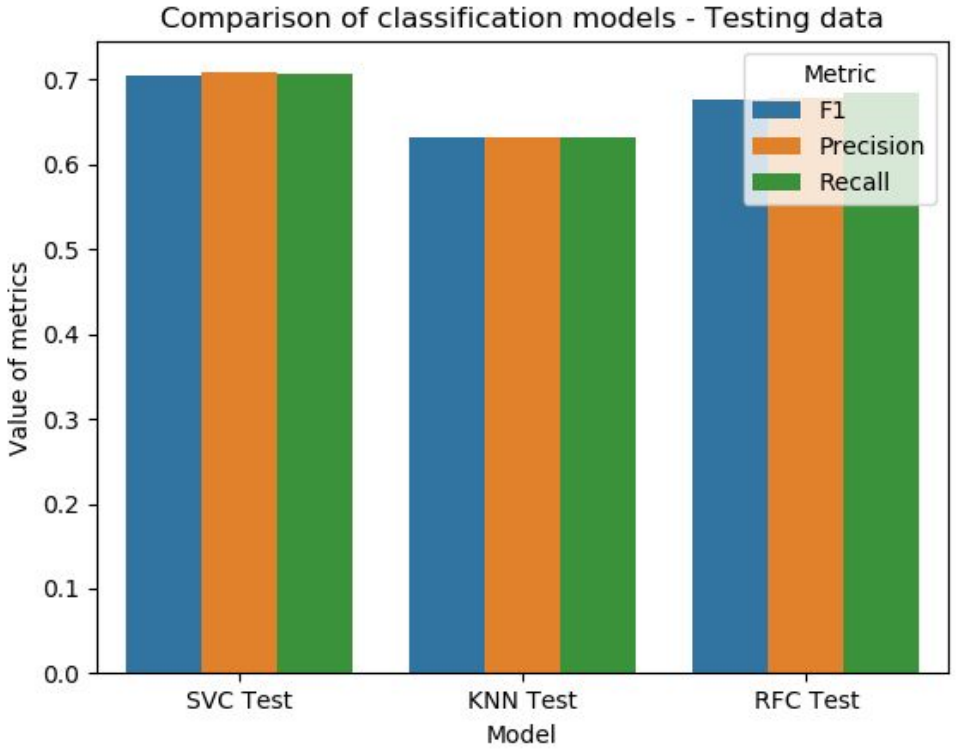
K-Nearest Neighbors: Value of K

Random Forest Classifier: number of trees, number of estimators, max depth

Comparing Classification Models

Out of our three models, it appears that Support Vector Classification has the edge in our metrics evaluating model accuracy

Next, we'll look at exactly how our model does and where it gets confused



SVC Performance

This confusion matrix shows where our model is most and least successful

Pop/R&B gives our model the most trouble - in particular, the model has a hard time telling between Pop and Rock music

Outside of Pop, the model does quite well, especially at identifying Rap

	Predicted Electronic	Predicted Pop/R&B	Predicted Rap	Predicted Rock
True Electronic	145	18	4	26
True Pop/R&B	33	91	9	61
True Rap	12	11	182	4
True Rock	17	37	2	148

Thank you!