

Spotify Project Report

The objective of this study is to analyze a social network of Spotify artists and predict the popularity via centrality measures and machine learning methods. The motivation behind this study comes from the limitations of the common methods of understanding Spotify artists popularity so the study will help find ways to predict popularity. The dataset consists of two files which are edges.csv and nodes.csv which comprise 156303 nodes (artists) and 300385 edges (collaborations). The edges.csv contain id_0 and id_1 which represent a collaboration between two artists. The nodes.csv contain columns that give information about Spotify artists which are spotify_id, name, followers, genres, and chart_hits. The followers and popularity column along with centrality measures will be used for predictive modeling.

The dataset is made of two files: nodes.csv and edges.csv. The nodes represent the artists and the edges represent collaborations. The nodes.csv consists of columns that give information about artists. Spotify_id is a unique identifier that each artist has. The name is simply the name of the artists. The followers says the amount of the followers that each artist has. The popularity is how popular an artist is based on Spotify. The genres give a list of genres that artists perform in. The chart hits say the amount of chart hits in countries. The edges.csv contain id_0 and id_1 which indicate a collaboration between artists. Overall, there are 156303 nodes and 300385 edges. The average number of followers of artists is 86224.26 and the average popularity score is 21.16. Both distributions of popularity and followers are positively skewed showing that the majority of artists do not have a large amount of followers and popularity score. The nodes.csv has been checked for missing values and it was found to have 136724 missing values which resulted in those nodes to be dropped to have a cleaned dataset.

The dataset was loaded using PySpark by converting both nodes_csv and edges_csv into spark dataframes called nodes_csv and edges_csv respectively. For the visualization and analysis parts of this study, a subset of random 20,000

artists were selected. A subset of 20,000 was created because working with the entire dataset would have been computationally intensive and would have taken a lot of the time for the code to run the entire dataset. At first, 1000 nodes were used to test the code and ensure it works smoothly. Then 10,000 nodes were tested and it took a bit longer, so a further increase would not be too bad. It has taken around less than an hour to run the code with 10,000 nodes. The visual graph of 10,000 nodes and 4819 edges was created. The degree distribution plot is similar to the previous two distributions.

Various centrality measures were calculated to identify influential artists. Degree centrality measures the number of collaborations that an artist has. The betweenness centrality measures how often an artist is on the shortest path between two other artists. Closeness centrality measures how close an artist is to other artists. The eigenvector centrality measures how influential an artist is based on the influence of their neighbors. The PageRank measures how influential an artist is based on the number and quality of in-degrees (incoming collaboration). After calculating the centrality measures of the artists, notable results have been revealed. Pitbull appears in all 5 of the centrality measures which means that he is the most influential of the network. Steve Aoki appears in 3 of the 5 centrality measures. MC RD, G.V. Prakash, R Kelly, Rick Ross and Konshens only appear as top 5 in 2 of the 5 centrality measures.

The average degree determines how connected a network is. Although, the average degree was not calculated by code, it can easily be calculated by this formula: $(\#edges * 2 / (\#nodes))$. From the formula: $(2 * 4819) / 20000 = 0.4819$ of the subset network and the entire network is $(2 * 300385) / 156303 = 3.84$. Since the average degree of both of the entire network and subset are small, that proves that the connectivity of the network is sparse and on average, artists have a small amount of followers. The amount of connected components was found to be 8807 and the diameter of the largest connected component is 21. The edge density was also found to be $2.41e-5$. All of this further proves that the connectivity of the network is very low.

PySpark's linear regression model was used to predict popularity with followers and centrality measures. First, the centrality measures had to be converted into spark dataframes so they could be merged with `spotify_id` and `followers` since they were already spark dataframes. Afterwards, any null values were replaced by 0 or with the median value if any were created. The dataset was split randomly by 80/20. Approximately 16000 observations are in the training set while 40000 are in the testing set.

The RSME was found to be 17.65 and the r^2 to be 0.0755. The r^2 value indicates 7.55% of variance can be explained by the features of the model and the high RSME means that there is a large difference between the actual and predicted popularity. From the coefficients obtained, degree centrality has the highest positive coefficient and both betweenness centrality and PageRank have the largest negative coefficients which indicates they have a strong correlation with popularity, but the very low r^2 may disapprove and show that these are not the strong indicators.

K-means clustering was able to be applied with PySpark. First, the features were merged and then standardized. Then, the K-means algorithm was applied. The optimal number of clusters was determined by calculating the silhouette score in a range of k-values from 2 to 10. From the k-scores, k=2 has the highest silhouette score so that indicates the optimal number of clusters is 2.

In this study, the dataset was loaded and cleaned, then a subset of 20000 was taken to visualize the social network and perform network analysis to find the most influential artist which was Pitbull. Afterwards, user popularity was predicted with centrality measures and followers, but the r-value was very low, so it was not a good model to use. Finally, K-means clustering was able to show that there are 2 clusters with their own characteristics. The results of this study could be used to find artists that people like and determine which artists have the most influence that traditional methods could not.