SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL488 Business Analytics Applied Project

Project Proposal

Jan 2021 Semester

Name:	Muhammad Syafiq Bin Md Yusof
PI No.:	N1882403
TG-GROUP:	T14
SUBMISSION DATE:	16/08/2021

Here are my main comments

1) ***Topic Formulation***: I don't know what the title of your project is – I am calling it "Hit Song Science in the Singapore Market". You discussed the business analytics problem well. I understood straight away what your postulate is, and what you were challenging through your research. I think, however, you need to be a bit explicit. You said you want to study music from Singapore and you said that you are combining all the previous vectors proposed by others in determining a songs popularity. Is this correct?
2) ***Literature Review***: You had referenced the relevant and appropriate literature to defend the hypothesis you put forth; I was happy to see that it was not too much and not too little. I learnt a lot from your literature research, and you can tell from my enthusiastic responses that I cant wait to see what you will discover
3) ***Data Understanding***: Very good data breakdown, analysis and preparation. I don't have any major issues with any of it.
4) ***Proposed Modeling***: While I understand what you are trying to do, I am a bit lost as I read this about what you want to predict.
5) ***Overall Presentation***: This is a strong proposal and your results will be very interesting to read about

I think like all things, what has happened is you started strong but kind of got lost In the weeds along the way. No fault of your won; the topic is very ambitious and highly technical. My 1 suggestion is to always direct the audience back to what you are finding and to frequently recap results from previous sections. Readability = "Scoreability" in my opinion

Table of Contents

Introduction

Historically, the music industry has been dominated by superstars the likes of The Beatles and Michael Jackson. These artistes had the backing of leading music production companies which catapulted their songs to chart-topping success. In the current digital era however, the multitude of content-sharing platforms has given independent artists an opportunity fame as well. Nevertheless, for these budding young artists, there remains a question of what kind of songs they should perform to become viral.

Music consumption was thought to be an entirely subjective matter in that the likeability of each song is determined by individual preferences. The study of popular music and its features; aptly known as Hit Song Science, aims to change that notion. Hit Song Science holds the assumption that there exist a set of underlying features that allows a song to achieve popularity. There is a growing body of literature surrounding the field due to its vast range of potential commercial applications (Merlock, 2020).

Previous studies have mainly focused on the global market in determining song popularity, with findings being generalized for individual countries. I argue that the Singaporean market has unique qualities that differentiate itself from others. As a result, the features of hit songs or the mechanisms that underly their popularity might not extend to the local context. This is in part due to the multiraciality and multilinguality of the Singaporean population which has contributed towards a vibrant and diverse musical landscape. For example, Singaporean rappers Yung Raja and Faris Jabba, have each seen their singles gain success both locally and regionally (Gwee, 2020). Their songs typically blend English and Tamil or Malay seamlessly in their lyrics, resulting in a unique spin on the rap genre that would not have been as well-received on a global scale.

Commented [MK1]: ok

Commented [MK2]: ok

Commented [MK3]: good

Commented [MK4]: this is the postulate - good

Commented [MK5]: great

Commented [MK6]: This would be your hyphothesis – good!

The current paper attempts to explore the features of hit songs within the local context. It begins with a review of the existing literature, highlighting the foundations and gaps in the domain. It then details a thorough analysis and concludes with a discussion for future research.

Commented [MK7]: Right on!

Literature Review

To understand why some songs are more popular than others, we must first understand the process of music perception. Scheld, Flexer and Urbano (2013) proposed a framework of various factors that can influence music perception shown in figure 2.1.

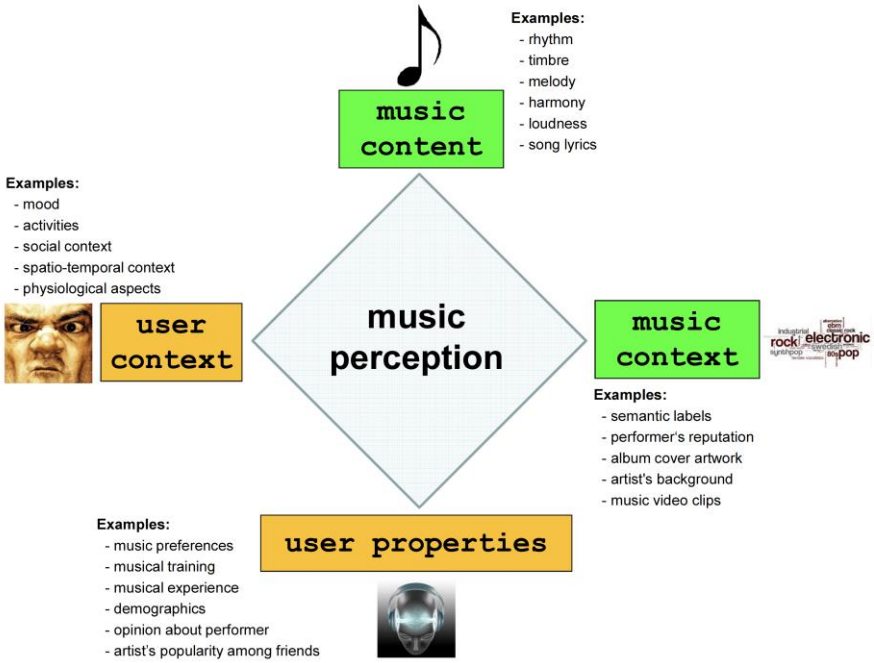

Figure 2.1 - Framework of music perception (Scheld et al. 2013)

The framework above outlines two broad categorizations of factors pertaining to the music itself and the user. Each category can be broken down further into its content; referring to elements that could be extracted directly from the audio signals of the song, and context; referring to other indirect elements surrounding the song. These content and context sub-categories exist for the user as well.

Nijkamp (2018) attempted to model the relationship between a song's audio features and its popularity. In a study using 1000 songs from various genres, the author performed a linear regression on a song's popularity; measured by its stream count on Spotify; with predictors such as the song's acousticness, tempo, and duration. The results indicated that half of the features were significantly correlated with popularity in the hypothesized directions. For example, more acoustic-sounding songs tended to be less popular. However, the relationships were generally weak. Overall, the predictors only accounted for 20.2% of the total variation, suggesting that there remain external factors that influence a song's popularity.

Another study by Pham, Kyauk, and Park (2016) conducted a more comprehensive study on predicting the popularity of songs using 2717 tracks and over 900 musical and non-musical features. The researchers built several classification models and were able to achieve a 76.2% accuracy using a Support Vector Machine (SVM) with a linear kernel. They also built regression models on the data to provide more insight into the relationships between the explanatory variables and song popularity. Their models indicated that a song's metadata, or features relating to the music context such as artist familiarity, tend to be more important than acoustic features in predicting song popularity. This finding was attributed to the information loss in reducing the variations of sounds in a song to only a single data to represent their musical qualities. The results of the study expanded upon the findings of Nijkamp (2018) and provided a reasonable explanation for the low explanatory power in his regression model.

Commented [MK8]: I guess the categories are user and music, but might be nice to spell it out for the reader?

Commented [MK9]: Can you give me some examples here? Might help me to visualize what you mean.

Commented [MK10]: This is an interesting result.

Commented [MK11]: ok

Commented [MK12]: So this supports the earlier observation by Nijkamp (2008) – interesting.

Commented [MK13]: See my earlier comment 😊

These studies did well to provide a starting point into investigating the predictors of song popularity. It is clear that musical features alone cannot represent the determinants of popular songs. They also highlighted the existing gaps in Hit Song Science. One gap that could potentially be further explored in the current research is the mechanisms that underly the relationship between a song's features and their popularity, with reference to the framework proposed by Scheld et al. (2013).

Commented [MK14]: Good, and which part of Scheld framework do you think is missing in your work?

Another group of studies took on a different approach in determining song popularity. Askin and Mauskapf (2017) argued that songs are cultural products and thus, its features are not appraised in a vacuum. Instead, it is the song's position within the larger market that determines how favourable they are evaluated by consumers. They posited that songs that are optimally differentiated – exhibiting features that are recognisable yet being different enough to avoid crowding the space with other songs, would experience more success. This hypothesis was tested by calculating each song's genre-weighted cosine similarity, as a measure of how similar their audio features were with other songs from their genre. The results of the study provided support for their hypothesis; songs that managed the similarity-differentiation trade-off well also performed better. Berger and Packard (2018) conducted a similar study, but performing analysis on the lyrics of songs instead. By computing each song's similarity in lyrical content with every other song in their dataset, they found support for the notion that songs containing lyrics atypical of their genre were more likely to achieve popularity.

However, these results were in contrast with Percino et al. (2014) who defined instrumental complexity as the variety and rarity of instruments appearing in a particular album. Their unique definition suggests that a more instrumentally complex album would sound more atypical of their genre, having been composed with a large variety of instruments that were not commonly found in music of the same genre. They found that album sales were negatively correlated with instrumental complexity – consumers preferred songs that were

Commented [MK15]: Very interesting

more familiar. At face value this, appears to contradict the findings by Auskin and Masukapf (2017) as discussed earlier. However, the authors noted that popular music has displayed homogenization over the last 5 decades, resulting in songs in the same genre becoming more formulaic as the genre sees more success. The study also did not address the similarity-differentiation trade-off to the same extent, leaving the possibility that the albums that were more instrumentally complex were too differentiated to be popular. Hence, it is unclear from this study whether musical familiarity can be considered a key factor of song popularity.

Thus far, studies in Hit Song Science have either focused solely on acoustic or lyrical content of songs. The current research aims to further build upon the findings of the aforementioned studies by combining their varied approaches in predicting song popularity. This paper presents a model of popularity based on song typicality both in terms of its music and lyrics, using music data from Singapore.

Data Understanding & Preparation

A dataset containing 3623 songs that appeared on Spotify (Singapore)'s top daily 200 hits from 2017 to 2020 was obtained from Kaggle. The dataset included 151 features, many of which were flags indicating a song's membership to a particular genre. This section details the steps taken to prepare the data for modelling.

First, it was found that the original dataset contained an excessive number of features. This was due to the levels in categorical fields such as genre, having been transformed into a tabular format. Thus, irrelevant fields were removed from the final dataset. Next, the artist_followers and days_since_released fields were found to have 1 and 63 missing data points respectively which were removed. Finally, 375 duplicate songs were discovered and subsequently discarded as well.

Commented [MK16]: In other words, they were popular but only to a sub-genre of the population – what we call colloquially "underground popular"

Commented [MK17]: Ok, so to understand what you are intending to do

1)You will be looking at Singapore data
2)You will be looking at studying all that you have mentioned above, but in the Singapore context.

Is this correct?

Commented [MK18]: There should be a reference here

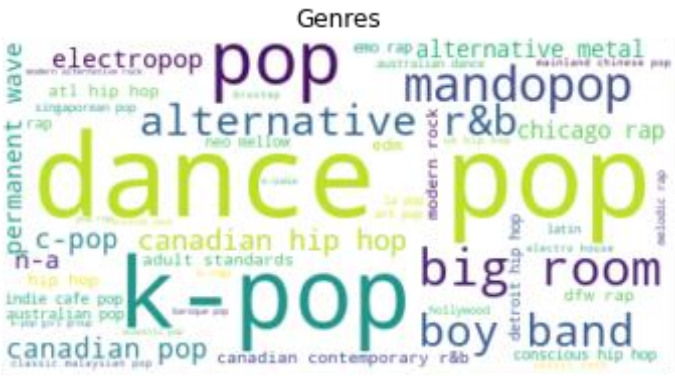

Figure 3.1 – Word cloud of Genres (original)

Next, an exploration of the genre column was conducted. Figure 3.1 shows a word cloud with each label scaled by their corresponding song counts. The genre labels provided by Spotify while detailed, were too granular for the purposes of the current analysis. There was also a large number of highly similar categories (e.g. mandopop and chinese-pop). Therefore, there was a need to reduce this number by regrouping the genres under more general categories. This process involved mapping each genre into a broader category based on certain keywords, resulting in a total of 26 newly labelled genres shown in Figure 3.2. From these 26 genres, the top 5 most popular (shown in Figure 3.3) were selected for further analysis in this paper.

Commented [MK19]: ok

Commented [MK20]: good

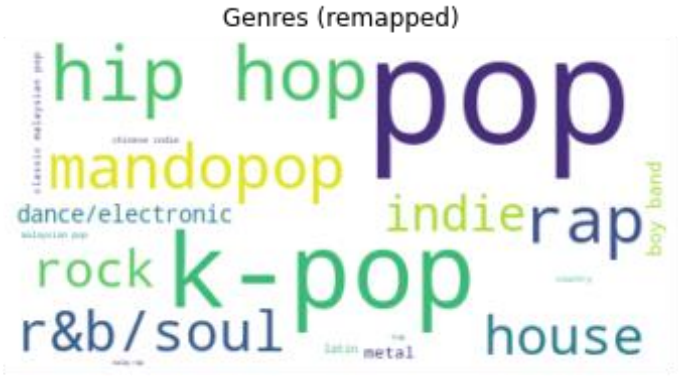

Figure 3.2 – Word cloud of Genres (Remapped)

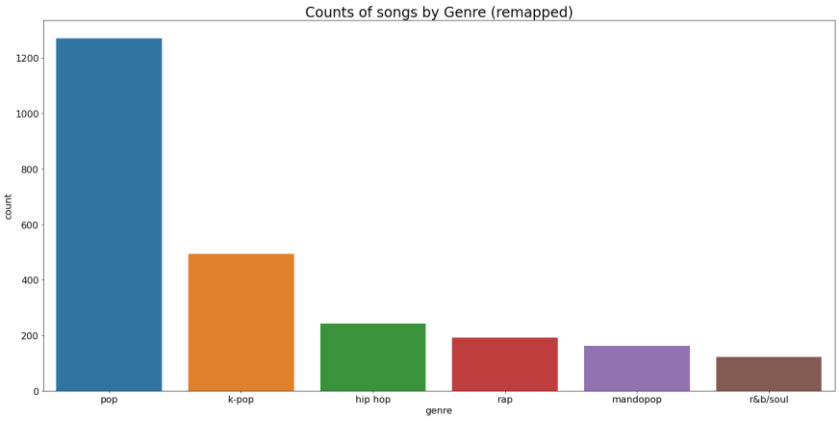

Figure 3.3- Song count of top 6 Genres (remapped)

Overall, songs of the Pop genre appear to dominate the Singaporean market with 6 of the top 10 being variations of the genre. C-pop (Chinese-pop), and Mandopop (Mandarin-pop) are well-represented in the bar plot here, reflective of the Chinese majority within the Singaporean population. K-pop (Korean-pop) also commanded the 2nd most popular genre, an unsurprising finding considering the large local fanbase.

Commented [MK21]: ok

Lyrics of the remaining 2322 songs were then retrieved via the Musixmatch API. Mandopop was replaced by R&B/Soul due to a large proportion of songs from the former not having English-translated lyrics readily available. Songs whose lyrics were not available via the API were subsequently omitted. This resulted in a final dataset containing 1919 songs with 33 features including lyrics. Figure 3.5 below shows the final song counts of each genre.

Commented [MK22]: What do you mean by this, when you say replace? You basically dropped mandopop you mean?

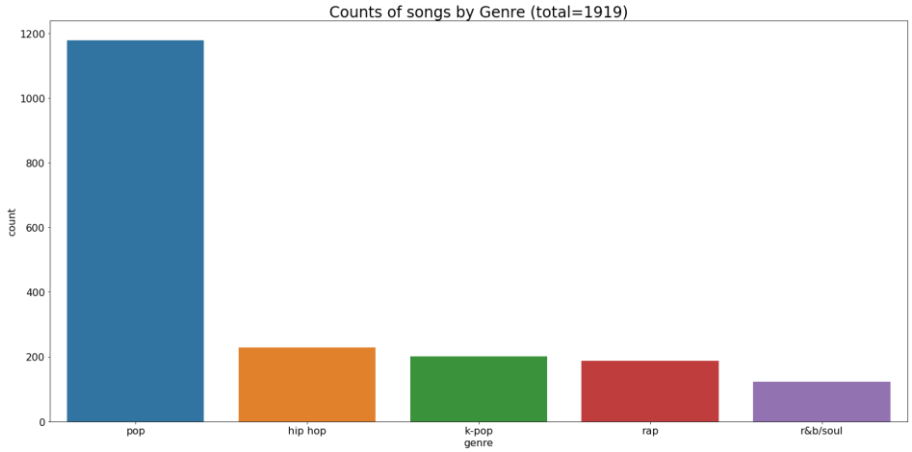

Figure 3.5 - Final song counts by Genre

Pre-processing was performed on the lyrics dataset. Figure 3.6 below outlines the general steps taken to prepare the lyrics for analysis.

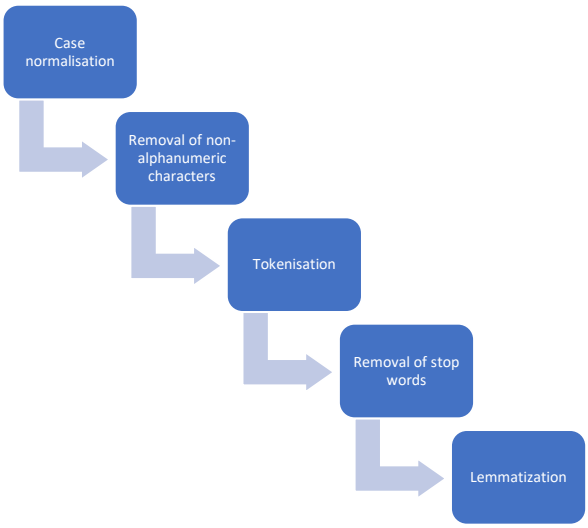

Figure 3.6 - Pre processing of song lyrics

Commented [MK23]: Maybe the analysis of lyrics should form its own subsection because the earlier section was more data wrangling of a dataframe type data set?

Figure 3.6 below outlines the general steps taken to prepare the lyrics for analysis. Text was first normalized to lower-case, with all punctuation, special characters, and digits removed. Next, tokenisation was performed to parse the text into individual terms. Stop words – extremely common words in the English language such as 'I' and 'the' that provide little meaning to the topic of the document were also discarded. This process had to be repeated many times, extending the list of stop words after each iteration to improve the meaningfulness of the remaining tokens. Finally, the text was lemmatized which involves transforming words into their root forms. For example, words in the continuous present tense such as 'walking', was reduced to 'walk' since they both convey the same basic meaning.

Commented [MK24]: Good job. What library did you utilize? If you used a different library, would this process have been simpler, more efficient? Mind you, I am just curious – just choose whatever method you want as long as it works 😊

The first variable of interest is acoustic typicality, a measure of how typical a song sounds in relation to other songs. This was imputed by taking each song's vector of 10 audio features. These features were first normalised from 0 to 1, to remove any biases from dimensions measured on a larger scale. Next, the genre weights were calculated by averaging all features of songs within genres, then finding the cosine similarities between each pair of genres. The resulting similarity matrix is shown in Table 3.1 here.

Commented [MK25]: What is this ? A frequency distribution? I don't understand

genre	hip hop	k-pop	pop	r&b/soul	rap
hip hop	1	0.981347	0.986559	0.980825	0.993894
k-pop	0.981347	1	0.992294	0.978077	0.98803
pop	0.986559	0.992294	1	0.994547	0.994867
r&b/soul	0.980825	0.978077	0.994547	1	0.991518
rap	0.993894	0.98803	0.994867	0.991518	1

Table 3.1 - Pairwise Genre Similarity

The values in the matrix indicate that the genres were mostly undifferentiated from one another. The same procedure was repeated with individual song vectors. Each song's feature vector was compared with every other song in the data to obtain their raw typicality score, then multiplied by their corresponding genre weights. The genre weights were included in line with Auskin and Masukapf (2017), they differentiate songs that belong to different genres even if their have similar sonic fingerprints. Finally, the resulting vectors were averaged to obtain the average genre-weighed typicality for each song. Figure 3.8 below shows the distribution of song typicality. The negative skew of this distribution suggests that the bulk of popular music is undifferentiated in the way they sound.

Commented [MK26]: Is this an expected result? Look forward to seeing where this leads

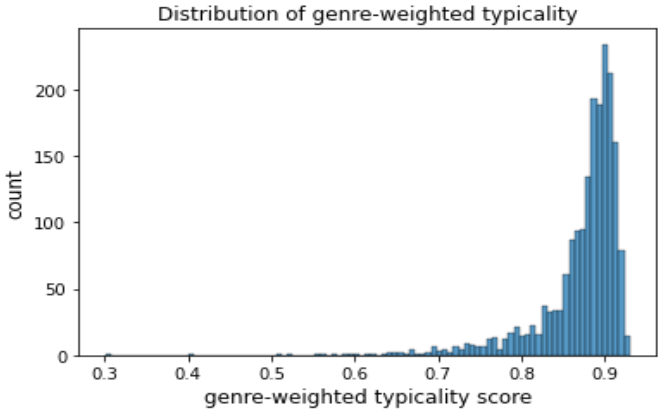

Figure 3.8 - Distrubution of song typicality

At the time of the submission of this proposal, the data preparation of lyrics is still incomplete. The final step is to transform the pre-processed lyrics will need to into a term frequency-inverse document frequency matrix. This will score each term on its frequency both within and between documents. The same process of calculating genre-weighted typicality can then be performed on the lyrics, with term frequencies representing the feature vectors instead. The average of these scores will return the lyrical typicality of each song.

Commented [MK27]: No worries

Finally, we explore the target variable; song popularity. The authors of the dataset highlighted the flaws of the original popularity dimension provided by Spotify. Spotify's measure of popularity is weighted towards songs that are more popular at the time of measurement. This meant that more recently popular songs will always be scored higher than all-time popular songs of the past that. Thus, the authors recalibrated the popularity dimension by accounting for the longevity of a song's appearance on Spotify's top 200 daily charts. They also scaled each song's popularity by an exponentially decaying weight in relation to their position on the chart, to further emphasize songs that were popular for extended periods. This resulted in a highly skewed distribution as seen in Figure 3.9. Hence, a logarithmic transformation was performed on the popularity feature to obtain a more normal distribution. Figure 3.10 shows the distribution of log transformation and Figure 3.11 shows the same distribution broken down by genres.

Commented [MK28]: Which authors? A bit lost here – you mean the Kaggle dataset?

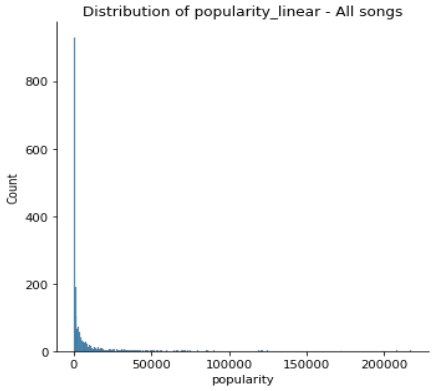

Figure 3.9 - Distribution of popularity (linear)

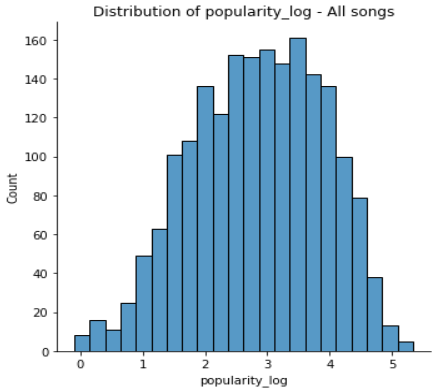

Figure 1 10 - Distribution of popularity (log)

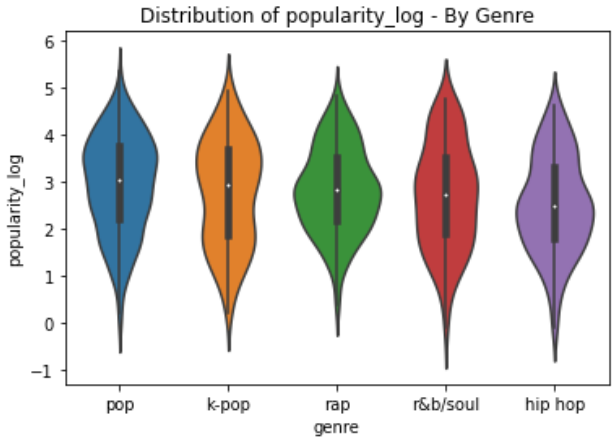

Figure 3.11 - Distribution of popularity by genre

Proposed Modelling and Evaluation

The dataset contains a total of 1919 records with 13 inputs. Since it is a reasonably sized dataset, I plan to employ a train-validate-test split as the main evaluation method. This involves partitioning the 1919 records into a training set, a validation set, and a testing set in a 60-30-10 ratio.

I propose a two-step approach for the prediction task. The first entails reframing the problem into a classification task by binarizing the popularity feature. Previous studies have employed a similar approach in selecting and adjusting to find the optimal threshold for a song to be labelled as 'popular'. Next, I can build a random forest to classify songs into hits vs non-hits based on the input features. A random forest is suitable here because it is a non-parametric model that does not assume the underlying distribution of the data. Random forests are also able to reveal a hierarchy of feature importance which will be useful in improving the model's performance.

Commented [MK29]: You lost me a bit here but I am not sure what you are trying to predict.

Commented [MK30]: I would say you need to buold maybe 2 or 3 different models and determine which is best, before being too definitive here.

The important features highlighted by the random forest can then be used as predictors in a polynomial regression model. A polynomial regression is a special variation of a standard linear regression, used when the relationship between a target variable and its predictors is not linear. This model is chosen here as I expect song popularity to have an inverted U-shape relationship with song typicality; songs that are positioned optimally on the similarity-differentiation spectrum are more likely to be popular.

Commented [MK31]: Although its not required, you should have a small conclusion with maybe a small timeframe, for the benefit of your reader.

References

Askin, N., & Mauskapf, M. (2017). What makes popular culture popular? Product features and optimal differentiation in music. *American Sociological Review*, *82*(5), 910-944.

Berger, J., & Packard, G. (2018). Are atypical things more popular?. *Psychological Science*, *29*(7), 1178-1184.

Gwee, K. (2020, October 12). *Yung Raja: Singaporean HIP-HOP STAR sparks joy with dizzying Tamil and ENGLISH RAPS*. NME. https://www.nme.com/en_asia/features/yung-raja-singaporean-hip-hop-star-tamil-and-english-raps-dance-song-interview-2020-2770593.

Merlock, F. (2020). *The Valuation of Songwriting Techniques: An Analysis of How Song Elements Affect Song Value* (Doctoral dissertation, University Honors College Middle Tennessee State University).

Nijkamp, R. Prediction of product success: explaining song popularity by audio features from spotify data, July 2018. *URL http://essay. utwente. nl/75422*.

Percino, G., Klimek, P., & Thurner, S. (2014). Instrumentational complexity of music genres and why simplicity sells. *PloS one*, *9*(12), e115255.

Pham, J., Kyauk, E., & Park, E. (2016). Predicting song popularity. *nd): n. pag. Web*, *26*.

Schedl, M., Flexer, A., & Urbano, J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, *41*(3), 523-539.