Music and Anxiety: an Examination

Data Science Major Capstone, Ayesha Tariq '24

WELLESLEY

WELLESLEY

Background

Mental health conditions are massively prevalent in society, and it is crucial to understand factors that may impact them. Music is one factor that has been studied to impact emotions, and therefore, it is important to understand its relationship with mental health. This study aims to understand the relationships between one particular mental health condition - anxiety - and its relationship with music. More specifically, it examines how different musical conditions may be related to levels of anxiety. This study aims to examine demographic characteristics as well as concepts relevant to music, such as major and minor keys, in relation to anxiety levels in humans.

Data

The data was collected from two sources. The first was from Kaggle, an online repository. This dataset consisted of several hundred responses to a survey inquiring about the respondents demographics, their individual listening habits (including their favorite genre, and their frequency of listening to various genres of music) as well as self-reported levels of four different mental health conditions: anxiety, depression, OCD, and insomnia. This second data source was from Spotify, which consisted of thousands of songs on the platform, their genre identifications, as well as characteristics of each song. Some of those characterizations included the following: the song's key, whether it was major or minor, its energy level, "danceability" level, tempo, and more. To merge the two datasets, the Spotify data was condensed into a few rows based on genre, where the characteristics of all the songs in a particular genre were averaged to produce a single value for each characteristic to represent the entire genre. Then, the datasets were joined by genre, and more particularly, by an individual's favorite genre. Each individual in the Kaggle dataset indicated their favorite genre, so for their favorite genre, the characteristics from Spotify were added.

Methods

To understand the relationship between the predictor variables (such as the demographic characteristics and listening habits of individuals, alongside the characteristics of music individuals listen to) and the outcome variable, level of anxiety, multiple linear regression models were conducted. First, models were created using only the listening habits and demographic characteristics of individuals as predictor variables. Then, to examine whether the addition of the predictor variables from the Spotify dataset may enhance the outcome of the models, a few predictor variables from the Spotify dataset were added. More specifically, the following predictor variables were used from both datasets: age, hours of music listened to per day, "instrumentalist" (whether an individual plays an instrument), "composer" (whether they are a composer), "majorness" (the amount of songs in a major key within the individual's favorite genre), "danceability," and energy level (the average energy level of songs of the individual's favorite genre).

Figure 2: Predicted and Actual values based on Model 5

	Predictors used	R^2	Adjusted R^2	BIC
Model 1	Age, Hours per day, Instrumentalist, Composer	0.06816	0.06253	1639.012
Model 2	Age, Hours per day	0.06954	0.05822	1650.142
Model 3	Age, Hours per day, Instrumentalist, Composer, energy level, danceability, majorness	0.09349	0.08247	1641.43
Model 4	Age, Hours per day, energy level, majorness	0.09628	0.07688	1657.834
Model 5	Age, Hours per day, majorness	0.09348	0.08524	1635.623
Model 6	Age, Hours per day, energy, majorness (as interaction instead of first order)	0.1246	0.08331	1693.688
Model 7	Age, Hours per day, majorness (as interaction instead of first order)	0.1099	0.09079	1652.761

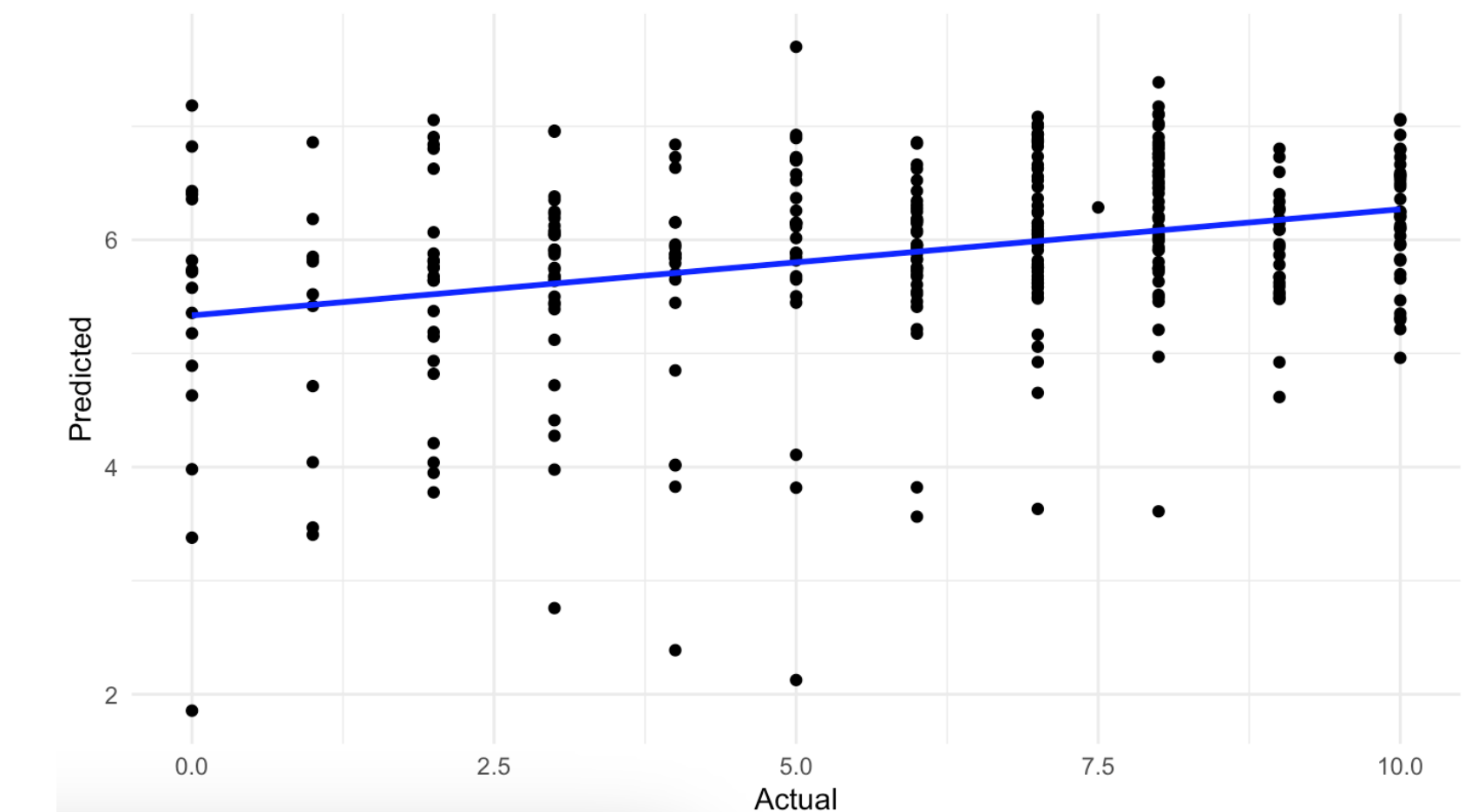Figure 1: R squared, Adjusted R squared, and BIC values for each model

Results

The analysis revealed that Model 5, which included the fewest predictors of only age, number of hours listened to music per day, and majorness, was perhaps the best model, with a lower BIC value and higher R squared and adjusted R squared values. However, none of the models had extremely strong values for R squared, adjusted R squared, and BIC. They all had quite low R squared and adjusted R squared values, and the BIC values did not differ by much.

Future Directions

In future directions, I would like to examine more of the mental health conditions within the Kaggle dataset. I would examine Depression, OCD, and Insomnia in addition to Anxiety. Further, a couple hundred data points were lost in the merging of the two datasets - any individual whose favorite genre was not represented in the Spotify dataset was removed. In order to prevent this from happening, I would like to take a different approach to merge the two datasets. Instead of merging them by an individual's favorite genre, I would like to create a score (for each characteristic, such as energy level) for each individual that represents all of the genres they listen to, not just their favorite genre. This would be done by understanding the listening frequency given for each individual and all the genres.