ANL488 FINAL REPORT

Automatic Classification of Hit Songs on Spotify using Acoustic and Topical Features

Submitted by

Name: Muhammad Syafiq Bin Md Yusof

PI Number: N1882403

SCHOOL OF BUSINESS

Singapore University of

Social Sciences

Presented to Singapore University of Social

Sciences in partial fulfilment of the

requirements for the Degree of Bachelor of Science

in Business Analytics

2021

Commented [MK1]: Well written, good length and pace with very logical arguments

Your conclusions and discussion is really well written and really brought the whole values of the thesis together I think. I don't have anything majorly negative to say.

Well done

Table of Contents

Abstract

This paper attempts to classify hit songs from non-hits using acoustic and lyrical features of songs. Three experiments were conducted with datasets containing music of varying geographic scopes and degrees of popularity. Measures of song typicality, both acoustically and topically, and their respective relationship to commercial success were of particular interest. Acoustic typicality was found to contribute to song popularity, although the specific direction of its impact changes based on other songs in the dataset. Additionally, differences in features that make for a hit song between the local and global markets were observed.

Commented [MK2]: good

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 to 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 a song is solely 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 exists 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).

Commented [MK3]: good

Previous studies have mainly focused on the global market in determining song popularity, with findings being generalized for individual countries. This presents the argument 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 [MK4]: intersting

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 [MK5]: good

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 factors that can influence music perception shown in Figure 1.1.

Commented [MK6]: ok

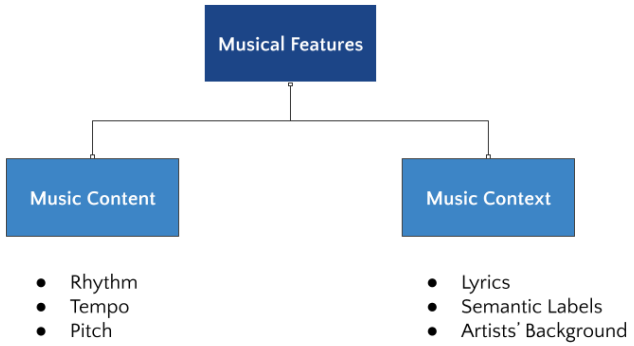

Figure 1.1 - Framework of music perception (Scheld et al. 2013)

Commented [MK7]: good

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.

Commented [MK8]: ok

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 [MK9]: ok

These studies did well to provide a starting point into investigating the predictors of song popularity. It is clear that musical features alone do not constitute all 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 [MK10]: ok

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 favourably 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.

Commented [MK11]: interesting

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 implies 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 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

Commented [MK12]: good

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.

Commented [MK13]: controversy – love it 😊

Thus far, studies in Hit Song Science have focused solely either 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 song popularity based on acoustic, topical, and typicality features, using music data from Singapore.

Commented [MK14]: good

Data Understanding & Preparation

Commented [MK15]: good, well written and documented

A dataset containing 3623 songs that appeared on Spotify (Singapore)'s top daily 200 hits from 2017 to 2020 was obtained from Kaggle (Pepe, 2021). The dataset included 151 features, many of which were binary flags (0/1).

Data Cleaning

It was found that the original dataset contained an excessive number of features. 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 duplicates and 553 outliers were discovered and subsequently discarded.

Next, it was found that genre labels provided by Spotify while detailed, were too granular for the purposes of the current analysis. Therefore, there was a need to reduce this number by regrouping the genres under more general categories, resulting in 26 genres. From these 26 genres, the top 10 most popular shown in Figure 2.1.1, were selected for further analysis in this paper.

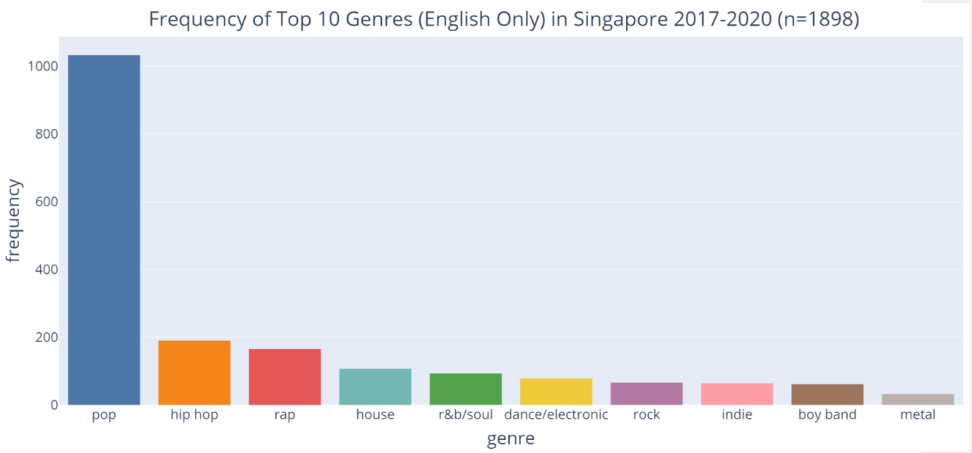

Figure 2.1.1 - Frequency of top 10 Genres excluding Foreign Languages (remapped)

Overall, songs of the Pop genre appear to dominate the Singaporean market, making up about 54.4% of songs in the data, which may be attributed to their simple melodies and lyrics that make them appealing to a wider group of listeners (Boyle et al., 1981). Non-English genres such as Mandopop (Mandarin-pop) and K-pop (Korean-pop) were also well-represented in the data but were excluded as the current study involves analysis of song lyrics.

Acoustic Features

The dataset contained 11 acoustic features that capture musical qualities of a song. These features and their descriptions can be found in Table 5.1 in the Appendix. Min-max normalization was performed on these variables to ensure uniformity of scale across all variables.

Text pre-processing of Lyrics

Lyrics of the remaining 1898 songs were then retrieved via the Musixmatch API. It should be noted that these were only snippets, or 30% of the full lyrics that were made available on the non-commercial license of the API.

Commented [MK16]: good caveat on the limitations

These lyrics then underwent text pre-processing; a process to represent text as numbers for meaningful analysis. Figure 2.2.1 below outlines the general steps taken to prepare the lyrics for analysis.

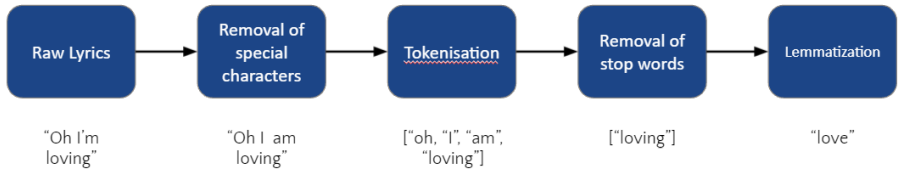

Figure 2.2.1 - Pre processing of song lyrics

The pre-processed lyrics were then transformed into a document-term matrix, resembling the one shown in Table 2.2.2 below. The values in the matrix represent the frequency of each term per document.

Document No.	"love"	"ice"	"tonight"
0	5	0	2
1	3	0	2
2	0	3	0

Table 2.2.2 - Document-term matrix

Topical composition of Lyrics

Topic modelling was performed on the TF-IDF using Latent Dirchlet Allocation (LDA). LDA assigns topic probabilities to each document, thus providing a general representation of topics found in each document (Blei et al, 2003).

LDA requires the user to specify the hyperparameter *k*, where *k* represents the number of topics to be extracted. Thus, 5 LDA models were built with *k* ranging from 6 to 10. The topics extracted from each model were then evaluated based on their coherence scores. The

coherence score measures the co-occurrence of the most frequent words per topic on a scale of 0 to 1; a higher coherence score implies more well-defined topics (Minno et al., 2011). Figure 2.3.1 below displays the coherence score for each LDA model.

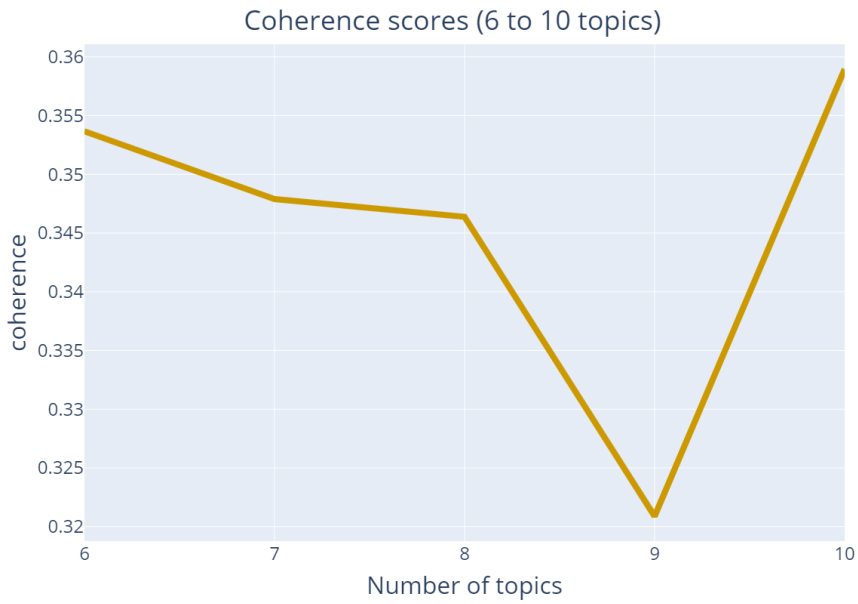

Figure 2.3.1- Coherence scores of k (6 to 10)

As evidenced by the coherence scores, *k*=10 appears to be the optimal number of topics. However, the extracted topics were found to be vague and thus, lacking in interpretability. Therefore, *k*=6 was selected as the next most optimal parameter, with the extracted topics observed to be much easier to interpret. The 6 topics were then inspected based on their most frequent words and assigned labels that reflect their meanings. Table 2.3.2 below presents the labels of the extracted topics and their most frequent words.

No.	Topic Label	Frequent terms
1	Romantic Love	“love”, “feel”, “heart”
2	Uncertain Love	“never”, “girl”, “love”
3	Wealth	“diamond”, “money”, “drip”

4	Anger/Aggression	“shit”, “bust”, “burn”
5	Sadness	“low”, “run”, “lonely”
6	Hope & Ambition	“motivate”, “bring”, “dream”

Table 2.3.2 – Frequent Terms of Extracted Topics

Lastly, each song was then measured on the degree of membership to each of the 6 topics, resulting in a topical composition matrix shown in Table 2.3.3 below. It should be noted that the sum of each song’s topical composition vector need not be equal to 1 as each feature is scored independently.

Document No.	Romantic Love	Uncertain Love	Wealth	Anger	Sadness	Hope
0	0.81	0.13	0	0	0	0
1	0	0	0.76	0.39	0	0
2	0	0.33	0	0.12	0.85	0

Table 2.3.3 – Topical composition matrix

Acoustic and Topical Typicality

Two new features were derived using the data prepared, namely acoustic typicality and topical typicality. The former measures the similarity in each song’s audio, while the latter measures the similarity in terms of the song’s lyrics, in relation to all other songs in the data.

With reference to Auskin and Mauskapf’s (2017) research discussed in the literature review section earlier, Figure 2.4.1 details the steps taken to compute the acoustic and topical typicality measures.

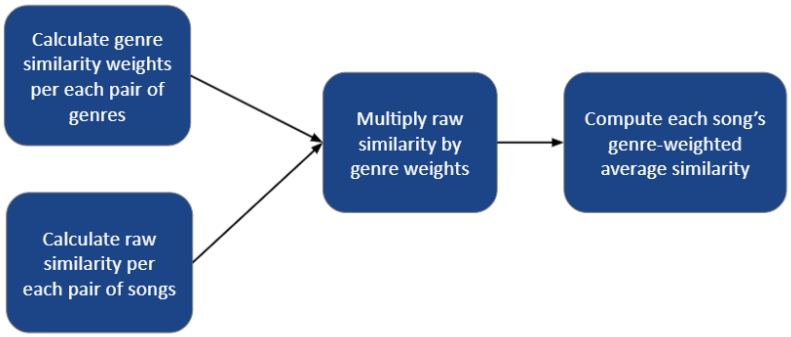

Figure 2.4.1- Deriving genre-weighted average similarity

First, the genre weights were calculated by averaging all features of songs within genres, then finding the cosine similarities between each pair of genres. A subset of the resulting similarity matrix is shown in Table 2.4.2 here.

Genre	hip hop	house	pop	r&b/soul	rap
hip hop	1	0.9626	0.9828	0.9776	0.9905
house	0.9626	1	0.9889	0.9707	0.9825
pop	0.9828	0.9889	1	0.9946	0.9949
r&b/soul	0.9776	0.9707	0.9946	1	0.9808
rap	0.9906	0.9825	0.9949	0.9808	1

Table 2.4.2 - Pairwise Genre Similarity Matrix

All values in the matrix were close to 1, indicating 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 to potentially differentiate songs that appear similar but do not belong to the same genre. Finally, the resulting vectors were averaged to obtain the average genre-weighed typicality for each song.

Commented [MK17]: ok

This entire process was repeated twice to obtain the average acoustic and topical typicality, whose distributions are plotted in Figure 2.4.3 below.

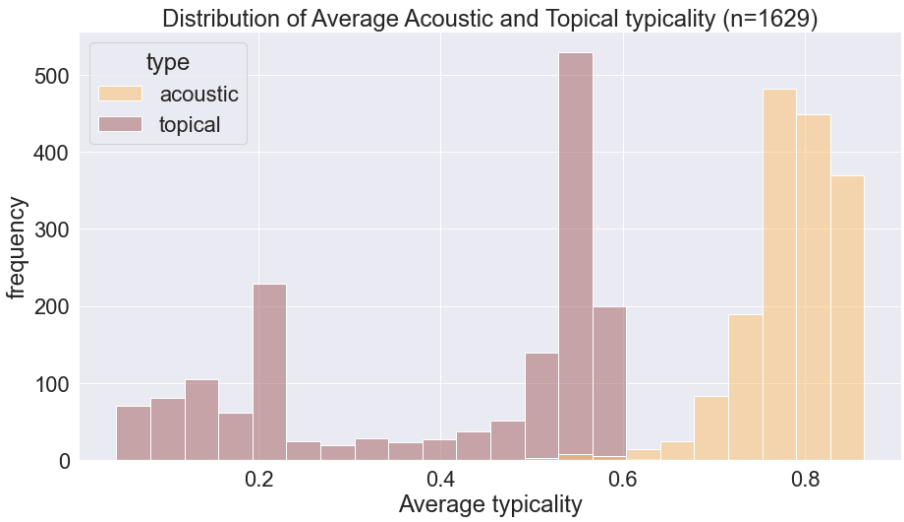

Figure 2.4.3 – Distributions of Acoustic Typicality and Topical Typicality

The distributions reveal that on average, songs are more similar to each other in terms of their sounds as compared to their lyrics. The negative skews present in both distributions suggest that the majority of songs are highly similar. A similar pattern was also found in Askin and Mauskapf's (2017) study.

Commented [MK18]: interesting

Commented [MK19]: ok, so your results are corroborated by literature. good

Song Popularity

Finally, we explore the target variable – song popularity. As mentioned previously, the dataset contains songs that have appeared on the Spotify's Daily Top 200 Charts in Singapore from 2017 to 2020. The separation between more popular vs less popular songs was made at the top 50 position, i.e. songs that have achieved a spot on the Top 50 (more popular) vs songs that have not (less popular). Figure 2.5.1 below shows the proportion of these classes.

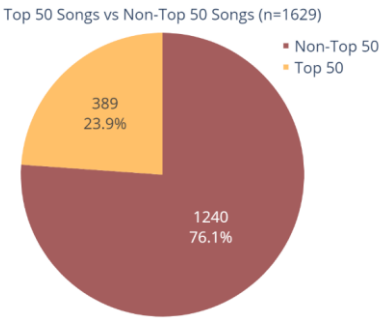

Figure 2.5.1 – Distribution of Top 50 vs Non-Top 50 Songs

It was observed that a class imbalance was present with the majority class, i.e., more popular songs, representing 76.1% of all songs. An imbalanced dataset may be problematic for classification models. These concerns will be addressed in the following section, along with the measures to mitigate their effects.

Commented [MK20]: good

Modelling

The final dataset contained 1629 songs and 58 features, although only 20 of these were used as inputs. Figure 3.1.1 below shows the set of input and target variables.

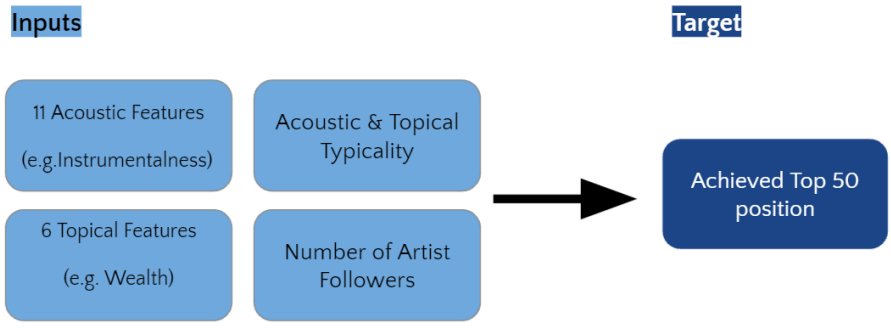

Figure 3.1.1 – Inputs and target variables

Number of Artists followers represents the number of Spotify users that subscribe to the Artist's music on the application, used as a measure of the artist's popularity.

Additionally, 2 more experiments were conducted, each with different datasets. Table 3.1.2 below summarises the parameters in each experiment.

Experiment	Dataset	Size	Target
1	Daily Top 200 Songs in Singapore (2017-2020)	1629	Top 50 vs Non-Top 50 Songs
2	Daily Top 200 Songs Globally (2017-2020)	11854	Top 50 vs Non-Top 50 Songs
3	Daily Top 100 Songs Globally on Spotify/Billboard/Shazam, and Non-Top 100 Songs (1985-2021)	2377	Top 100 vs Non-Top 100 Songs

Table 3.1.2 – Parameters of Experiments 1 to 3

Experiment 2 was included as a control group to Experiment 1; to potentially highlight any unique traits Singaporean listeners might have in terms of music consumption. Experiment 3 used a dataset (Theodoropoulos, 2021) containing songs that have been highly popular (Top 100) and songs that were never popular (Non-Top 100). This is in contrast to Experiments 1 and 2 which attempted to classify songs between extremely popular (Top 50) and slightly less, albeit still popular (51st to 200th). The hypothesis for this experiment was that with a larger separation between the categories, model performance would improve, and the effects of the inputs would be more pronounced. It should also be noted that Experiments 2 and 3 did not make use of any topical features including topical typicality as inputs as many songs were written in foreign languages.

Commented [MK21]: so this was a smaller dataset than experiment 2?

Commented [MK22]: I think this could have been more elegantly written. It's a bit confusing as it is

Model Evaluation Method

Three models were selected for comparison, each representing a particular type of model shown below in Table 3.2.1.

Model	Model Type	Description
Logistic Regression (LR)	Parametric	A variation of Linear Regression used when the target variable is categorical instead of numeric. Parametric because it assumes a linear relationship between inputs and log(odds) of the target.
K-Nearest Neighbours (KNN)	Non-parametric	Groups observations into *k* groups based on distance measures. Non-parametric as it does not assume distribution of data.
Extreme Gradient Boosting (XGB)	Ensemble (Boosting)	Sequentially builds decision trees that learns from the mistakes of previously built trees. An ensemble model as it combines many weak learners into one strong learner.

Table 3.2.1 – Descriptions of Selected Models

As mentioned in the previous section, the imbalance in classes can hinder classification performance (Ling & Sheng, 2010). Oversampling and undersampling are two techniques that can address this problem. Oversampling, specifically with Synthetic Minority Oversampling Technique (SMOTE), artificially inflates the size of the minority class by duplicating its instances. Conversely, undersampling removes instances of the majority class that are highly similar to those of the minority class (Yap et al., 2013). This creates a larger separation between the categories with the aim of improving the classification performance.

Commented [MK23]: good

These models were then evaluated on 3 performance metrics, namely Precision, Recall and F1-scores. Precision refers to how well the model is able to capture true hit songs amongst

those that it predicts to be hits. A high precision rate means that of all songs predicted to be hits, there is a high degree of certainty that they are true hits. Recall refers to how well the model is able to capture true hits amongst actual hits. A high recall rate means that there are very few true hit songs that were left undetected. The present study does not consider either to be more important hence, the inclusion of an F1-score. Formally, the F1-score is the harmonic mean of the precision and recall rates. More simply, it is a performance metric that equally balances both precision and recall.

The modelling evaluation process is as follows:

1. For each type of model (LR, KNN, and XGB), build 3 models, with No Resampling (-NR), Undersampling (-U), and Oversampling (-O).
2. For each model type, compare and select candidate model based on Precision, Recall & F1-score.
3. Select overall best model from candidate models.

Figure 3.2.2 below summarises the modelling evaluation process.

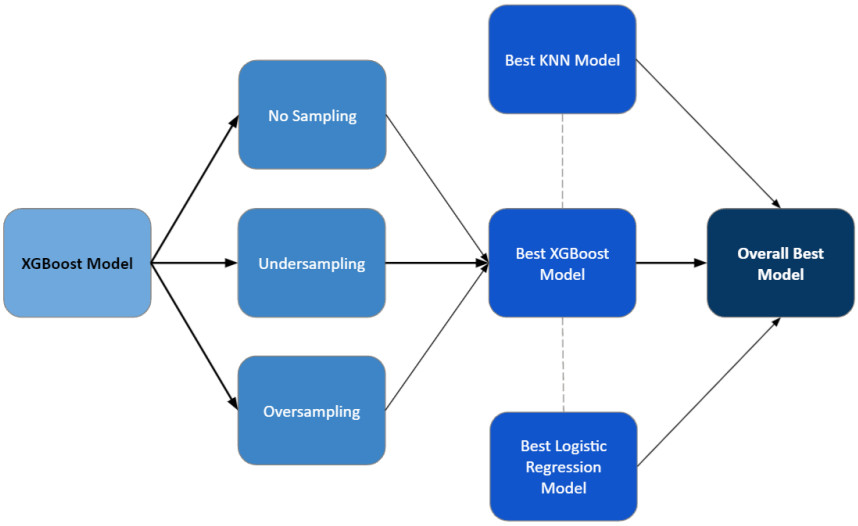

Figure 3.2.2 – Overview of Model Evaluation Process

Commented [MK24]: good

Results

Table 4.1.1 below presents the performance metrics for each model per experiment. Figures highlighted in blue represent the candidate models, while those in orange denote the overall best model for each experiment. As shown below, the XGB-U model performed the best in all 3 experiments.

Model	Resampling	Experiment 1			Experiment 2			Experiment 3		
		Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
LR	None	0.2784	0.5593	0.3711	0.4273	0.6796	0.5243	0.7066	0.7506	0.726
	Under	0.278	0.5668	0.3723	0.4244	0.6932	0.526	0.7389	0.7226	0.7281
	Over	0.2654	0.5374	0.3545	0.4211	0.6807	0.5196	0.8684	0.6374	0.7326
KNN	None	0.3356	0.0811	0.1292	0.5435	0.3307	0.4079	0.6968	0.819	0.7492
	Under	0.1474	0.337	0.2025	0.5324	0.4156	0.4642	0.7315	0.769	0.746
	Over	0.2921	0.6911	0.41	0.433	0.6818	0.5292	0.8856	0.6671	0.7596
XGB	None	0.247	0.9597	0.3928	0.5595	0.522	0.5383	0.8532	0.7153	0.7758
	Under	**0.2463**	**0.9745**	**0.3932**	**0.4288**	**0.7192**	**0.5363**	**0.878**	**0.7077**	**0.7825**

	Over		0.2588	0.8276	0.3941		0.4337	0.673	0.522		0.878	0.6746	0.7612

Table 4.1.1 – Summary of Performance Metrics from Experiments 1-3

Commented [MK25]: good

Commented [MK26]: good result and expanation

Experiment 1 – Top 50 Songs in Singapore

In experiment 1, the objective was to build a classification model that could predict whether a song has achieved peak popularity by appearing on Singapore's Top 50 position at least once. LR-U, KNN-O and XGB-U were the candidate models for this experiment.

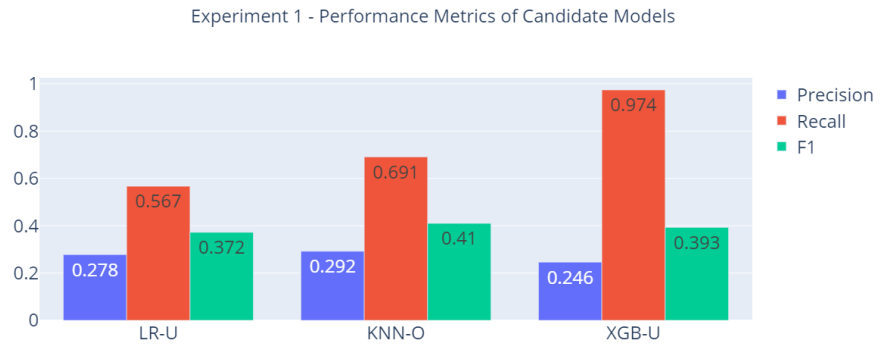

Figure 4.2.1- Performance metrics of models in Experiment 1

Figure 4.2.1 above presents the Precision, Recall and F1-scores of the 3 candidate models. LR-U was not comparable to the others due to its much lower Recall rate. While KNN-O had the highest F1-score of 0.41, XGB-U outperformed it on Recall by a substantial margin (+0.25), with only a slight cost to its Precision (-0.05). This meant that the XGB-U misclassified slightly more Non-Top 50 songs as hits, but correctly captured a much larger proportion of true Top 50 songs. Since the massively improved Recall rate outweighed the marginally lower Precision, XGB-U was selected to be the overall best model for this experiment.

The confusion matrix of the XGB-U model is presented below in Figure 4.2.2. The matrix compares the actual vs predicted hits and non-hits.

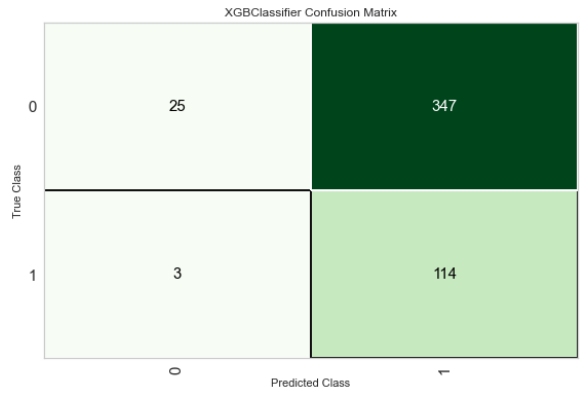

Figure 4.2.2- Confusion matrix of XGB-U model in Experiment 1

The matrix indicates that the model clearly favours hit songs, with very few predictions of non-hits, thus resulting in the high recall and low precision rates. The model appears to classify most songs as hits by default, which is unexpected given that the hit songs represent the minority class. This behaviour is made more evident when inspecting the Shapley Additive Explanations (SHAP) values for the XGB-U model shown in Figure 4.1.3.

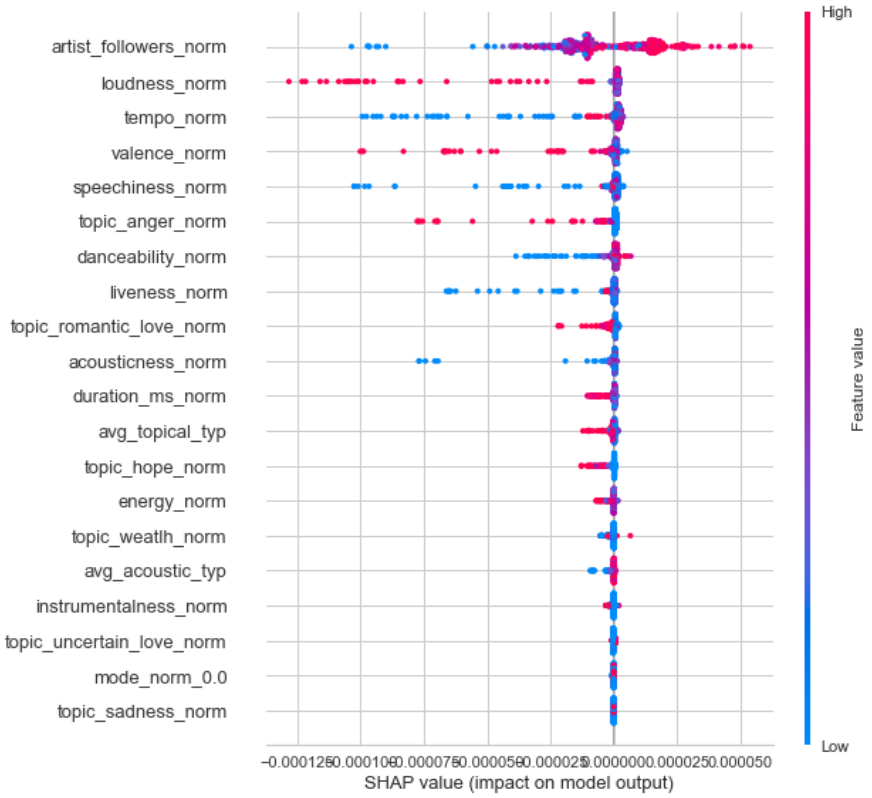

Figure 4.2.3- SHAP values of features from XGB-U Experiment 1

The SHAP values provide insight into each feature's contribution to the model's prediction (Merrick & Tally, 2020). The x-axis on the graph denotes the effect on the log-odds of a song has achieved the Top 50 position. The y-axis lists the features ranked by importance in descending order, from top to bottom. The most important features were number of artist followers and loudness, while mode and the topic of sadness were the least important. The SHAP values also reveal the impact of high (in red) and low values (in blue) per feature on the probability of a song being a Top 50 hit. For example, the plot indicates that songs by artists with a large following were more likely to be hits, since most red data points lie on the positive side of the x-axis, and vice-versa for the blue data points. This is in

line with general intuition, as it is reasonable to expect songs by popular artists to be popular as well. Besides number of artist followers, no other feature contributed positively to the probability of a song being a hit. This can be seen with most features' data points either clustering around the mean or scattered towards the negative side of the x-axis. This observation dovetails the results seen in the confusion matrix, explaining why the model favours hit songs so heavily.

Commented [MK27]: basuically the prediction is incorrect is what you are saying. A lot of false negatives

Overall, the models derived from this experiment performed poorly. While the best model obtained a decent Recall rate, Precision and F1 scores were consistently low across all models.

Experiment 2 – Top 50 Songs Globally

In experiment 2, the aim was to classify songs into Top 50 vs Non-Top 50 in the global market, thereby surfacing any traits that might be unique to Singaporean consumers. The same models from experiment 1 were selected as candidate models, with their performance metrics shown in Figure 4.3.1 below.

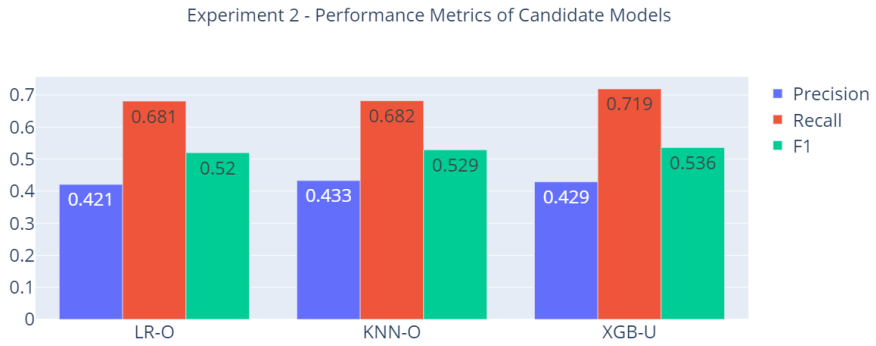

Figure 4.3.1 – Performance metrics of Candidate Models in Experiment 2

Again, XGB-U was selected as the best model among the candidates, having the best F1-score and recall. Generally, there were significant performance gains, specifically in precision rates from the previous experiment. Performance across the models was more consistent this time, with smaller differences between them. The improved performance may be attributed to a larger dataset, being twice as large of that from experiment 1. Despite these improvements however, the precision rates were still underwhelming, with the highest of the models only boasting a 0.433.

Commented [MK28]: ok

The confusion matrix of the XGB-U model shown in Figure 4.3.2 further highlights the improvements in the model. There is a much smaller proportion of non-hits misclassified as hits (39.9%), compared to the 93.3% observed in experiment 1.

Commented [MK29]: this is good

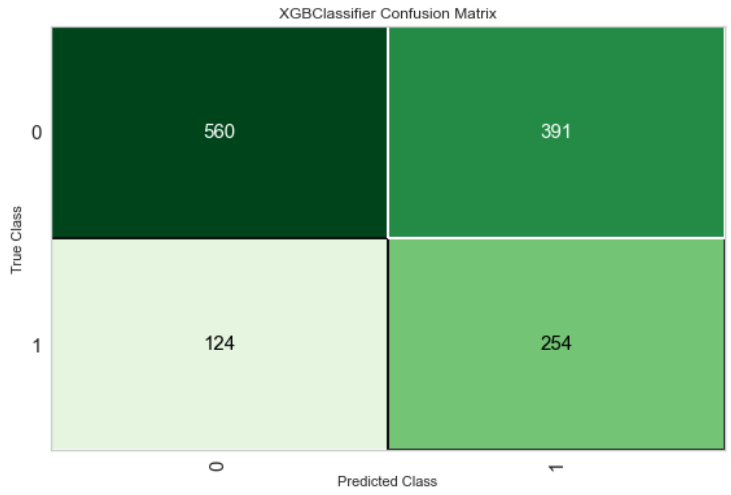

Figure 4.3.2 – Confusion Matrix of XGB-U model in Experiment 2

From the SHAP values presented in Figure 4.3.3 below, a much greater spread of data points is observed along the x-axis across all features.

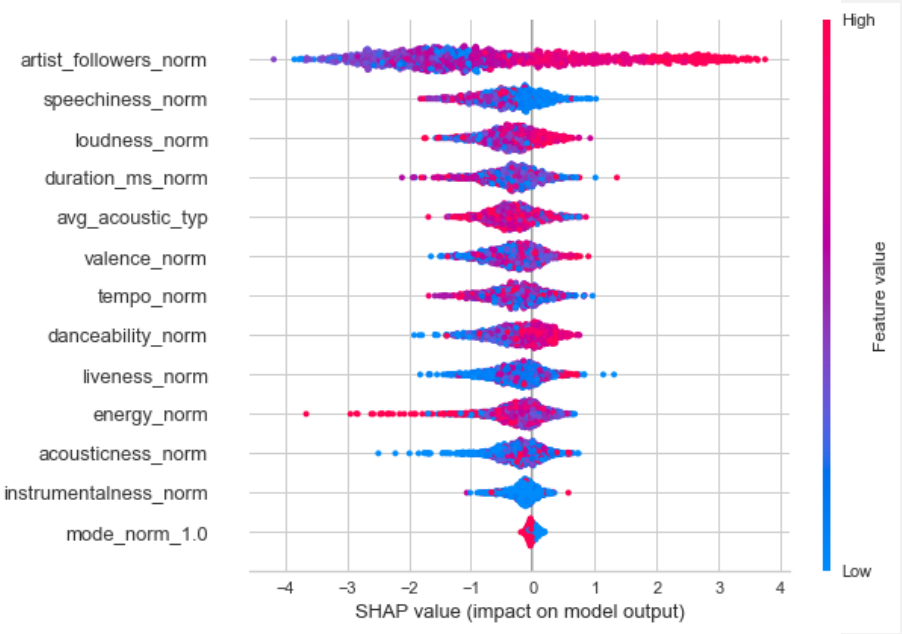

Figure 4.3.3 – SHAP values of features from XGB-U model in Experiment 2

Number of artist followers remains the most important feature, with hits more likely to be sung by more popular artists. Speechiness and loudness were also indicated to be important features. Speechy tracks such as those that contain poems, or spoken word were less likely to be hits, while louder tracks tend to be more popular. Additionally, more typical songs were more likely to be hits, although this result is less obvious due to most songs having high typicality, as seen in the mass of data points in red along that dimension.

The results of this experiment differed significantly from those in experiment 1. The same features that were not useful in classifying between hits and non-hits in the Singaporean market, proved to be more important in the global market. One possible reason for this observed contrast is due to a difference in preference of popular music. Figure 4.3.4 below presents the difference in acoustic composition between hit songs that were only popular in the Singaporean market and those that were only popular in the global market.

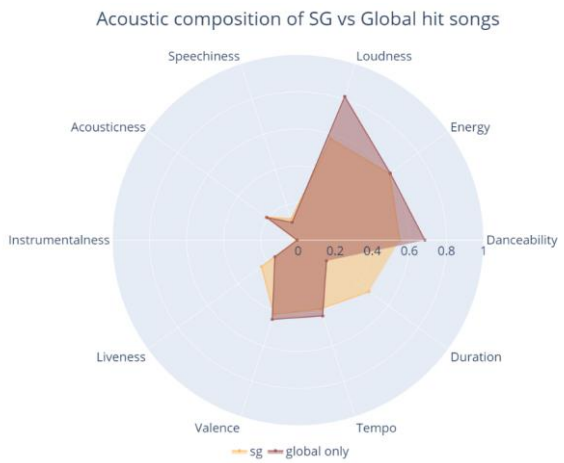

Figure 4.3.4 – Acoustic Composition of Singaporean-only vs Global-only hit songs

The chart suggests that generally, hit songs exclusive to the Singaporean market tend to be more moderate in their acoustic composition, particularly in the dimensions of loudness and danceability, whereas hit songs in the Global market tend to skew towards the extremes. The tendency for hit songs in Singapore to be closer to the middle values, thus less differentiated from non-hits, might explain why the models in experiment 1 struggled in the classification task.

Commented [MK30]: this is an interesting observation.

Experiment 3 – Top 100 Songs Globally

In experiment 3, classification was performed on songs that have appeared on the Top 100 charts vs songs that have never appeared on any charts globally. It should be noted that this is a significant change in the data used in experiments 1 and 2. In the earlier experiments, classification was done on extremely popular songs (Top 50) vs slightly less popular (Top 51st to 200th). The hypothesis here is that model performance will improve with the larger separation between the classes. The performance metrics of the candidate models are shown in Figure 4.4.1 below.

Commented [MK31]: Good

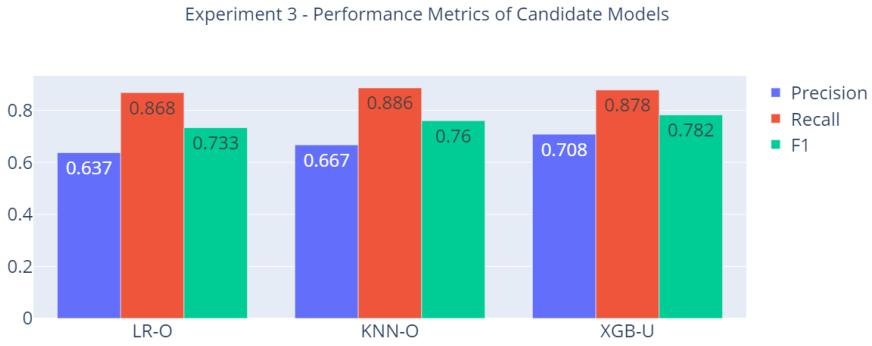

Figure 4.4.1 – Performance Metrics of Candidate Models in Experiment 3

In line with the hypothesis, there were significant improvements in the absolute values and consistency metrics across the candidate models. The confusion matrix of the XGB-U model shown in Figure 4.4.2 below highlights how much better the model is at classifying between hits and non-hits. There is a significantly smaller proportion (10.3%) of non-hits being misclassified from previous experiments (>= 39.9%)

Commented [MK32]: excellent

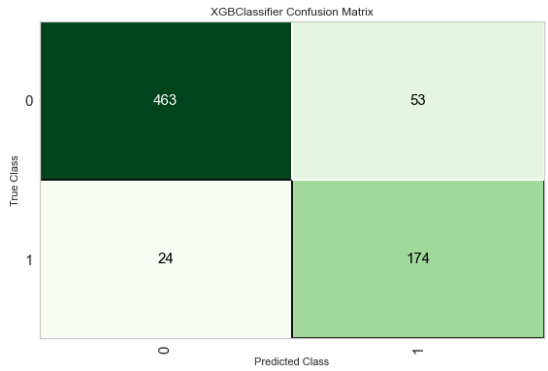

Figure 4.4.2 – Confusion Matrix of XGB-U model in Experiment 3

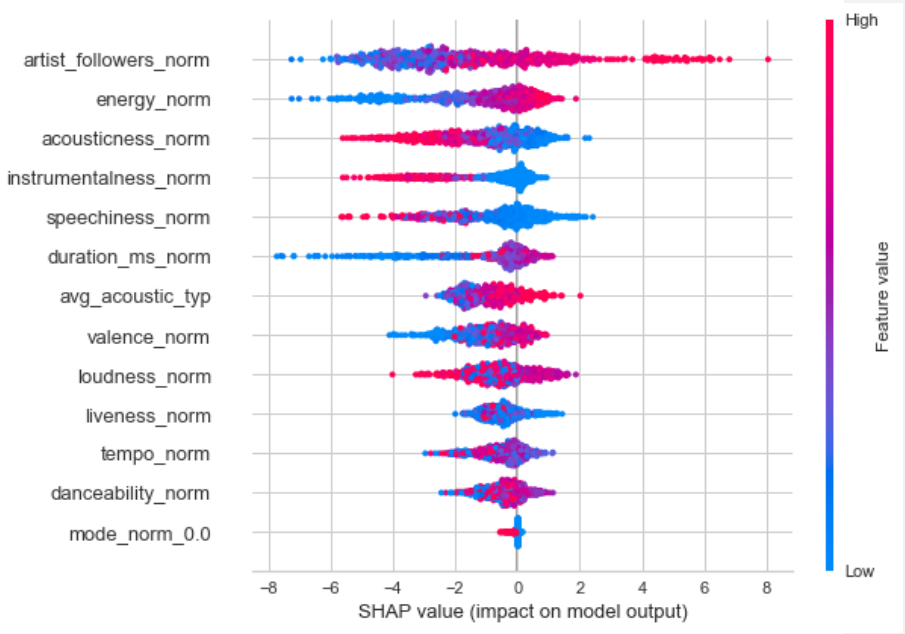

Figure 4.4.3 – SHAP values of features from XGB-U model in Experiment 3

From the SHAP values in Figure 4.4.3, number of artist followers was once again ranked as the most important feature, with the direction of impacts consistent with findings from previous experiments. Energy, acousticness, and instrumentalness were also revealed to be important features. More popular songs tend to be high in energy, contain more electronic than acoustic sounds. They are also more likely to be vocal tracks; songs with lyrics; instead of instrumental tracks such as classical pieces. Some of the more interesting findings were that both high and low values in duration, loudness, and danceability contributed to a lower probability of a song being a hit. Instead, songs with values closer to the mean for these features were more likely to be hits.

Commented [MK33]: I always considered this very strange. You need to be popular first in order to have a hit song. You don't need a hit song in order to be popular.

So, talent is a 2nd order phenomenon – make sure your instagram account and PR machine is working overtime to get you those hits 😊

Commented [MK34]: Also interesting.

Discussion

Theoretical Implications

The results derived from the 3 experiments provide several insights into Hit Song Science. Acoustic features were useful in classifying hits vs non-hits, particularly when there is a greater separation between the classes, as demonstrated in experiment 3. In attempting to classify between the higher and lower ends of already highly popular songs (Top 50 vs Top 200), acoustic features were found to be inadequate. This could be due to the homogeneity in the acoustic and topical compositions of popular music, making it difficult to discriminate them based on musical features alone. In this case, as suggested by past studies, non-musical features proved to be much more important in contributing to song popularity. For example, number of artist followers was shown to consistently be the most important predictor of popularity in all 3 experiments. This could be attributed to fans behaving more like active propagators as opposed to passive consumers of music in the digital era, where media can easily be shared with others (Shin & Park, 2018). Thus, having a large established fanbase accelerates this sharing, allowing the songs from already popular artists to garner even high volumes of attention.

A similar pattern was found when comparing hit songs in the local and global markets. Neither acoustic nor topical features seemed to affect song popularity, instead number of artist followers was shown to be the most impactful factor in determining hit songs locally. This finding suggested that Singaporean listeners could be evaluating music in a manner that is unique from the general global audience. One possible explanation is the extreme prevalence of Korean-Pop (K-Pop) among Singaporean listeners. The present study had excluded this genre from analysis, but it was found that K-Pop was the 2nd most consumed genre locally. Shin and Park (2018) described the K-Pop industry as highly commodified, with production companies and artistes having a great amount of influence, more so than music of other genres.

As such, a greater emphasis is placed on pre-release marketing in order to generate attention from the fanbase. In such an industry, non-musical features, specifically those relate to marketing efforts may account for a greater proportion of song popularity. These include the dollar amount spent on pre-marketing or budget for the music video and albums. It is possible that the dominance of K-Pop locally has influenced Singaporean consumers to employ a different set of criteria, one that is more heavily skewed by non-musical factors when evaluating songs. Alternatively, it could also be due to the high proportion of social media users. Müller (2021), found that 84.4% of Singaporeans were users of social media. In line with the view that music fans are active disseminators of songs by the favourite artistes, it could be reasoned that in countries with high social media penetration rates, sharing of music is proportionally higher as well. As such, non-musical features may predict song popularity better than musical features in countries where sharing of information is more rapid.

The two main features of interest introduced in this study, namely acoustic and topical features and their relationships with song popularity were inconclusive. Topical typicality was only examined in the first experiment, in which it was not found to be an important predictor. This could be attributed to the current's study use of lyric snippets instead of full lyrics. The limited portion of lyrics may not accurately represent the topic of the songs.

Despite only accounting for a small effect, topical typicality was shown to move in the hypothesized direction, with less typical songs being more likely to be a Top 50 hit. This finding is in line with that from Berger & Packard (2018), who found a positive link between lyrical differentiation and commercial success. However, the explanation behind this link remains unclear. While it is possible that these songs gained popularity due to consumers' genuine preference for novelty, it could also simply be due to a direct consequence of being different. The use of peak chart position as a measure of popularity in the current study did not lend itself well to uncovering these underlying relationships. For example, a song could achieve

an extremely high peak position shortly after its release but quickly drop off from the charts in the following weeks. In such a case, the short-lived popularity of the song would be better attributed to listeners' curiosity or other psychological states, rather than a genuine liking towards the track's novelty. Future studies could uncover these processes further by incorporating more comprehensive measures of song popularity.

Interesting relationships were gleaned between acoustic typicality and song popularity. The results indicate that acoustic typicality affects popularity in different ways depending on how popular the songs already are. When comparing between Top 50 vs Non-Top 50 hits, less typical songs were more likely to be in the Top 50. However, when comparing between Top 100 hits and non-hits, the result was reversed; songs that were more typical were also more likely to be hits. This set of results implies that consumers generally prefer songs that sound more familiar or typical, but songs that sound more unique among those already popular see the most success. The current findings provide support for claims that song popularity is determined by a track's position, relative to other songs in the feature space (Askin & Mauskapf, 2017). However, the present study did not find an inverted U-shaped relationship between acoustic typicality and popularity as was previously discovered. This could be attributed to the authors using the peak position as an ordinal measure of popularity, while the present study utilised a more simplified binary measure.

Acoustic features' contributions to song popularity were inconsistent across the 3 experiments, validating the difficulty of using acoustic qualities to predict hit songs as discovered in previous studies (Lee & Lee, 2015; Nijkamp, 2018). Based solely on the results from experiment 3, it was found that hit songs tend to sound more electronic as opposed to acoustic. Paradiso (1997) noted that electronic musical instruments such as synthesizers allowed for musicians to produce more complex sounds, which may have contributed to the prevalence of electronic sounds in popular music. Less instrumental songs, or songs that

contain more vocals, were also more likely to be hits. Nichols et al. (2009) found that syllabic stress in lyrics tend to coincide with melodic peaks and note durations. This tight coupling of lyrics and melody may provide for a richer listening experience, making them more popular than instrumental songs which do not contain any lyrics.

Practical Implications

Experiment 3 produced a well-performing model to classify popular and unpopular songs. This section presents practical uses for the model in the music industry. Figure 5.1.1 below outlines a simplified product life cycle of a song or album (DeArcangelis, 2016).

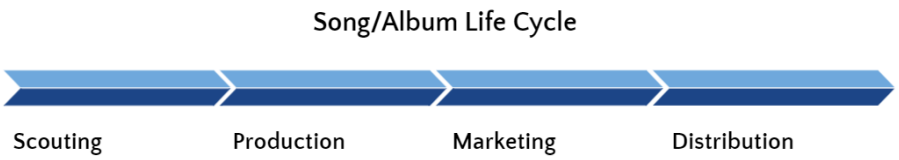

Figure 5.1.1 – Product Life Cycle of Song/Album

In the early stages of a song's life cycle, record labels scout for talents that to join the company. The scouting process may involve holding auditions, or simply browsing content-sharing websites like YouTube to find amateur musicians who show potential. The classification model may be able to expedite that process by accurately identifying songs with attributes that are likely to make them popular.

In the production stage, the artist works with music producers to create a song. Here, the SHAP values derived from the model highlight the most important features that determine a hit song. For example, it was shown that songs with higher energy and less instrumentalness were more likely to be hits. Artists could tune their songs accordingly, such as incorporating ad-libs, to reduce the instrumentalness of their tracks.

Commented [MK35]: good

Next, marketing of the album refers to the creation of promotional content and advertisements surrounding the album. The model can be used here to facilitate resource allocation, in determining which albums are more likely to reach popularity, thus given more resources to promote.

Commented [MK36]: good

Finally, in the distribution phase, the album is released through music distributors, which in the modern state of the industry refer to music streaming services such as Spotify and Apple Music. Companies may want to enter into partnerships with record labels for albums that are likely to be hits. This would give them the exclusive rights to distribute the album on their platform, providing them a unique advantage over their competitors.

Conclusion

The current study attempted to classify hit songs vs non-hits in the Singaporean and global markets using both acoustic and lyrical features of a song. While there is some support for the claim that Singaporeans evaluate hit songs differently from that of a global audience, the processes and determinants that they use remain unclear, and could be examined in future research. Non-musical features such as those pertaining to the artists and production companies were identified as good candidates for potential predictors.

Song popularity was shown to be determined partly by how typical, or familiar a song sounds, with the direction of its impact being highly dependent on the songs it is being compared to. There was also no evidence to suggest that neither the typicality nor the topics of song lyrics themselves, contributed to song popularity. However, the present study was limited in obtaining full lyrics for songs due to licensing issues and songs in foreign languages. Future studies could address this by obtaining the full-length lyrics of songs, or translation of foreign songs used in this analysis.

Finally, there was partial support for some acoustic features to have an impact on commercial success, but results remain largely inconsistent. This could be due to the inadvertent loss of information when distilling a song's musical qualities to only a few data points. Researchers with more in-depth knowledge of music theory could explore the use of more complex musical features also provided by Spotify, such as chroma and harmony. These features intrinsically contain more acoustic information, and thus may be more well-suited to predict song popularity.

(6399 words)

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.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Boyle, J. D., Hosterman, G. L., & Ramsey, D. S. (1981). Factors influencing pop music preferences of young people. *Journal of Research in Music Education*, *29*(1), 47-55.

DeArcangelis, C. (n.d.). *The 5 phases of a recording's life cycle that you need to know*. Sonicbids Blog - Music Career Advice and Gigs. Retrieved November 7, 2021, from https://blog.sonicbids.com/the-life-cycle-of-a-recording.

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.

Lee, J., & Lee, J. S. (2015, October). Predicting music popularity patterns based on musical complexity and early stage popularity. In *Proceedings of the Third Edition Workshop on Speech, Language & Audio in Multimedia* (pp. 3-6).

Ling, C. X., & Sheng, V. S. (2010). Class Imbalance Problem.

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).

Merrick, L., & Taly, A. (2020, August). The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 17-38). Springer, Cham.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011, July). Optimizing semantic coherence in topic models. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 262-272).

Müller, J. (2021, April 07). Number of social network users in Singapore. Retrieved from https://www.statista.com/statistics/489234/number-of-social-network-users-in-singapore/#:~:text=Singapore: number of social network users 2017-2025&text=This statistic shows the number,from 4.74 million in 2019.

Nichols, E., Morris, D., Basu, S., & Raphael, C. (2009, October). Relationships between lyrics and melody in popular music. In *ISMIR 2009-Proceedings of the 11th International Society for Music Information Retrieval Conference* (pp. 471-476).

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

Paradiso, J. A. (1997). Electronic music: new ways to play. *IEEE spectrum*, *34*(12), 18-30.

Pepe Python (2021). Spotify HUGE database – daily chart over 3 years, Version 1. Retrieved from https://kaggle.com/pepepython/spotify-huge-database-daily-charts-over-3-years/metadata.

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.

Shin, S., & Park, J. (2018). On-chart success dynamics of popular songs. *Advances in Complex Systems*, *21*(03n04), 1850008.

Spiros Theodoropoulos (2021). Hit song science – 34740 songs (+spotify features), Version 1. Retrieved from https://www.kaggle.com/multispiros/34740-hit-and-nonhit-songs-spotify-features.

Yap, B. W., Abd Rani, K., Abd Rahman, H. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)* (pp. 13-22). Springer, Singapore.

Appendix

No.	Feature	Description
1	Acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
2	Danceability	How suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
3	Instrumentalness	Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness: value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0
4	Speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks
5	Valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)
6	Loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db

7	Duration	The duration of the track in milliseconds
8	Energy	Measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy
9	Tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration
10	Liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live
11	Mode	Indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0