Exploring the Influence of Intrinsic, Extrinsic, and Crowdsourced Features on Song Popularity

Sandra Angela Berjamin, Angeli Dianne Mata,
Paolo Montecillo, and Rafael Cabredo

De La Salle University Manila
{sandra_angela_berjamin, angeli_dianne_mata,
gio_montecillo, rafael.cabredo}@dlsu.edu.ph

Abstract. Hit songs from popular music artists have been investigated to help uncover the pattern underlying the unique appeal of their tracks. Given this, intrinsic, extrinsic, and crowdsourced features have been identified as some of the necessary information in determining the popularity of a song. Each of these features alone is lacking in reaching the said objective. As a result, the combination of these features was hypothesized to improve the estimation of the performance of a track. Structural equation modeling was done to check the impact of each of the features to the said performance. Then, the comparison of the random forest, support vector machine, and boosting trees techniques to predict the 10th week of streams for each of the songs was done. In conclusion, the extrinsic, and crowdsourced features were discovered to be the most important, and all three modeling techniques used performed similarly to each other.

Keywords: music analysis, intrinsic, extrinsic, structural equation modeling, hedonic consumption constructs

1 Introduction

The various features of hit songs have been observed and analyzed to help determine the pattern behind the popularity of certain musical tracks. The significant information mentioned from prior research can be divided into four groups: intrinsic, extrinsic, emotional, and crowdsourced. Intrinsic audio features are those that can be extracted from the track itself. However, the usage of the standard acoustic features alone was lacking in predicting music popularity [1]. Extrinsic features are characteristics of the song that do not originate from the track itself. Nonetheless, this does not consider the personal experience consumers may have with music [2]. Music can be classified based on the emotions felt by the listeners, which is known to be influenced by several factors such as personality, environment, and listening mood [3]. Despite this, it is important to note that to acquire the overall popularity of a track, the general opinion of a large group should be considered, not the detailed emotional experience of a few entities. The information gathered from a crowd is generated from an efficient and responsive process which enables it to achieve high efficiency at only low costs [4].

J. Caro et al. (eds.), *Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2023)*, Atlantis Highlights in Computer Sciences 20,
https://doi.org/10.2991/978-94-6463-388-7_24

A study [5] used a method wherein crowdsourcing was added to a model that used both extrinsic and intrinsic features. This additional information significantly improved the model's predictions about a track's chart success. Despite this result, there remains a lack of research that further tests this method in the field of hit song science. Since it is still underutilized and lacks resources, crowdsourcing is not reaching its potential to determine the crowd's wisdom [6].

All of the features mentioned above contribute to the popularity of a song, but using each feature independently is inadequate. Combining intrinsic, extrinsic, and crowdsourced features can be seen as an effective way to determine whether a song will be a hit. Given this, the main objective of this research was to add crowdsourced information to intrinsic and extrinsic features in analyzing the popularity of a song to improve the predicting performance of machine learning techniques.

By conducting this research, the range of values of some intrinsic features that are common for the hit songs analyzed may be used by artists as a guide in producing songs. Additionally, those involved in the marketing field of the music industry may obtain better insights on how to effectively distribute and market tracks based on the identification of the most important extrinsic and crowdsourced features. Lastly, this research could be the basis for future multicultural studies.

2 Methodology

The research has four phases, namely data collection, pre-processing, model building, and analysis.

2.1 Data Collection

General Data The songs used and their essential information (i.e. song's title, artist, and stream count) were gathered from the *Spotify Charts*, specifically, the *Weekly Top Songs* in the Philippines. This study only took into consideration the first week's top 50 songs. This is to limit the songs to be observed by the participants in crowdsourcing. Regarding the weekly stream count for all 10 weeks, it was decided to divide the streams into 10 categories in order to give way for relatively small differences in values.

Intrinsic Data The intrinsic features of each song from the chart were collected using the *Spotify* API. Only the following features that were said to have great influence in popularity based on previous works [7,8] were used: tempo (beats per minute), valence, speechiness, loudness, key, liveness, duration, and danceability.

Extrinsic Data The extrinsic features revolved around the chart history of an artist, the size of the music label (i.e. major, independent, major-independent) associated with the song and the artist, the presence of an artist collaboration, and the consumer's awareness of the song and artist. In numerical form, one (1) represented independent, five (5) represented major-independent and 10 represented major. In determining the artist's chart history until the last week of data collection, the total number of weeks the artist remained on the Philippines' *Weekly Top Artist* chart from *Spotify* was used. The information on music labels is publicly available on *Spotify.* For artist collaboration, one (1) showed the presence of such while zero (0) portrayed the absence of it. Lastly, the consumer awareness variable of an artist, and song was determined on the 10th week of data collection using the *Google Trends search volume index* with a period of 10 weeks or the whole time span of data collection.

Crowdsourced Data The forms for the crowdsourced data were made using *Google Forms*, and were released using convenience sampling. It was posted in the *DLSU Community Forum* and the public *Facebook* pages of the researchers. The data came from 200 people who voluntarily agreed to participate and satisfied all the selection criteria: have lived in the Philippines for at least 8 months, whose age is at least 13 years old, could give their email address should they wish to be contacted, and the general area they are currently living in. Each song was evaluated twenty times, yielding 1,000 responses, of which each participant only evaluated five songs. During each data collection session, 30-second breaks were included in the *Google Forms* as a precautionary measure to help participants avoid exhaustion while filling out the form.

To have a more complete view of the participants' backgrounds, they were also asked about some significant information relating to their consumption of music, which may affect their evaluation of the songs. This included their musical affinity [9], which determined how significant music was to each individual. They were also asked about their background in any formal music training and musical employment. Each participant also performed the Short Test of Music Preferences (i.e. STOMP) to rate each of the 14 music genres [9] to show their music preference. Each person's music consumption preference [5] was also taken into consideration to see their preference for physical consumption (e.g. purchasing official CDs or vinyls) and digital consumption (e.g. streaming or downloading).

While answering the survey, the participants were expected to be in a quiet location free from any disturbances while wearing headphones or earphones for a more immersive listening experience. Each participant evaluated the first 90 seconds of a song, the minimum time it takes to form one's first impression of a certain thing or concept [10]. They then rated the song's overall affective response, which showed the initial impression of the participants to the song and the need to re-experience, which measured the desire of the participant to experience the track again. Furthermore, since the songs were not all newly released and the popularity of the different artists vary, each participant were asked of their familiarity with two aspects: the song, and the artist [11].

To preserve the data quality of the evaluation for each song, instructional manipulation checks [12] was applied throughout the said process. A unique text appeared at the end of all the 90-second video snippets of each of the 50 songs. The said text was required to be noted down by the participant, as a way to check if they have finished the whole video. Each media was posted on *Vimeo*, wherein one can post videos without the playbar. With this advanced setting, the participants did not have the ability to fast forward to the end of the video in order to shorten the time it took for them to finish the evaluation. In all song evaluations where the participant's input did not match the unique text were excluded from any further processing and analysis.

2.2 Pre-processing

Exploratory Data Analysis (EDA) was performed on each of the features to understand better the data gathered, namely (1) weekly rank, and streams, (2) intrinsic data, (3) extrinsic data, and (4) crowdsourced data. The evaluation of the different pairs of the said categories of data was also done to evaluate the possible relationships between each of them.

With the help of appropriate summary statistics and visualization techniques, the correlation between features were identified more easily. Also, through EDA, the features and their corresponding data types, as well as whether they contain null values and required some form of standardization, were investigated. To normalize all the cardinal data accordingly, different techniques were experimented with, namely the utilization of (1) the *StandardScaler()* function, (2) Max Absolute, and (3) Min Max Scaling. After building the models, it was later discovered that the *StandardScaler()* function gave the best results.

2.3 Model Building

In this phase, models were created from the pre-processed data acquired from the previous step. There were two major stages done for this step: (1) the usage of the Structural Equation Modeling (SEM) technique, and (2) the comparison of the Boosting Trees (BT), Random Forest (RF), and Support Vector Machine (SVM) modeling techniques.

Structural Equation Modeling SEM is a data analysis method combining simultaneous regression equations with factor analysis. Due to this, it is frequently applied in the field of social work, which involves the investigation of complex structures of concepts (i.e., cognition, affect, and behavior) that cannot be easily measured accurately with a single item from a survey [14]. Given this, it was seen fit for this research to make use of SEM to check the impact of each of the features on the overall performance of the songs since the scope of the study deals with constructs under the topic of hedonic consumption (i.e., the analysis of an event of consumption as a subjective experience rather than an information-processing circumstance), namely the overall affective response and the need to re-experience.

Two separate models were created regarding the values obtained from the pre-processed data. The first model only contained features from intrinsic, and extrinsic data while the second model had features from intrinsic, extrinsic, and crowdsourced data. Model specification (i.e., where the hypothesized relationships between the different variables were set), model estimation (i.e., where the parameters were estimated), and model evaluation (i.e., where the model was checked regarding its overall fit, and the significance of the parameters utilized) were performed for each of these models.

Boosting Trees, Random Forest, and Support Vector Machine To predict the tenth week of streams for each song, BT, RF, and SVM methods were used given the first 9 weeks of streams as they are said to be the most commonly used machine learning models in terms of predicting music popularity and were among the ones that performed best in certain studies [1,7]. BT is an ensemble learning algorithm that adds decision trees one at a time, wherein the errors from the previous trees given increased weight when building the next tree [15]. Similar to BT, RF utilizes multiple decision trees but it centers on the concept of developing unique and independent trees instead of continuously improving the model through the iterative process of adjusting the weights of residuals [16]. On the other hand, SVM are learning systems that transforms the inputted features into a high dimensional space through the usage of a kernel function [17]. The SVM classifier is known as a hyperplane that is optimized to divide observations into their respective classes according to their features, which are distinct patterns seen in the input data. The said features are used as coordinates on the high dimensional space based on their corresponding relationships to one another [18].

First, the data was divided wherein 75% was used for training while the remaining 25% was used for testing. Using the said models, six different groups of inputs were used. These selection of features were namely (a) only intrinsic data, (b) only extrinsic data, (c) only crowdsourced data, (d) intrinsic and extrinsic data, (e) intrinsic, extrinsic, and crowdsourced data, and (f) the best features according to the results of SEM. The training and testing accuracy were evaluated, and compared to determine the best model among the three [7]. The best modeling technique and selection of song features to predict a song's popularity were then determined through further investigation of the data collected and models created.

2.4 Analysis

The last phase was the complete analysis of the data collected from all the previous steps. The model evaluation step in SEM helped in the analysis and comparison of the resulting models using the SEM modeling technique. Moreover, it allowed the determination of which features played significant roles in the popularity of a song. Besides the analysis provided by SEM, the statistics and visualizations created from the EDA were further investigated and examined. The underlying information that were previously observed were clarified with the help of the information provided by the SEM analysis as the relationships of the most essential features with the popularity of a song were confirmed. To determine which model performed the best, the overall accuracy produced by the generated models were compared.

3 Results and Discussion

As previously mentioned, the collected data and different pairs of their main categories were analyzed through EDA. Afterwards, the results of the models were compared and investigated further. The following section shows the significant findings discovered in the whole process.

3.1 EDA: The Intrinsic Features

Among all the intrinsic features, speechiness had the lowest value of standard deviation. This hinted that a certain range of this specific intrinsic feature may be a possible similarity to a majority of the 50 songs. Additionally, key had three top values (i.e. the keys of F#, B, and G#), wherein almost 40% of the 50 songs had.

3.2 EDA: The Extrinsic Features

First, it was defined that mainstream artists were those who have remained in the Weekly Top Artists Philippines Chart from *Spotify* ever since the chart was first created, which was 73 weeks before the end of the data collection period. Given this, there was a difference between the number of songs in the charts for mainstream artists, and for non-mainstream artists. For the number of artists who had more than one song in the Top 50 charts for the first week, there were more mainstream artists compared to non-mainstream artists. Besides the reputation of the artist, the type of music label was also analyzed. There were more songs in the charts from a major music label compared to the other two categories, which were independent and major-independent.

3.3 EDA: The Crowdsourced Features

In the year 2023, the song evaluation process started at January 31 and ended at February 24.

The ratings of the sub-features under overall affective response had a similar range to each other. The same could be said for the sub-features under the need to re-experience.

In order to check the similarity of how participants rated each song, the standard deviations of the sub-features for the overall affective response and for the need to re-experience were extracted. The results helped conclude that the participants tended to rate each song similarly for all the sub-features under each main crowdsourced feature.

3.4 EDA: The Relationship of Extrinsic and Crowdsourced Features

One significant difference discovered was that the participants' artist and song familiarity with non-mainstream artists were generally higher compared to that of mainstream artists. This was a particularly interesting finding since one would assume that mainstream artists and their songs should be more well-known compared to artists who were otherwise. Additionally, there was a difference in terms of the ranges between the crowdsourced features when songs were divided into the different types of music labels, as seen in the boxplots. This finding was confirmed when the standard deviation of each of the crowdsourced features for the songs under each type of music label was computed. The major label had a relatively low standard deviation values, which showed that songs that were released by a major label were rated the most similarly according to their overall affective response. The same could be said about the ratings under the need to re-experience. To further investigate this finding, the mean for each sub-feature was also computed. This helped understand what type of rating did the participants give songs under each type of music label. Noting the standard deviation values from the previous step, the data showed that songs under major labels tended to be rated the highest by the participants in terms of the overall affective response. Again, the same could be said about the responses to the need to re-experience. This showed the possibility that major labels made good use of their abundant resources to understand the taste of the general public enough to let the charts show a pattern that is advantageous to them.

3.5 SEM: The Intrinsic, Extrinsic, and Crowdsourced Features

To determine the influence each of the intrinsic, extrinsic, and crowdsourced features to the popularity of a song, two models were created. Model B had intrinsic and extrinsic features as input while Model A had the addition of the crowdsourced features. For both of these models, the extrinsic features were seen as the important feature in determining the performance of a song, as seen in Table 1. This was because features are deemed as significant in SEM when they have a p-value less than 0.5.

In terms of the intrinsic features, Table 2 presents the results. The most significant features were found to be loudness, speechiness, duration, valence, and danceability. The estimates in SEM show how significant each of the said features are. Given this, valence had the highest estimates for both models,

Table 1. The p-value to the "chart success" variable

Feature	Model A	Model B
Intrinsic	0.092000	0.052000
Extrinsic	0.000007	0.000010
Crowdsourced	0.269000	N/A

making it the most important intrinsic feature. The negative estimates of the duration variable represented its possible relationship to chart success, wherein the longer a song is, the lower the chance for the song to be successful in the charts.

For the extrinsic features, only the artist collaboration and music label features showed some sign of significance to both models, as shown in Table 3.

For the crowdsourced features, the results can be seen in Table 4. All sub-features of the overall affective response were deemed as significant. However, the dull-exciting and forgettable-memorable categories were observed to have the least significance. All sub-features of the need to re-experience were seen as significant as well, all being almost equally significant. Additionally, artist familiarity was not seen at all as significant to impact the popularity of a song, which somehow contradicted the results of the extrinsic features wherein artist collaboration was the most significant feature.

Table 2. The p-values to the "intrinsic" variable and estimates (for significant features only)

	Model A		Model B	
Intrinsic Feature	**p-value**	**Estimate**	**p-value**	**Estimate**
Valence	0.0002	0.971	0.0003	0.996
Loudness	0.002	0.745	0.002	0.799
Speechiness	0.016	0.575	0.013	0.621
Duration	0.0005	-0.894	0.0007	-0.913
Danceability	-	1.000	-	1.000
Liveness	0.375	-	0.435	-
Tempo	0.951	-	0.891	-
Key	0.957	-	0.977	-

Table 3. The p-values to the "extrinsic" variable and estimates (for significant features only)

	Model A		Model B	
Extrinsic Feature	**p-value**	**Estimate**	**p-value**	**Estimate**
Artist Collaboration	0.003	0.023	0.004	0.022
Music Label	-	1.000	-	1.000
Chart History	0.778	-	0.680	-
Google Trends	0.699	-	0.727	-

3.6 BT, RF, and SVM: The Intrinsic, Extrinsic, and Crowdsourced Features

Many experiments were done involving the three different modeling techniques and different selection of features, the results of which can be seen in Table 5. According to the testing accuracy of the models trained and tested, the RF and BT algorithm performed the best and, generally speaking, the models performed better with only the intrinsic and extrinsic features as input. This may have been because RF and BT are both based on the concept of trees. This characteristic allowed them to create a large amount of trees with high-depth level and learn all the training space having all the possible combinations. In terms of the low training accuracy for BT, it might be possible that this model technique could not capture the complexity of the data, which then led to a lower training accuracy compared to their testing accuracy. Another interesting finding was that the values for the training accuracy and testing accuracy for BT was the same for all of the varying selection of inputs.

3.7 Additional Experiments

After building the initial models in the study and understanding their results, various attempts were made to improve and understand the developed systems more.

Table 4. The p-values to the “crowdsourced” variable and estimates (for significant features only)

Crowdsourced Feature	p-value	Estimate
AR [bad-good]	0.002	2.414
AR [unpleasant-pleasant]	0.002	2.407
AR [distasteful-tasty]	0.002	2.396
AR [tasteless-tasteful]	0.002	2.374
AR [untalented-talented]	0.002	2.238
AR [boring-interesting]	0.002	2.183
AR [unimaginative-creative]	0.003	2.100
AR [forgattable-memorable]	0.003	1.905
AR [dull-exciting]	0.005	1.679
NR[share to my friends]	0.002	2.349
NR[listen to similar songs]	0.002	2.294
NR[add to my playlist]	0.002	2.276
Song Familiarity	-	1.000
Artist Familiarity	0.496	-

Additional Feature and Reduction of Stream Categories Another feature was added, which was the weekly rate of change in streams of each song. The initially generated models only depended on the value of the stream category for every week to determine the position of the track in the hit charts. This new feature was hypothesized to give the models a better idea of the weekly trend in terms of the performance of each song.

The decrease in stream categories was done as well due to the data having unequal number of songs per stream category as seen from Table 6. There were ten categories previously, but they were reduced to two as seen from Table 7: the first contains data from previous stream categories 9 and 10, and the second contains data from previous stream categories 8 and below. This division was done because there is a high number of songs in the 9 and 10 category as compared to the other categories. In addition, there were a few categories—specifically 3, 2, and 1—where no song instances were present.

Using the new data in SEM, it was observed that extrinsic and crowdsourced features have significant roles in determining the chart success of a song as the p-

Table 5. The results of RF, SVM, and BT

Model	Features Involved	Training Accuracy	Testing Accuracy
Random Forest	Intrinsic	97.30	84.62
	Extrinsic	91.89	84.62
	Crowdsourced	91.89	76.92
	Intrinsic + Extrinsic	100.00	92.31
	Intrinsic + Extrinsic + Crowdsourced	91.89	92.31
	Best Features	100.00	84.62
Support Vector Machines	Intrinsic	78.38	61.54
	Extrinsic	75.68	61.54
	Crowdsourced	100.00	38.46
	Intrinsic + Extrinsic	100.00	76.92
	Intrinsic + Extrinsic + Crowdsourced	100.00	53.85
	Best Features	100.00	76.92
Boosting Trees	Intrinsic	72.97	92.31
	Extrinsic	72.97	92.31
	Crowdsourced	72.97	92.31
	Intrinsic + Extrinsic	72.97	92.31
	Intrinsic + Extrinsic + Crowdsourced	72.97	92.31
	Best Features	72.97	92.31

values of the two features are both less than 0.05 as seen in Table 8. Meanwhile, the intrinsic features still have the least impact on a song's performance.

In terms of the RF, SVM, and BT models, the new modified data was observed to have increased the performance of the three models as their test accuracy were high as they were usually 100 percent as shown in Table 9. However, having these high accuracies could also be the result of having a small test dataset of only 10 instances as the training accuracies were not 100 percent. The importance of the features for the best models in RF and BT was examined in order to have an idea how the models made use of the features in predicting the 10th week stream category. It was seen that in RF, all features were utilized but the most significant features were the stream categories of weeks eight and nine.

Table 6. Values of the Initial Stream Category

Stream Category	Initial Count
10	8
9	26
8	1
7	7
6	2
5	4
4	2
3	0
2	0
1	0

Table 7. Values of the Reduced Stream Category

Stream Category	Initial Count
1	34
0	16

Table 8. The p-value of models from Initial Test, Experiment with Additional Feature, and Reduction of Stream Categories

Feature	Initial p-value	p-value after Experiment
Intrinsic	0.092000	0.802247
Extrinsic	0.000007	0.000844
Crowdsourced	0.2690000	0.049040

On the other hand, only the stream category of week nine was used in the BT model as all the other features have feature_importance value of 0.

Deletion of Weekly Stream Category In addition to the previous experiment, the stream category features of weeks two to nine were deleted in order to avoid the models solely focusing on the previous week categories for the prediction of the 10th week stream category. Due to removing some stream information, the test accuracy of RF, SVM, and BT models decreased as seen from Table 10. The three traditional ML models have a relatively similar performance as their accuracy are near to each other although there is a possibility that this similarity might be due to the small dataset. In terms of the importance of features in the RF and BT models, it was observed that the weekly rate of change features have a constant relatively high importance but other features such as artist_familiarity, unimaginative_creative, and forgettable_memorable were also found to have high significance.

Averaging the Performance of Models The previous experiments all ran each of the models only once. In this experiment, each of the models was run three times, in which every run used a different split of training and testing datasets. The accuracies obtained for each run of the model were then averaged. This was done to make up for the lack of data the study has. The average prediction accuracy in the running of the different models are presented in Table 11. As seen from the values, the SVM model obtained the highest training and testing

Table 9. The Performance of Initial Test and Experiment with Additional Feature

	Initial Results		**Experiment Results**	
Model	**Train Accuracy**	**Test Accuracy**	**Train Accuracy**	**Test Accuracy**
Random Forest	100.0	76.9	97.5	100.0
SVM	48.6	61.5	100.0	100.0
Boosting Trees	78.4	76.9	97.5	100.0

Table 10. The Performance of Initial Test and Experiment with Deletion of Weekly Stream Category

	Initial Results		**Experiment Results**	
Model	**Train Accuracy**	**Test Accuracy**	**Train Accuracy**	**Test Accuracy**
Random Forest	100.0	76.92	92.5	70.0
SVM	48.64	61.53	99.17	83.33
Boosting Trees	78.37	76.92	95.0	63.33

accuracy out of the three models, although all the test datasets had only 10 instances and all of the models had a relatively similar training performance due to their values being near to each other.

Deletion of the Stream Information from Weeks 1 to 9 and the Correlated Features One consistent result that could be seen from the previous experiments was the fact that features regarding the stream information of the previous weeks (i.e. weekly rate of change and stream categories from Weeks 1 to 9) were consistently determined as some of the most significant features. Given this, they were deleted exclusively for this experiment. This was done in order to let the models focus on the intrinsic, extrinsic, and crowdsourced information extracted for each of the songs to predict the 10th week of streams.

The correlation between features within each feature group was also examined. This was done to remove features that have a high correlation with another feature, as highly correlated features may have had a significant influence in the building of the ML models. The threshold for defining high correlation was set at 0.97 based on the observation of the correlation between the crowdsourced features. In the intrinsic and extrinsic feature groups, it was observed that there were no highly correlated features. Meanwhile, in the crowdsourced feature group, it was found that distasteful_tasty, tasteless_tasteful, unpleasant_pleasant, untalented_talented, and bad_good were highly correlated. In addition to this, share_friends, add_playlist, and listen_similar were also found to be highly correlated with one another. Only one feature from the group of highly correlated features were retained.

Due to removing all stream information, the test accuracy of RF, SVM, and BT models decreased as seen from Table 12. Similar to the results of previous experiments, the three traditional ML models have a small difference in terms of their performance as their accuracy are near to each other (i.e. ranging from 60 to 70). However, as mentioned in previous sections, the small dataset may be one reason for this occurrence. In terms of the importance of features in the RF and BT models, it was observed that the crowd's rating of artist_familiarity had the highest significance. This shows potential on how information from the crowd can be used to predict a song's popularity. Additionally, among the remaining features after the deletion of those correlated to another, the one with the highest significance was the boring_interesting feature according to the results of the RF model. This may be due to the clarity of the words used to rate

Table 11. Average of Performance of the Best Models

Models	Train Accuracy	Test Accuracy
Random Forest	92.5	70.0
Support Vector Machines	99.2	83.3
Boosting Trees	95.0	63.3

this specific feature, compared to the other subfeatures under the overall affectiveness category (i.e. namely bad_good, dull_exciting, unimaginative_creative, and forgettable_memorable).

Addition of Stream Categories From having only 2 stream categories, this experiment turned them to three instead. Those categorized as high-performing songs belonged to category 10 from the initial dataset. Those under the mid-performing songs were songs that initially belonged to category 9 while the rest of the songs were labelled as low-performing songs. The total count of each of these stream categories can be seen in Table 13. This new division was done to further test the capability of the inputted features and models to predict the popularity of songs.

In terms of the RF, SVM, and BT models, this new division of stream categories greatly decreased the performance of the three models as shown by the relatively low test accuracies in Table 14. This may be due to the overfitting of the models to the training dataset since there was a more limited amount of information for each stream category compared to when there was only two stream categories.

4 Conclusion

Through SEM, the extrinsic features were deemed to be the most significant features in determining the performance of a song. Moreover, the most important intrinsic feature were loudness, speechiness, duration, valence, and danceability. For extrinsic features, the most essential ones were artist collaboration and music label. Song familiarity, overall affective response and the need to re-experience were the most critical crowdsourced features. After feeding the SEM model more information about the weekly performance of the song and reducing the stream categories to only two values (i.e. low and high performing), the extrinsic and crowdsourced features were both seen as significant features in determining the performance of a song in the 10th week. Also, the random forest, support vector machines, and boosting trees models all generated a similar range of training and testing accuracies for each experiment. Given this, all three modeling techniques

Table 12. The Experiment with Averaging Performance of Models and Experiment with Deletion of All Stream Information

	Experiment #5 Results		**Experiment #6 Results**	
Model	**Train Accuracy**	**Test Accuracy**	**Train Accuracy**	**Test Accuracy**
Random Forest	92.50	70.0	79.17	66.67
SVM	99.17	83.33	70.0	70.0
Boosting Trees	95.0	63.33	82.5	60.0

Table 13. Count of Newly Divided Stream Categories

Stream Category	Count
high-performing	8
mid-performing	26
low-performing	16

Table 14. The Performance of Experiment with Averaging Performance of Models and Experiment with Three Stream Categories

	Experiment #5 Results		Experiment #7 Results	
Model	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
Random Forest	92.5	70.0	92.5	56.7
SVM	99.2	83.3	95.0	73.3
Boosting Trees	95.0	63.3	73.3	43.3

can be said to have a comparable predicting power in terms of a track's performance in a particular week. This supports the conclusion from previous studies that no specific modeling technique displays notable prediction accuracy since the selection of data inputted into the models impacts their performance significantly [7] [1]. Furthermore, the removing of all features regarding the stream information from the previous weeks showed that the most important features were under the crowdsourced category of inputted data, specifically those that was relatively easy to identify and measure by the participants due to the simplicity of the concept and words used. All of these results led to the conclusion that extrinsic and crowdsourced features have the most impact in the prediction of the performance of a song, which confirms the findings of certain studies [5] [13].

The current study may be seen as the initial step to the usage of crowdsourced data in the analysis of the popularity of songs in the Philippines. However, the findings in this research should be considered carefully due to the small sample size (i.e. in terms of the number of songs, ratings, and participants for the crowdsourcing aspect), and the lack of diversity among the participants (i.e. most of which are college students). Moreover, this study only made use of one specific time period (i.e. a span of 10 weeks), which begs the question if there is a specific duration of time to best predict a song's popularity. It is highly suggested for future studies to increase the number and genre of songs involved and consider the usage of all types of songs, not just hit songs, in order to give the model a better idea of the difference between a track with a good performance in the top charts to that with bad performance.

References

1. J. Lee and J. Lee, "Music popularity: Metrics, characteristics, and audio-based prediction," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3173-3182, 2018.
2. M. Lorenzen and L. Frederiksen, "The management of projects and product experimentation: examples from the music industry," *European Management Review*, vol. 2, no. 3, pp. 198-211, 2005.
3. C.C. Liu, Y.H. Yang, P.H. Wu and H.Chen, "Detecting and classifying emotion in popular music," *9th joint international conference on information sciences*, 2006.
4. K.B. Wilson, V. Bhakoo and D. Samson, "Crowdsourcing: A contemporary form of project management with linkages to open innovation and novel operations," *International Journal of Operations & Production Management*, 2018.
5. D.M. Steininger and S. Gatzemeier, "Digitally forecasting new music product success via active crowdsourcing," *Technological Forecasting and Social Change*, vol. 146, pp. 167-180, 2019.
6. K. Wazny, "Crowdsourcing ten years in: A review," *Journal of global health*, vol. 7, no. 2, 2017.
7. A. Gao, "Catching the Earworm: Understanding Streaming Music Popularity Using Machine Learning Models," *E3S Web of Conferences*, vol 253, pp 03024, 2021.
8. Y. Ge, J. Wu and Y. Sun, "Popularity prediction of music based on factor extraction and model blending," *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, pp. 1062-1065, 2020.
9. P.J. Rentfrow and S.D. Gosling, "The do re mi's of everyday life: the structure and personality correlates of music preferences," *Journal of personality and social psychology*, vol. 84, no. 6, pp. 1236, 2003.
10. J.E. Lydon, D.W. Jamieson and M.P. Zanna, "Interpersonal similarity and the social and intellectual dimensions of first impressions," *Social cognition*, vol. 6, no. 4, pp. 269, 1988.
11. J.L. Tam, "Brand familiarity: its effects on satisfaction evaluations," *Journal of Services Marketing*, 2008.
12. D.M. Oppenheimer, T. Meyvis and N. Davidenko, "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of experimental social psychology*, vol. 45, no. 4, pp. 867-872, 2009.
13. M.O. Silva and M.M. Moro, "Collaboration-Aware Hit Song Analysis and Prediction," *Anais Estendidos do XXVII Simpósio Brasileiro de Sistemas Multimídia e Web*, pp. 11-14, 2021.
14. N.K. Bowen and S. Guo, "Structural equation modeling," Oxford University Press, Incorporated, 2011.
15. C.D. Sutton, "Classification and regression trees, bagging, and boosting," *Handbook of statistics*, vol. 24, pp. 303-329, 2005.
16. T.G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, pp. 139-157, 2000.
17. N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods," Cambridge University Press, pp. 93-124, 2000.
18. D.A. Pisner and D.M. Schnyer, "Machine Learning," Academic Press, pp. 101-121, 2020.
19. M.B. Holbrook and J. Huber, "Separating perceptual dimensions from affective overtones: An application to consumer aesthetics," *Journal of Consumer Research*, vol. 5, no. 4, pp. 272-283, 1979.

20. R.B. Zajonc, "Feeling and thinking: Preferences need no inferences," *American psychologist*, vol. 35, no. 2, 1980.

A Appendix

The following section presents the exposition of the questions employed in the surveys given to all the participants of the study.

To understand the backgrounds of the participants, their musical affinity [9] was first measured through their response using the 7-point Likert scale, where 1 represents "strongly disagree" and 7 is "strongly agree", to the statement "Music is important to me." They were also asked of the amount of their formal music training and musical employment with the usage of the response categories [5] seen in Table 15.

Each participant also performed the Short Test of Music Preferences (i.e. STOMP) in which they used the 7-Point Likert scale, where 1 represents "strongly dislike" and 7 is "strongly like," to rate each of the 14 music genres [9] shown below.

1. Alternative
2. Blues
3. Classical
4. Country
5. Electronica / Dance
6. Folk
7. Heavy Metal
8. Rap / Hiphop
9. Jazz
10. Pop
11. Religious
12. Rock
13. Soul / Funk
14. Soundtracks

Table 15. Musical Training/Employment Categories [5]

Category	**Years of Experience**
1	None
2	less than 1 year
3	1 to 5 years
4	6 - 10 years
5	11 - 15 years
6	16 - 20 years
7	more than 20 years

Each person's music consumption preference [5] were also taken into consideration through the 7-Point Likert scale having the values of 1 to represent physical consumption (i.e. purchasing official CDs, vinyls) and 7 to represent digital consumption (i.e. streaming, downloading).

In terms of the evaluation of a song, each participant was asked to rate the song's overall affective response, which was measured through the 7-Point semantic differential and was subdivided into 9 categories [19] as presented below.

1. bad-good
2. distasteful-tasty
3. dull-exciting
4. tasteless-tasteful
5. unimaginative-creative
6. untalented-talented
7. unpleasant-pleasant
8. forgettable-memorable
9. boring-interesting

They also rated the song's need to re-experience, which was measured through the 7-Point Likert scale with 1 as "strongly disagree" and 7 representing "strongly agree." This feature is further divided into three statements, which can be seen below.

1. "I want to listen to other songs of the same artist"
2. "I would like to share this song with my friends"
3. "I would love to add this to my favorite playlist"

Furthermore, it is important to consider the familiarity bias, wherein the familiarity with a specific concept or product may affect an individual's evaluation of it [20]. Since the evaluated songs were not all newly released (i.e. some may have been in the charts for more than a few weeks) and the popularity of the different artists vary, each participant were asked of their familiarity with two aspects: the song and the artist. This was done through the 7-point scale with 1 representing "not familiar at all" and 7 as "very familiar" [11].