Introduction/Business Problem

The music business has completely changed in the last 20 years. While many look at the rise of streaming as a leading factor in this, it is only a symptom of a larger shift towards live performances being the main revenue streams for musicians. As noted by Amy Wang for Rolling Stone, “Live events are quickly shaping up to be the most lucrative space for musicians in the digital-music era, and for good reason: As listeners become inundated with cheap access to music provided by streaming services, dedicated music fans crave more intimate experiences with their favorite artists.” Whereas artists used to tour in promotion of their records, records are now released as advertisements for tours. Much of the planning and promotional effort surrounding music focuses on how to ensure tours do two things: expand an artist’s fanbase, and make a profit. My capstone project will help touring and local artists, signed and independent, to book shows in the Austin, Texas area. The stakeholders will be the record labels, managers, and bands that are booking the shows. As a musician, I know first hand how difficult it can be for artists at many levels to book gigs in the Live Music Capital of the world. Being booked at a venue that doesn’t draw people who are into your genre, or performing at the wrong size venue, can make capitalizing off of the tour difficult. As an outsider, it can be difficult to know the lay of the land, and googling where to play almost never works. In my project, I will build a recommendation system that helps artists of any level find the right venues to play.

Data

For my data, I will be using three databases, two of which have robust API’s, and one that I will be interfacing with manually to create my own data frame. The first database is the Foursquare database, which I will use to find an up-to-date list of Austin venues and their metadata, including ratings, reviews, and location. I will expand upon this with another database of Austin venues from Indie on the Move, where I will gain additional insight regarding the genres a venue hosts, as well as its capacity. I will use the above information to create my own dataframe of Austin venues and carry out clustering algorithms and visualization functions. Finally, I will be interfacing with Spotify’s API in order to get information about the given client/band, including streaming figures, genre, and unique monthly listeners. I will use all of the above to find a short list of recommendable venues for the band to book a show at.

Stream of consciousness

The first step of the whole project will be collecting the data nad cleaning it up for the venues. Then I can find the relationships between all fo the selected features in the data section to create venue recommendations by size (which will be divided into a certain amount of bins based on my own personal assumptions as an artist and as someone who has booked shows) and genre to account for all of the specified genres that are listed in the venues database that I am appending to the list of venues that I get from Foursquare. The main thing will be the genre and the size of the venues, given that these are the two key features for the given artist that I will be trying to help out with the thing, but the recommendation will be DRIVEN by many of the other features such as proximity to other venues. So for the genres, there will be pre determined discreet genres that the artist has to present me. As far as the size of the band, I will simply ask them to estimate the amount of streams they have on their biggest song or something, or to give the amount of streams on their top five songs on spotify in case I can not actually get that list from the spotify API (which I SHOULD be able to, we shall see how henry can help me with this!) In that event I will need just the artist’s spotify ID. Or their name, if there is a way to look up the artist’s ID but I do not think that there is. It seems that it would be easier for artists to calculate their stream totals on their own buit I think it would look better to show interfacing with the API. OR perhaps their monthly listeners! And then use this info, based on artists and where they HAVE played in recent stops to Austin, to create a predictive model of where someone will play based on their genre and listener size. How do I determine genre concretely for an artist? Do I allow them to identify themselves within a given list of things? I guess I do not even need to know the band name. Genre and listeners is enough. Do I have to manually collect a set of artists and where they have played as its own dataframe? Representing a diverse list of genres?

So, get venue data. Then get some sample artist training data. Then weighing against certain features based on each bucket of venue size, make that recommendation. Haversine Distance. For smallest size venues, proximity to other venues is a large factor as are built in crowds. Medium small and medium venues, genre and rating are the most important. How to weight these bins actually? Local Large venues, user ratings are really important. Large venues, capacity is the most important thing to look at. I think that THREE venue recommendations would be the best way to go about this as an end deliverable to the Stakeholders.

Pearson Correlation

Spearman’s Correlation

Non Parametric vs Parametric. Parametric → test data follows some sort of distribution. T-test normal, linreg normal errors. Non Parametric → no assumptions of distributions. Usually Non P partner for every Parametric Version. Non P more conservative sometimes. P value tells how likely false positive (roughly).

Proximity rating. Hyperparameter. How many in x mile radius? Avg distance to x nearest venues? X being the hyperparameter. Play around to find a good x for linear fit.

Univariate feature selection.

Generally pick features with a high variance. Helps distinction of observations.

Genre- how genre specific is the venue. How many genres do they play that arent my genre.

1. Collect venue data.

2. Create df of concerts for predicting venue capacity based on spotify listeners.

3. Proximity Rating, creating new column in venue df.

4. “Venue Score” which will be some function of user rating, num user ratings (maybe do the log), prox rating, ~~genre specificity,~~ etc. Independent of artist. Multiply all together maybe?

5. For given artist (genre, spotify listeners), return top 3 venues that play that genre, within a 20 percent cap margin, descending order of “Venue Score.”

Foursquare API instructions:

https://labs.cognitiveclass.ai/tools/jupyterlab/lab/tree/labs/DS0701EN/DS0701EN-2-2-1-Foursquare-API-py-v2.0.ipynb?lti=true

Link:

https://www.rollingstone.com/pro/features/how-musicians-make-money-or-dont-at-all-in-2018-706745/

1. Introduction (Nothing has to go here)

1.1 Background

The music business has completely changed in the last 20 years. While many look at the rise of streaming as a leading factor in this, it is only a symptom of a larger shift towards live performances being the main revenue streams for musicians. As noted by Amy Wang for Rolling Stone, “Live events are quickly shaping up to be the most lucrative space for musicians in the digital-music era, and for good reason: As listeners become inundated with cheap access to music provided by streaming services, dedicated music fans crave more intimate experiences with their favorite artists.” Whereas artists used to tour in promotion of their records, records are now released as advertisements for tours. Much of the planning and promotional effort surrounding music focuses on how to ensure tours are booked appropriately. A large part of this comes down to selecting the proper venue.

1.2 Problem

My capstone project will help touring and local artists, signed and independent, to book shows in the Austin, Texas area. This project aims to create a recommendation model for a given musical act based on specific selected features of the act in conjunction with identified and relevant venue factors.

1.3 Interest

The stakeholders will be the record labels, managers, and bands that are booking the shows. Given the state of the world in 2020, with the COVID-19 pandemic all but shutting down live performances, labels are looking for more security in booking their shows so that they get a lot more money and do not have to worry about the performance as much. They want to control the variables and avoid booking bad shows. There are several features across paid-for platforms that help labels know what cities to tour in (Bandcamp pro allows you to map the streams and purchases of users to better find your audience, as does spotify). Once they find those cities, the process becomes a little bit more difficult to navigate, which is where this project can come in handy. I am limiting the scope of the project to the Austin area.

2. Data

2.1 Data sources

I will be using three databases, two of which have robust API’s, and one that I will be interfacing with manually to create my own data frame. The first database is the Foursquare database, which I will use to find an up-to-date list of Austin venues and their metadata, including ratings, reviews, and location. I will expand upon this with another database of Austin venues from Indie on the Move, where I will gain additional insight regarding the genres a venue hosts, as well as its capacity. I will use the above information to create my own dataframe of Austin venues and carry out clustering algorithms and visualization functions. I will use input data of band genre and monthly listeners from spotify as a user input to operate off of. I will use all of the above to find a short list of recommendable venues for the band to book a show at.

Also make training data frame to figure out what artists performed what capacities they played at to accurately make the bins.

2.2 Data cleaning

3. Methodology

3.1 Proximity rating for venues. Make Prox into a new feature.

3.2 Feature Selection for Correlation Plot. Check model redundancy and drop unneccesary features. (PCA → Principal Component Analysis. Way to check redundancy.)

3.3 Relationship between num user ratings and prox score? Redundancy?

4. Results

4.1 Regression models

4.1.1 Applying standard algorithms and their problems

4.1.2 Solution to the problems

4.1.3 Performances of different models

4.2 Classification models

5. Discussion

6. Conculsion/Future Directions