RISING STAR OR JUST A DREAM?

PREDICTING ARTISTS' SUCCESS BY USING MACHINE LEARNING APPROACHES ON FEATURES IDENTIFIED IN SPOTIFY DATA

MSIN0097 Predictive Analytics
Group 9
Group Assignment – Executive Report

Word count (excl. tables, figures, Appendix, Bibliography): 2181

Link on faculty: https://ucl.my.faculty.ai/project/33feeda8-632d-4f5f-82e1-2eb25ab63c89/workspace

Table of Contents

1. Introduction

Spotify is a platform that allows music streaming to personal devices and tracks user behavior preferences observed over time. Users can create own playlists, "like" songs, which allows Spotify algorithms to recommend tracks based on past behavior, or to enable artists to be more visible to users, by adding them to the top playlists, with highest numbers of followers and streams.

For this report, we will be investigating whether certain artist features are boosting probability of success for artists who have made it to the top four playlists. We will perform an in-depth exploratory analysis to identify any patterns and insights and apply machine learning models to research if success can be predicted by including important features in the model.

This may help musicians or production houses determine if newcomers to the industry have a higher chance of becoming successful, if they possess the features that have a direct impact on the model performance.

2. Definition of Success

We define "success" as the presence of an artist in the key four playlists identified by Warner Music (hereby called a successful artist). These playlists[1] include Hot Hits UK, Massive Dance Hits, The Indie List and New Music Friday. Due to duplicate playlists names and IDs in the original dataset, we chose the playlist IDs corresponding to the above top four playlists based on the highest number of streams for the playlist with the same name.

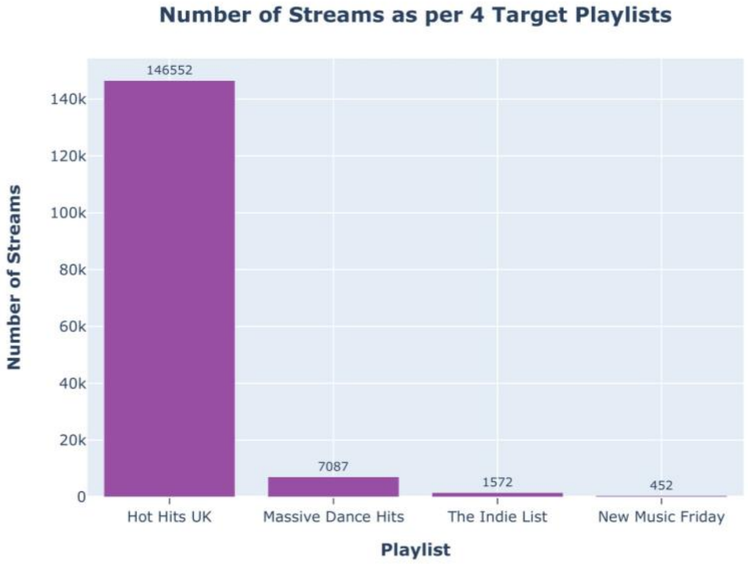

Figure 1: Number of streams for the 4 target playlists

[1] Hot Hits UK id: 6FfOZSAN3N6u7v81uS7mxZ, Massive Dance Hits id: 37i9dQZF1DX5uokaTN4FTR, The Indie List id: 37i9dQZF1DWVTKDs2aOkxu, New Music Friday id: 37i9dQZF1DX4JAvHpjipBk

3. Insights from Data Exploration & Feature Creation

Playlist behavior before and after inclusion into top-4 playlist

By identifying the first date that each of the successful artists joined the top-4 playlists, we can subsequently identify the historical number of playlists that each artist was part of, prior to becoming "successful". This helps us observe if the journeys are consistent across most artists in terms of growth momentum, or if some individuals are more likely to become successful, depending on the number of prior playlists that they have been included on.

The right-hand side of Figure 2 confirms there is no recognizable pattern for building growth momentum right before becoming successful. Some artists experience a downward trend days before becoming part of the top playlists, so there is no clear identifiable turning point for an artist's chances of success to increase.

The grey lines in the below figure show that for some artists, the presence across the number of playlists remains unchanged, whilst for others, the average number of playlists they appear in after becoming 'successful' increases considerably[2].

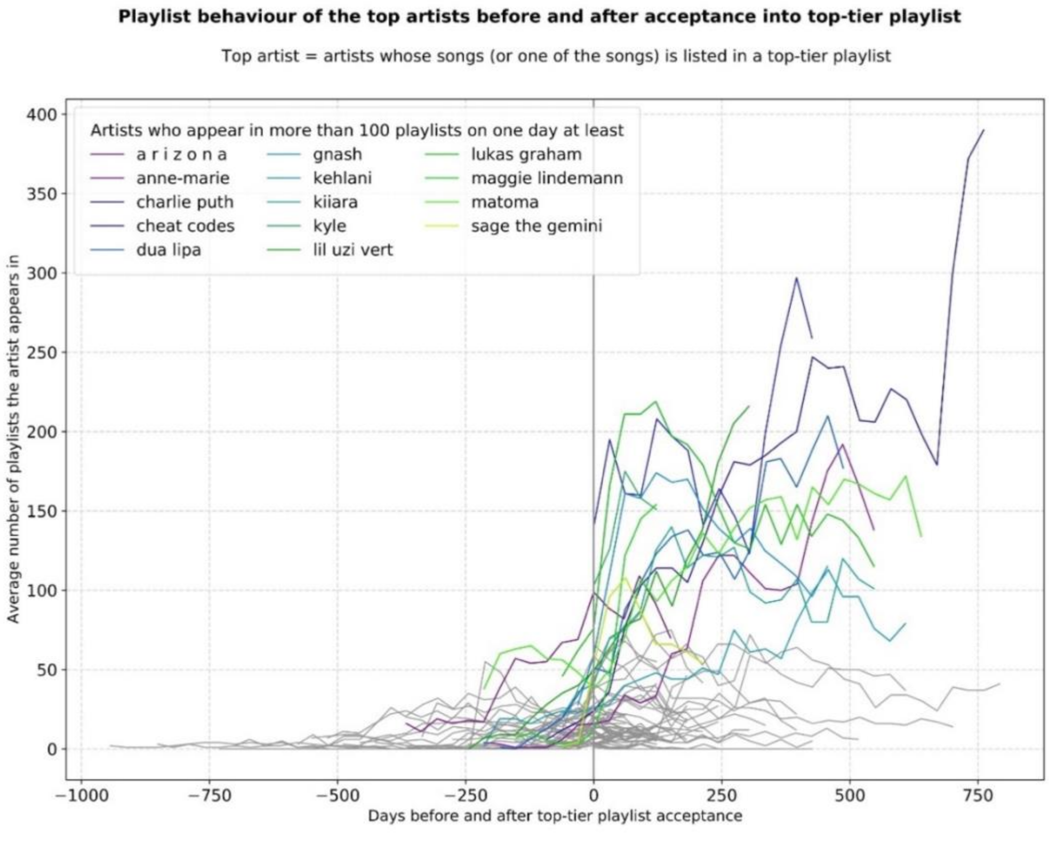

Figure 2: Number of times artists appear on a playlist before and after acceptance into a top-4 playlist.

[2] colourised lines in the chart

Focusing on the average number of streams per song for each artist leads to the same insights as described in the chart below.

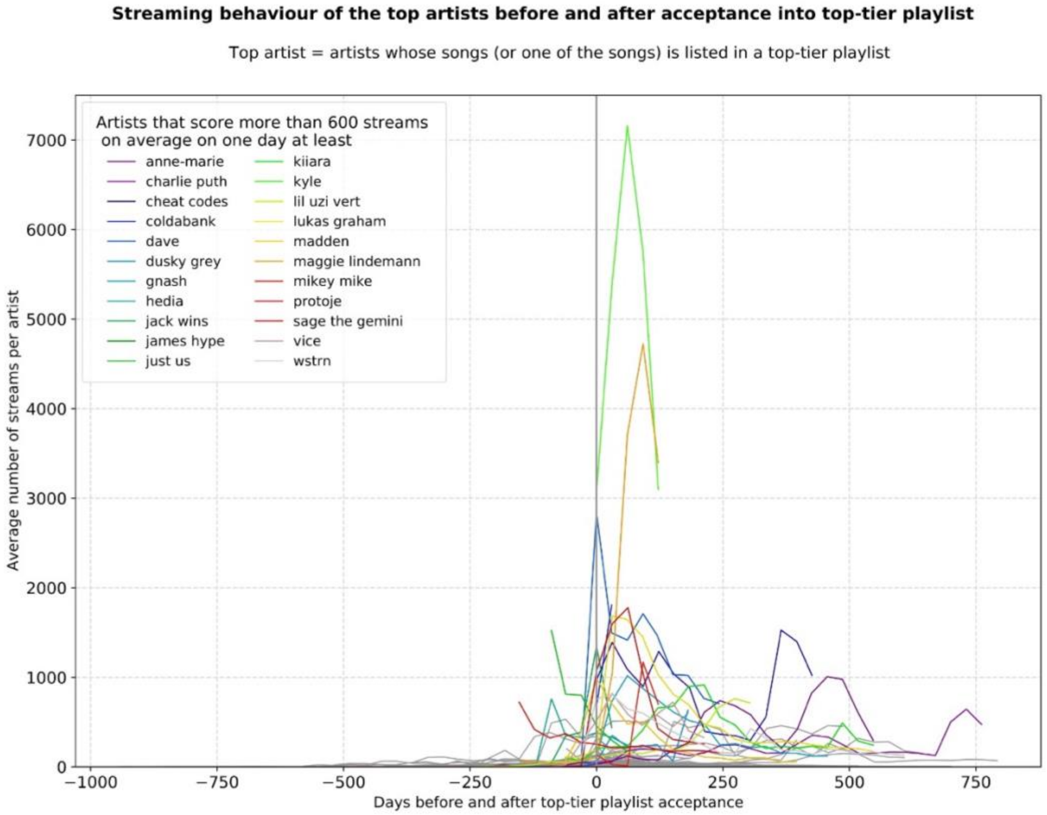

Figure 3: Average number of streams per day before and after acceptance into a top-4 playlist.

Gender domination

In the underlying dataset, 51.5% and 47.5% of the customers are from female and male listeners respectively, whereas 1% have not indicated their gender. Studies have shown that female listeners tend to share more if they find something appealing (Acquisti, 2006). Since this behavior would increase the likelihood of an artist's success, we analyse the artists of the top-4 playlists[3] by differentiating between the listener's gender (defined as gender domination).

For successful artists, the gender categories are almost equally distributed having 47% female-dominated artists and 53% male-dominated artists.

Among all artists, there is a tendency towards male-listener-dominated artists with 64.6%, and 35.3% for female-listener-dominated artists.

However, analysing the distribution for the top 25 artists based on how <u>often</u> they appear in <u>different playlists</u>, 68% of those artists are female-dominated, whereas only 32% of the artists are male-dominated. This is consistent with

[3] Bar-chart for dominant gender for successful artists and all artists can be found in the Appendix.

the 'sharing' behaviour mentioned above. Since being listed on many playlists can contribute to getting accepted into the top-4 playlists, the information of gender domination can serve as a valuable feature.

Generation domination

70.2% of the listeners are aged 21-37 years ("generation millennials"), which represents the largest group. This is followed by the 37-53 years old age group ("generation X", 15.2%), (0)-21 years old ("generation Z", 10.1%) and 53+ years old (generation boomer, 4.3%). 0.36% of the listeners did not state their age.

Analysing the artists' predominant listeners by generation (defined as generation domination), we get the following distribution:

Generation	Percentage share
Millennials	85.8%
Generation X	8.3%
Boomer	4.7%
Generation Z	1.1%
Unknown age	0.15%

Table 1: Age domination distribution across all artists

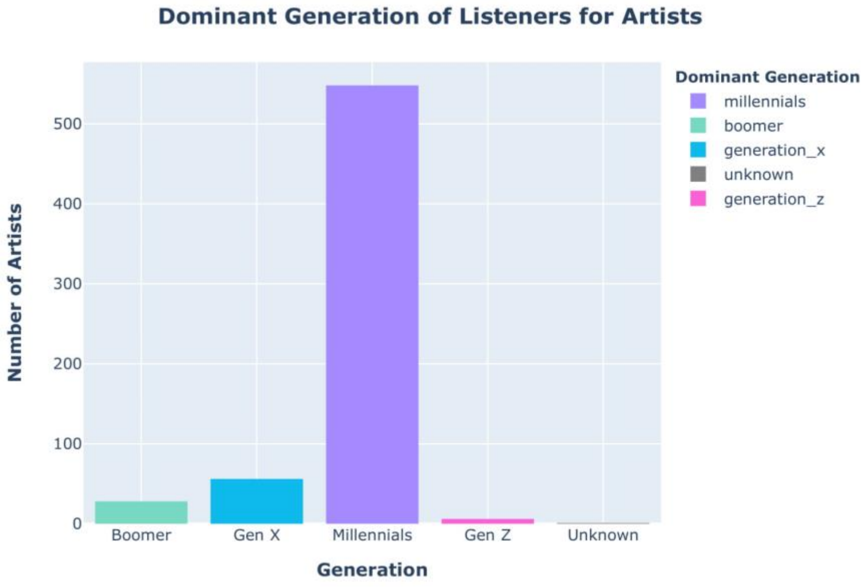

Figure 4: Distribution of generation domination across all artists with absolute values

Considering the successful artists, their listeners are dominated by the millennial generation (100%), who are more likely to use Spotify due to their ease of adopting new technologies. (Chan-Olmsted, 2020) For aspiring artists this could imply that chances of becoming successful would increase by targeting the dominant generation's preferences.

Season domination

Season domination refers to the season in which the artist is listened the most (either winter, summer, spring or autumn). We observe that among the successful artists, spring dominates with 51%, followed by summer with 31%:

Season domination	Percentage share
Spring	50.7%
Summer	31.0%
Winter	12.7%
Autumn	5.6%

Table 2: Season domination distribution among the artists in the top-4 playlists[4]

Contrastingly, looking at the season distribution of all artists, it becomes flatter:

Season domination	Percentage share
Spring	39.0%
Summer	37.2%
Winter	16.5%
Autumn	7.3%

Table 3: Season domination distribution across all artists

[4] Bar-chart for dominant seasons for successful artists can be found in the Appendix.

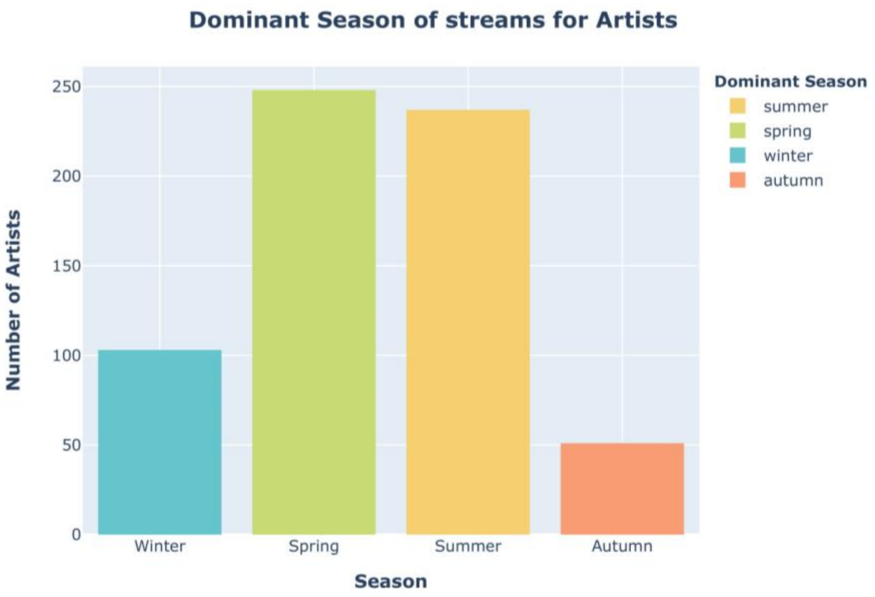

Figure 5: Distribution of season domination across all artists with absolute values

This insight suggests that songs belonging to successful artists are characterised by a corresponding mood for the respective season. Data on listener behaviour with regards to seasons can promote better illustration of a song's mood, and therefore in predicting the artist's future success.

Features per artist

Artists who feature other artists on their songs, might benefit from better success probabilities since they can leverage networking effects (fan base from other artist). The correlation coefficient between the number of features per artist and average stream count per artist overall is 0.3, which is a moderate, yet positive relationship. However, by focusing on the successful, we observe a correlation coefficient of 0.13, which indicates no strong relationship between these two factors.

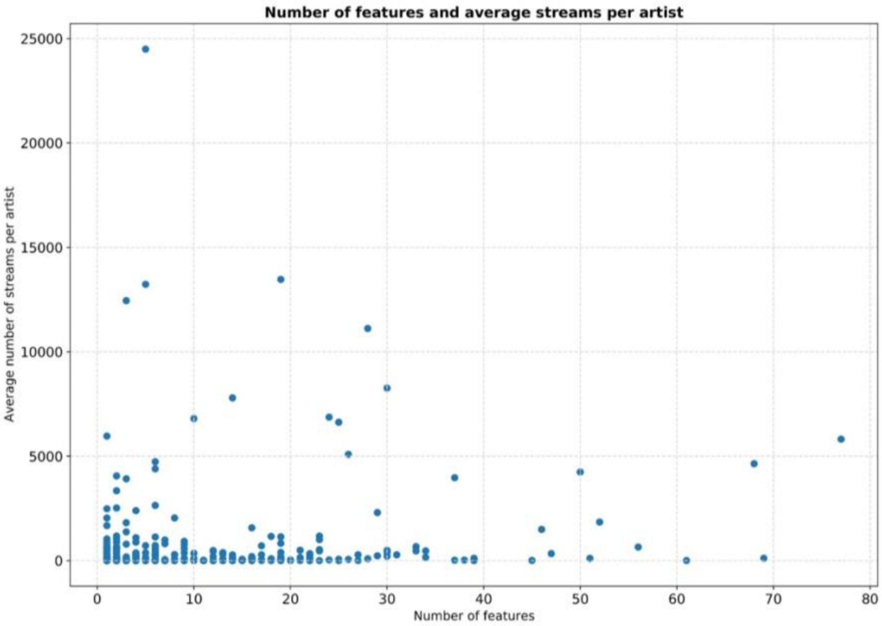

Figure 6: Average number of streams and number of song features per artist across all artists

Number of Songs

The number of songs captures the 'ambition' of an artist: the more songs an artist produces, the higher the likelihood of creating a hit or increasing recognition. Figure 7 indicates that successful artists tend to produce more songs than the overall average of 6. However, this can also be skewed by the fact that some artists have a longer tenure in music compared to others with fewer songs.

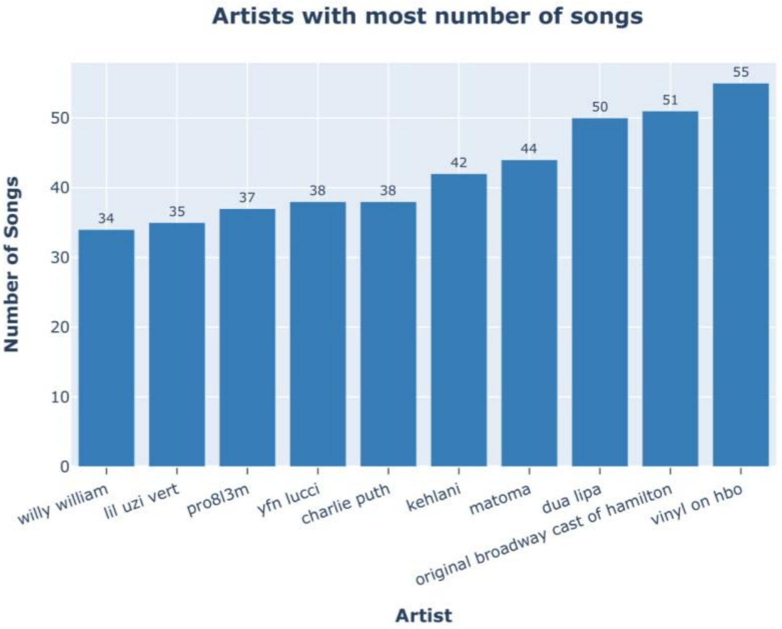

Figure 7: Number of songs per artist for artists with most songs.

Number of Playlists an Artist appears in

The number of playlists an artist appears in might be important: More appearances in various playlists could mean higher likelihood of reaching the top-4 playlists due to network effects. The chart below shows the successful artists vs number of playlists they appear in.

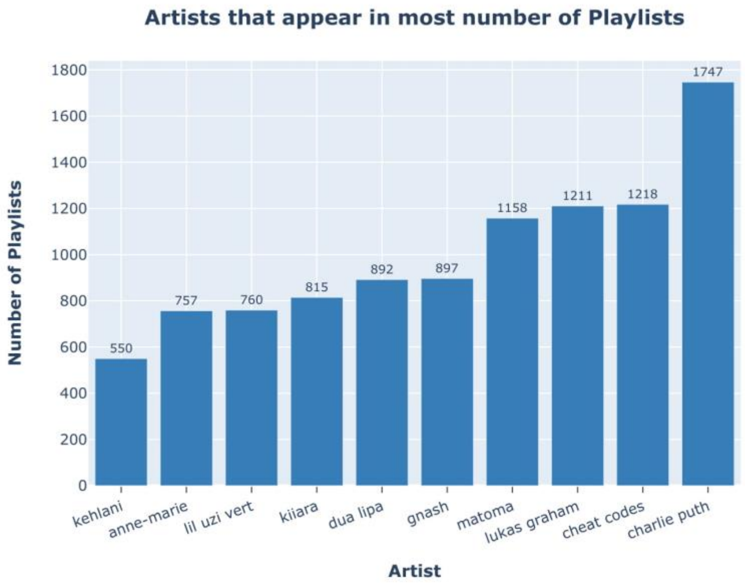

Figure 8: Top artists based on how many times they appear on a playlist.

Passion Scores per Artist

The passion score is a ratio of the stream counts and unique listeners of an artist. A high score is most likely achieved if the 'fan base' of the artist is relatively small, but streaming frequency is high, which may apply to new artists (see chart below) who are often 'shared' by users. This score serves as a good feature for identifying the 'sharing' potential of an artist and, thus, their predicted future success as their visibility increases.

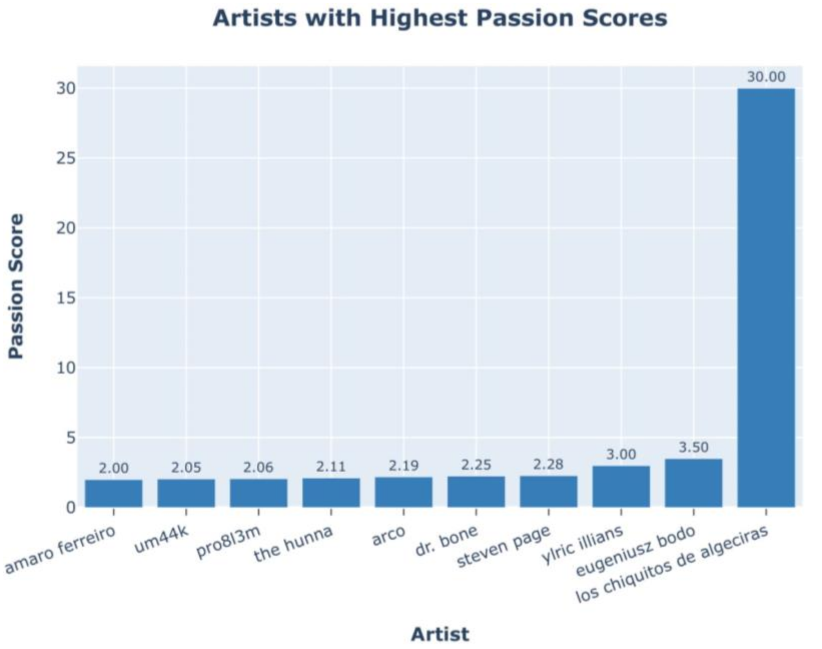

Figure 9: Artists with highest passion scores

Average Stream Length

The average stream length is the average customer playtime of a track per artist. This feature indicates if a song is listened until the end (e.g. high average stream time) and, thus indicating user's level of engagement. There might be cases where a song is highly advertised (generating high number of streams), but only listened partially. However, this may be biased by the length of the song itself, which can be seen in the chart below.[5]

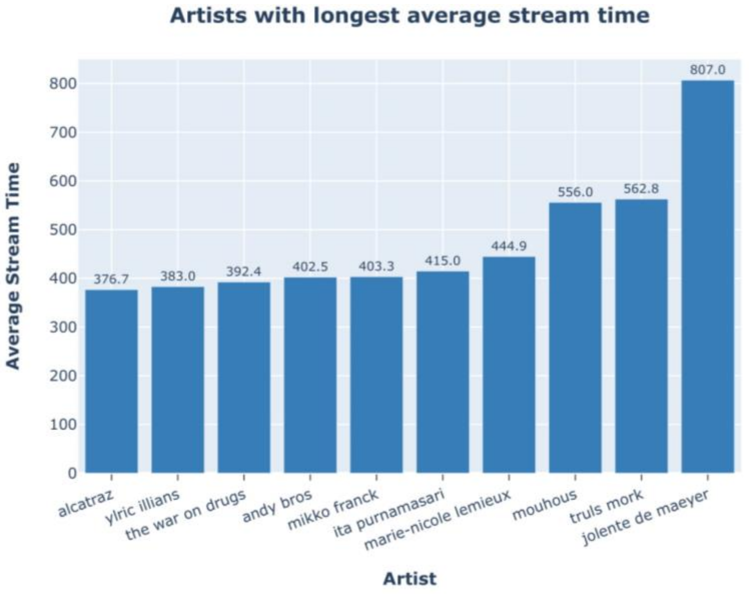

Figure 10: Top artists based on average stream time (units in seconds)

[5] Here, Jolente de Maeyer scores the highest. Her violin songs are lengthier than the average pop song. However, this also indicates that they are listened to the end.

Number of Repeat Streams

The number of repeat streams is the number of times a customer plays the same song more than once, disregarding time window. As the chart below shows, most successful artists have highest repeat streams, which outlines the importance of this feature.

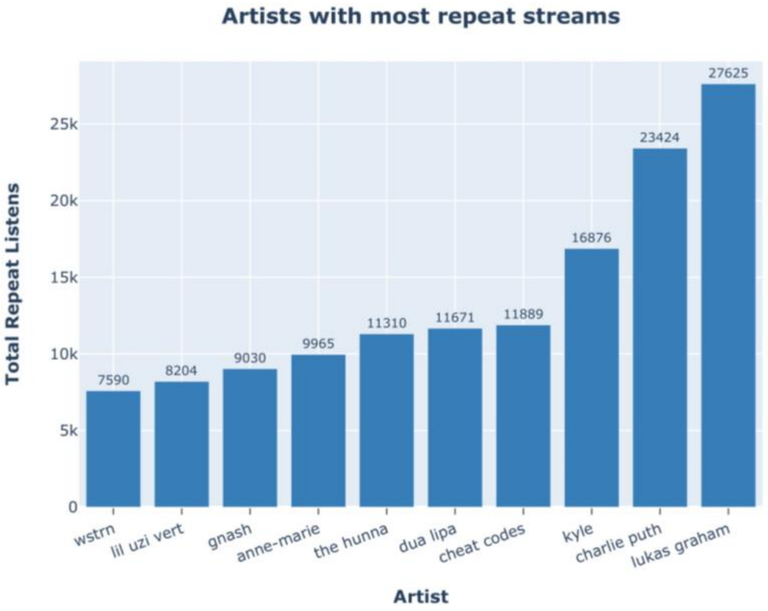

Figure 11: Top artists based on most repeat streams.

Weekday domination

Weekday domination illustrates the day on which the artist is listened to most. The tables below show that Friday peaks higher in the case of the successful artists (35.7% vs. 23.0%). This feature could be used for deducing the overall mood of a song, as Fridays are likely to prompt festive behaviours (in anticipation of the weekend) and therefore contribute to an artist's success if a song of a particular theme is released on a specific weekday.

Weekday domination	**Percentage share**
Monday	35.7%
Tuesday	5.7%
Wednesday	10.0%
Thursday	2.9%%
Friday	35.7%
Saturday	10.0%
Sunday	0%

Table 4: Weekday domination distribution across artists who appear in top-4 playlists.

Weekday domination	Percentage share
Monday	41.0%
Tuesday	4.7%
Wednesday	11.0%
Thursday	4.2%
Friday	23.0%
Saturday	14.6%
Sunday	1.6%

Table 5: Weekday domination distribution across all artists.

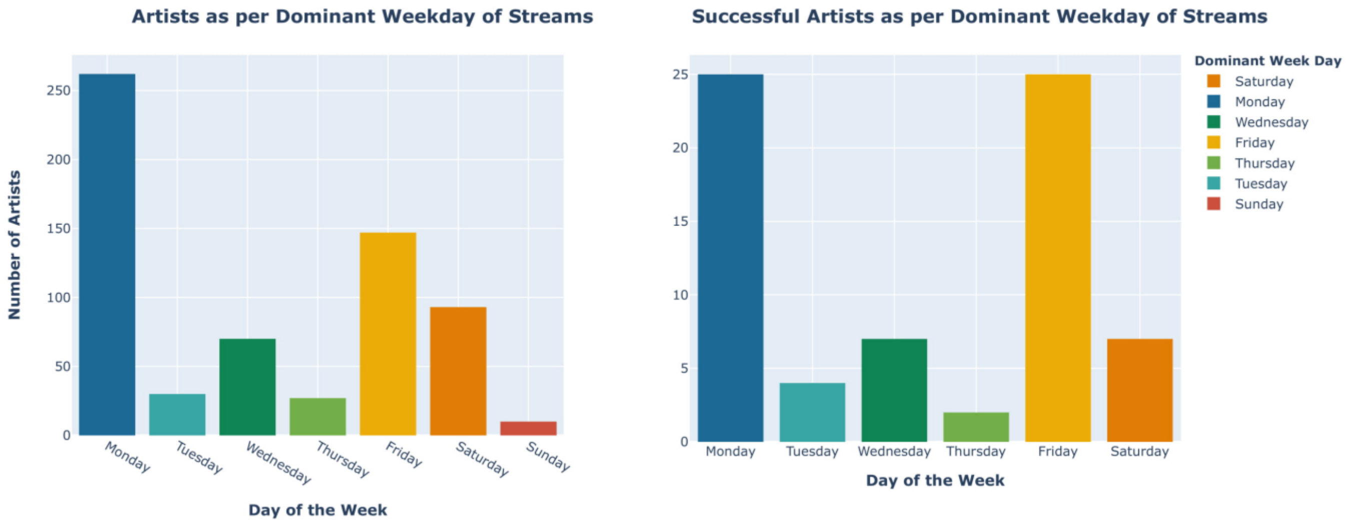

Figure 12: Distribution of dominant weekday for all artists (left side) and successful artists (right side)[6] with absolute values.

Figure 10 showcases that Mondays and Fridays are the dominant weekdays for artists, which can be due to Spotify's built-in playlist updating schedule. 'Discover Weekly' is a regularly updated playlist on Mondays, which explains the lower proportion (9.5% = 41% vs. 35.7%) of Monday-dominant artists who are successful, compared to 17% of the Friday-dominant successful artists, given that our measure of success for the playlists does not currently include 'Discover Weekly'. Our definition of success focuses on the 'New Music Friday' playlist instead.

[6] Sunday has 0% successful dominant-weekday artists. Refer to Table 4.

Time of Day

Time of day domination refers to the time when the artist is listened the most. The tables reveal that for successful artists the dominant time of day is mostly afternoon, whereas the distribution across all artists is flatter. This can again reveal information of the song's mood.

Time of Day domination	Percentage share
Morning	2.9%
Afternoon	95.7%
Evening	1.4%
Night	0%

Table 6: Day-phase domination distribution across artists who appear in top-4 playlists.[7]

Time of Day domination	Percentage share
Morning	16.0%
Afternoon	74.5%
Evening	7.5%
Night	2.0%%

Table 7: Day-phase domination distribution across all artists.

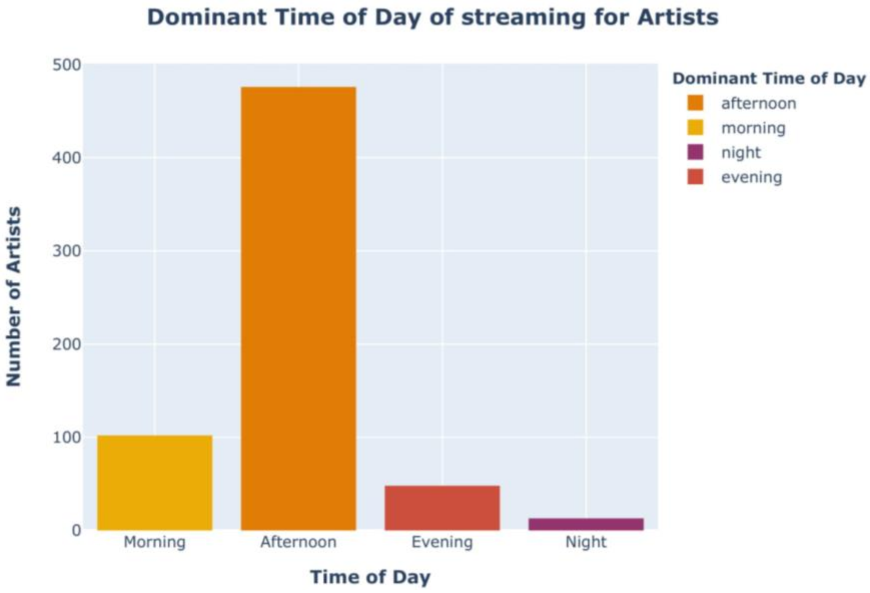

Figure 13: Time-of-day domination distribution across all artists with absolute values

[7] Bar-chart for dominant time of day for successful artists can be found in the Appendix.

4. Feature Engineering

From the insights gained above, we computed additional "summary" features from the dataset provided by Warner Music. Our aim is to create unique features for each artist and use these in our models to predict whether an artist would be successful.

Feature Name	Description
Number_songs	Number of songs per artist in the dataset
Playlists	Number of playlists an artist appears in
Passion_score	The average passion score of each artist (streams/users)
Avg_stream_time	Average stream time per artist
Repeat_count	Number of customers listening to same song more than once per artist
Gender_domination	The gender that streams each artist the most
Generation_domination	The generation that streams each artist the most
Season_domination	The season where each artist is streamed the most
Weekday_domination	The weekday where each artist is streamed the most
Dominant_dayphase	The time of day the artist has been streamed the most
Featuring_artists	Average featuring artists per song per artist

After checking multi collinearity, we noticed that number of streams and number of unique users per artist had very high variance inflation factor (348 and 300 respectively).

This implies high correlation with one or more of the other features and would make the final predictions very sensitive to minor changes in the models. Additionally, passion score is a function of both streams and unique users, so we removed those variables before running the models.

5. Model Performance Evaluation

We focus on (1) accuracy and (2) precision as our two metrics of interest. The accuracy score describes the model's ability to correctly predict success, whereas precision helps identify the chances of artists actually becoming successful. Warner Music is eager to focus on discovering artists who have the potential of becoming successful, in order to maximise ROI and reduce risk of financial loss due to misplaced investments. Therefore, in the recall-precision trade-off, we prioritise higher precision values, with the aim of avoiding false positives, saving valuable resources and ensuring higher hit rate.

Metric	**Description**[8]
Accuracy	(TP +TN)/(TP+FP+FN+TN): is the proportion of correct predictions out of total predicted values.
Precision	TP / (TP + FP): Precision measures the model's ability to identify success i.e out of the total predicted success cases, how many actually turned out to be successful
ROC Area under curve	AUC ROC measures the degree of separability. It communicates the model's capability of correctly predicting unsuccessful artists and correctly predicting successful artists.
Recall	TP / (TP + FN): Recall measures the model's ability to correctly identify the relevant data – the True Positives.
F1	$2 * \frac{Precision*Recall}{Precision+Recall}$: F1 is the harmonic mean between Precision and recall and is mainly used in comparing the performance of various models.

Table 8: Description for metrics used to evaluate the model's performance.

After finetuning our chosen machine learning models for predicting artists' success, we compare their performance by using the above metrics:

Model	**Performance**				
	Accuracy	ROC Area Under Curve	Precision	Recall	F1
Logistic regression	0.94	0.85	0.73	0.73	0.73
Random forest	0.88	0.65	0.40	0.36	0.38
AdaBoost classifier	0.85	0.59	0.27	0.27	0.27
Gradient boosting classifier	0.90	0.70	0.50	0.45	0.48
XGBoost	0.88	0.85	0.45	0.82	0.58

Table 9: Model performance results.

[8] TP: true positive, TN: true negative, FP: false positive, FN: false negative

Table 9 summarises the model evaluation results by comparing the values from the 'actual' test set to the predicted values. We observe that Logistic Regression records an Accuracy of 94,3% and a Precision of 72,7%, whilst also maintaining other performance metrics high. This implies that our finetuned model accurately predicts the artists' success, thus maximising our ability to make viable recommendations for Warner Music. With the below Confusion Matrix, we can confirm that the model has a very low rate of predicting 'success' when the true outcome is 'not success', which is in line with our objective. Consequently, having a high accuracy as well as a relatively high precision score contributes to making safer investment decisions as the number of false positives can be kept relatively low.

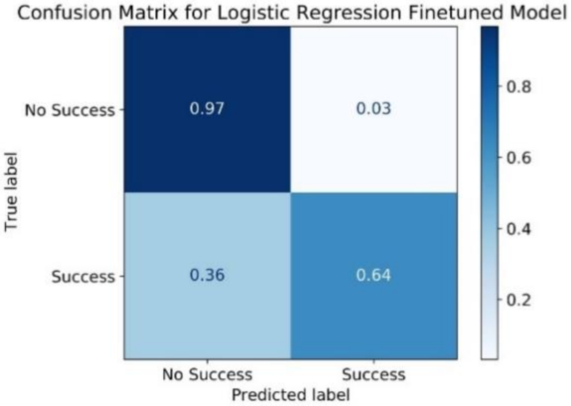

Figure 14: Confusion matrix (i.e. Error Matrix) for the Logistic Regression model with tuned parameters.

6. Feature importance and Implications

The figures below summarise the top & bottom 10 features from the Logistic Regression model with regards to feature importance. These insights contribute to identifying customer segments and time periods that would be most prevalent, in order for Warner Music to enable aspiring artists to plan releases accordingly, and focus on driving success for their existing artists.

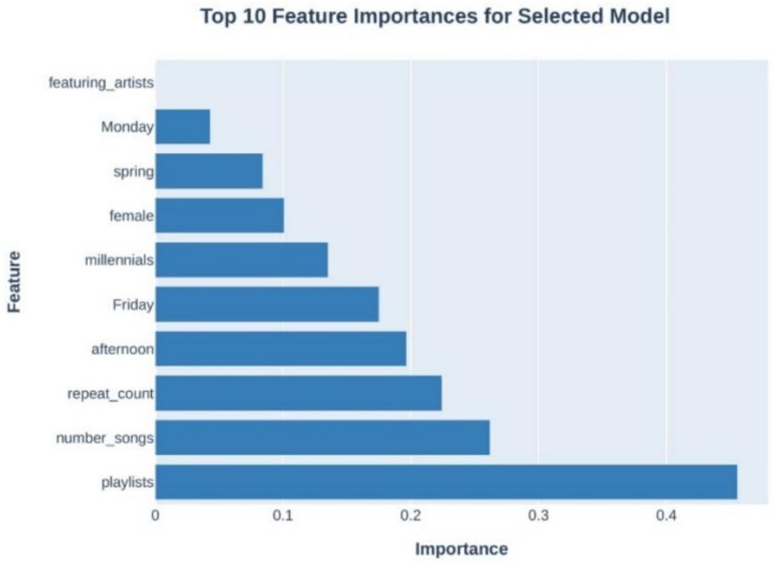

Figure 15: Top 10 most important features

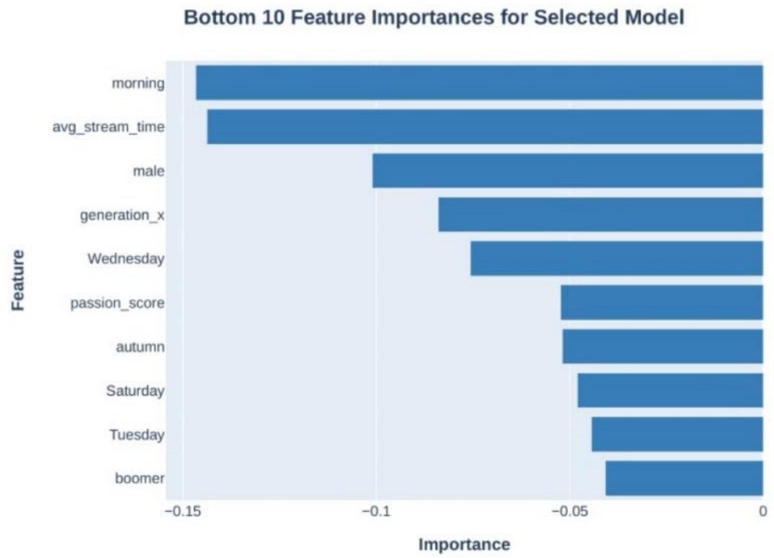

Figure 16: 10 least important features

Above, we can observe that female millennial customers are the most important customer segment, whilst new song releases and marketing efforts should be scheduled for Friday afternoons. This is supported by the stronger 'sharing' behaviour [9] of female listeners, as previously stated. In addition, we see that the number of songs an artist releases

[9] Sharing on social media or sharing with close network

and number of playlists they appear on will increase their likelihood of becoming successful. This confirms the value of encouraging social behaviours, like 'sharing' amongst networks, which can be achieved via targeted marketing strategies (e.g., aimed at female listeners).

7. Conclusion and further improvements

Conclusion

The purpose of this report is to predict whether an artist will become successful by analysing calculated artist features from a selected dataset containing streams from Spotify. We are able to predict an artist's success with an accuracy of 94.33% and a precision of 72.72%, which is acceptable. We, also, identified key customer segments, namely millennial females, for Warner Music to focus on with regards to promoting their existing artists and enhancing their success.

Lastly, Warner Music could have more success by focusing marketing efforts for artists on Friday afternoons. Given the Friday domination of successful artists, we can deduce that this may be driven by the mood or genre of the song itself, so we would recommend that Warner Music conducts further research into establishing dominant song moods, in order for aspiring artists to match consumer preferences and increase chances of success.

Improvements

Although we reached a satisfactory precision rate, there are still other features or improvements we could consider for future research:

1. We disregarded regional differences, as a large company would likely not focus on one segment in a specific region in the UK. With further data, exploring trends between different countries or geographical features and their "key customer segment" would be recommended.
2. Since a set sample size was chosen, we could not capture all historical data[10]. Hence, the model might lack generalization due to not consider a fully representative data set.
3. Incorporating features outside of the Spotify context, such as social media metrics, would contribute to a more holistic model and would allow for better incorporation of real-world dynamics of the music industry.

[10] Histogram for yearly stream counts showing uneven split can be found in the Appendix.

8. Appendix

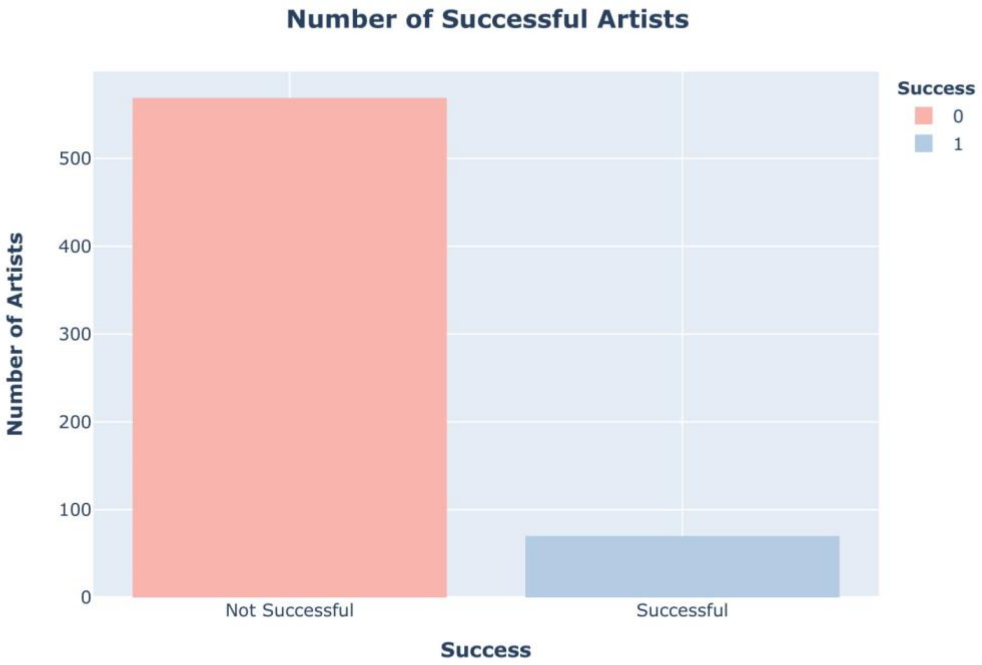

Figure 17: Proportion of successful and not successful artists

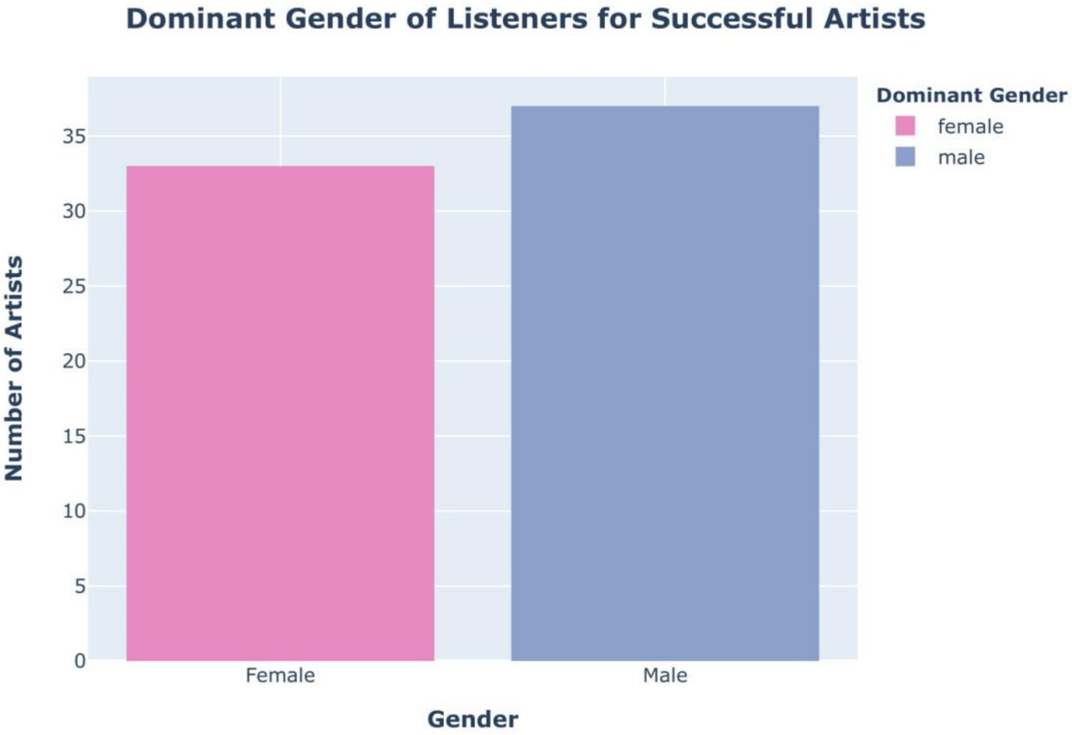

Figure 18: Dominant gender split for successful artists.

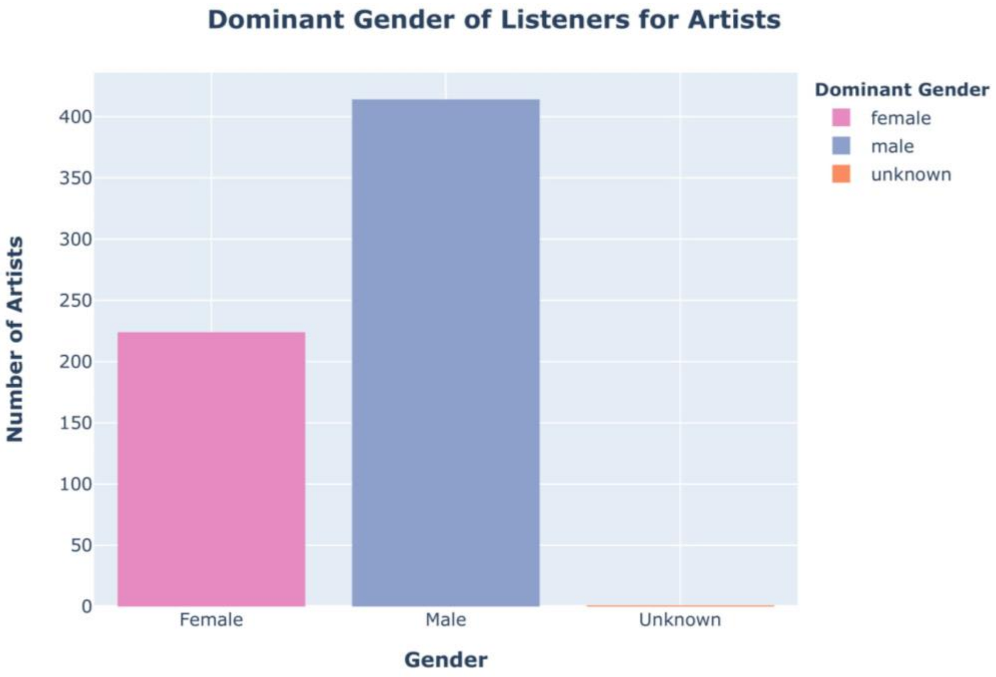

Figure 18: Dominant gender split for all artists.

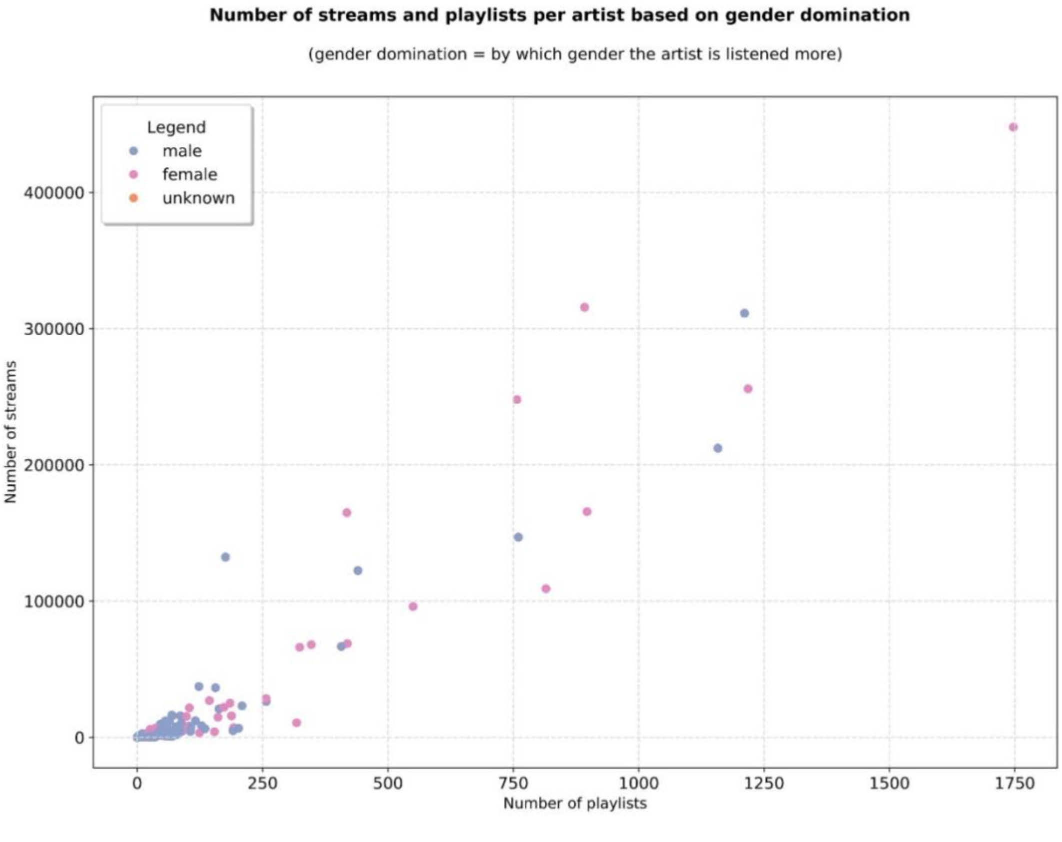

Figure 19: Scatter plot of number of streams against number of playlists based on gender domination.

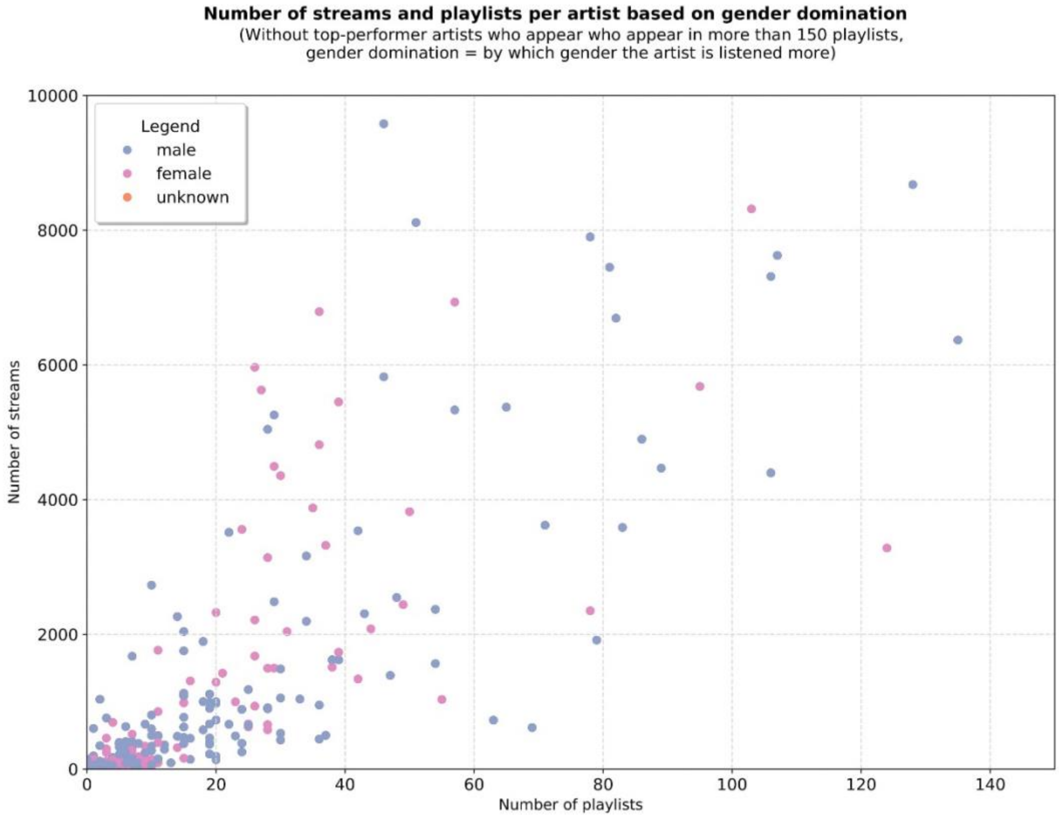

Figure 20: Scatter plot of number of streams against number of playlists based on gender domination, excluding artists who appear in more than 150 playlists.

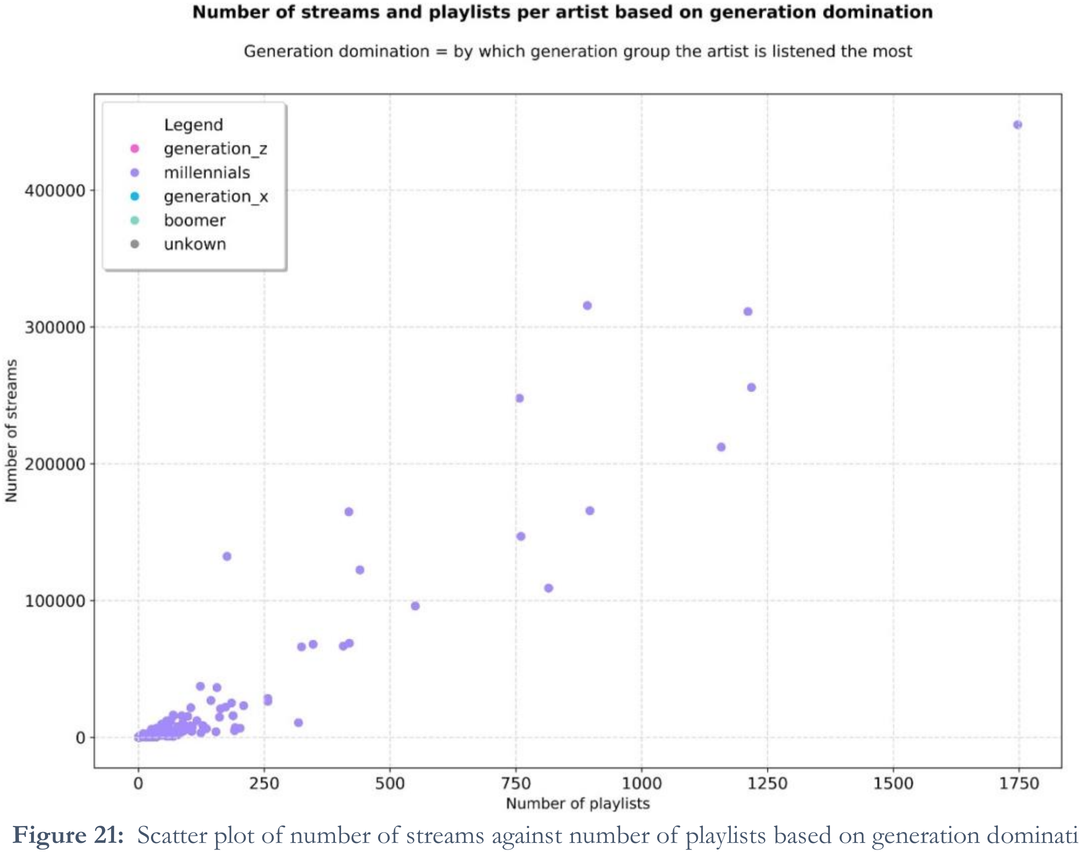

Figure 21: Scatter plot of number of streams against number of playlists based on generation domination.

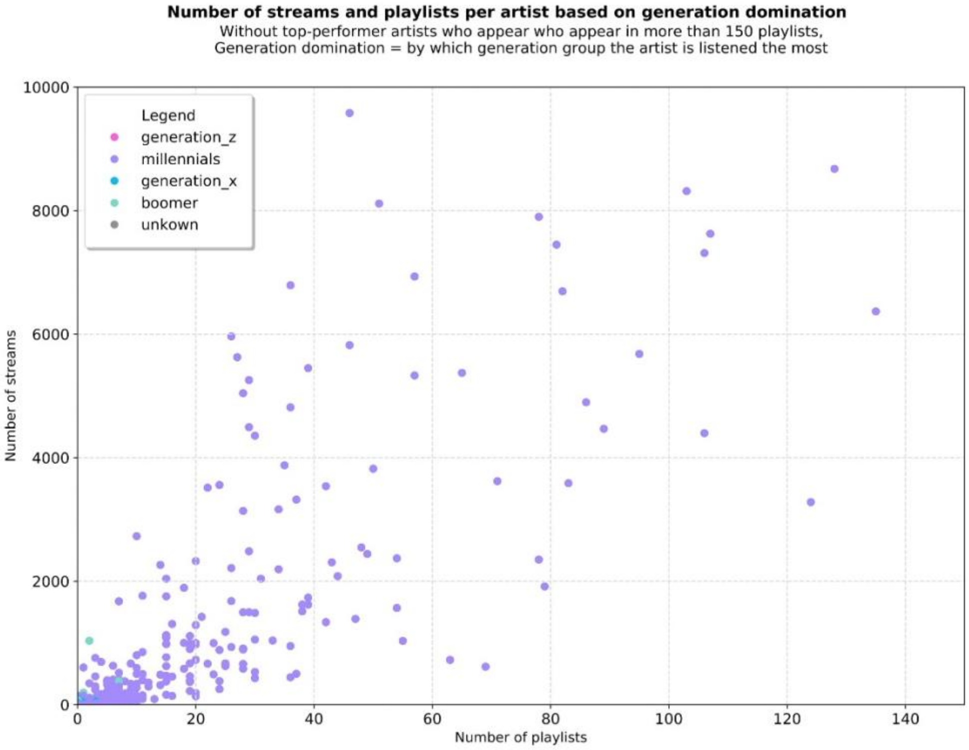

Figure 22: Scatter plot of number of streams against number of playlists based on generation domination, excluding artists who appear in more than 150 playlists.

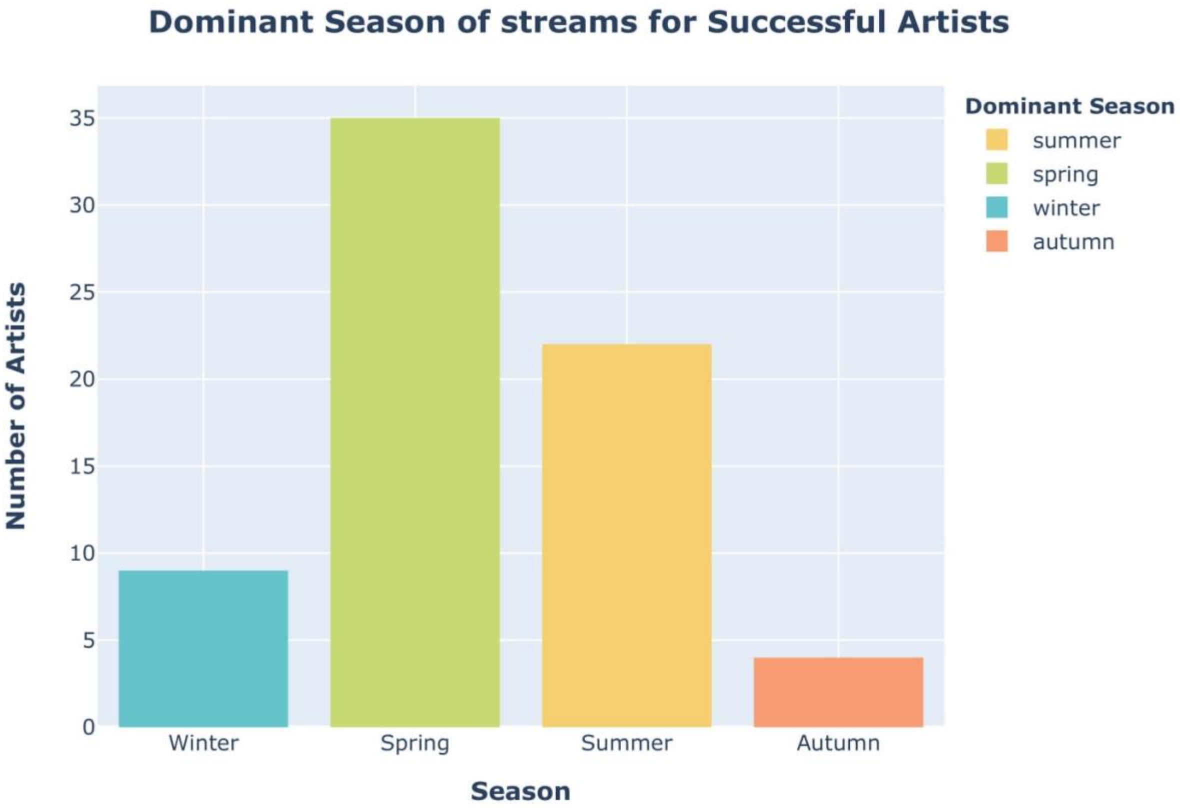

Figure 23: Number of successful artists as per each dominant season.

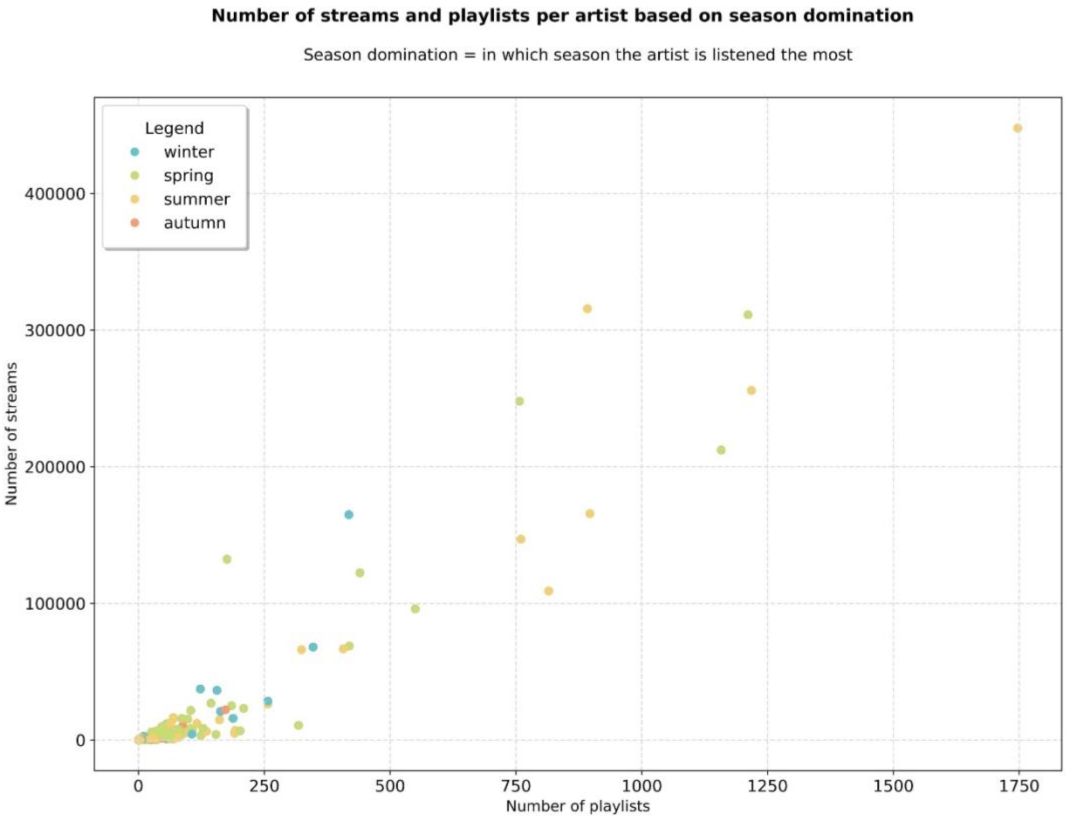

Figure 24: Scatter plot of number of streams against number of playlists based on season domination.

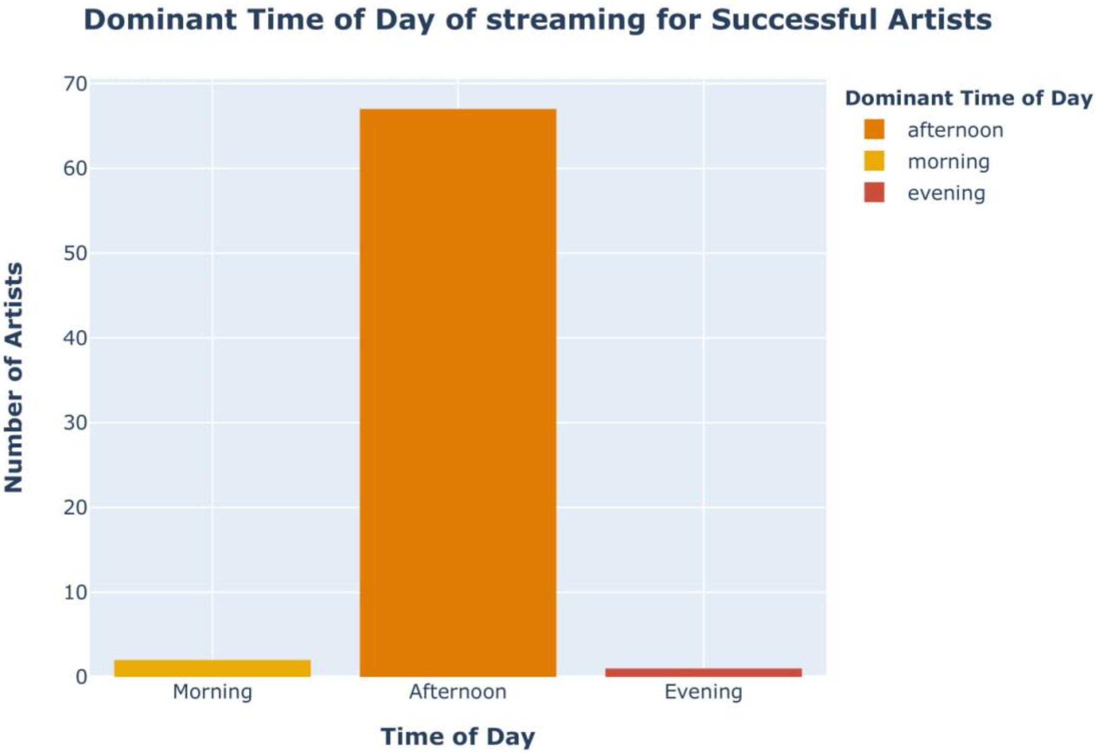

Figure 25: Number of successful artists as per dominant day-phase.

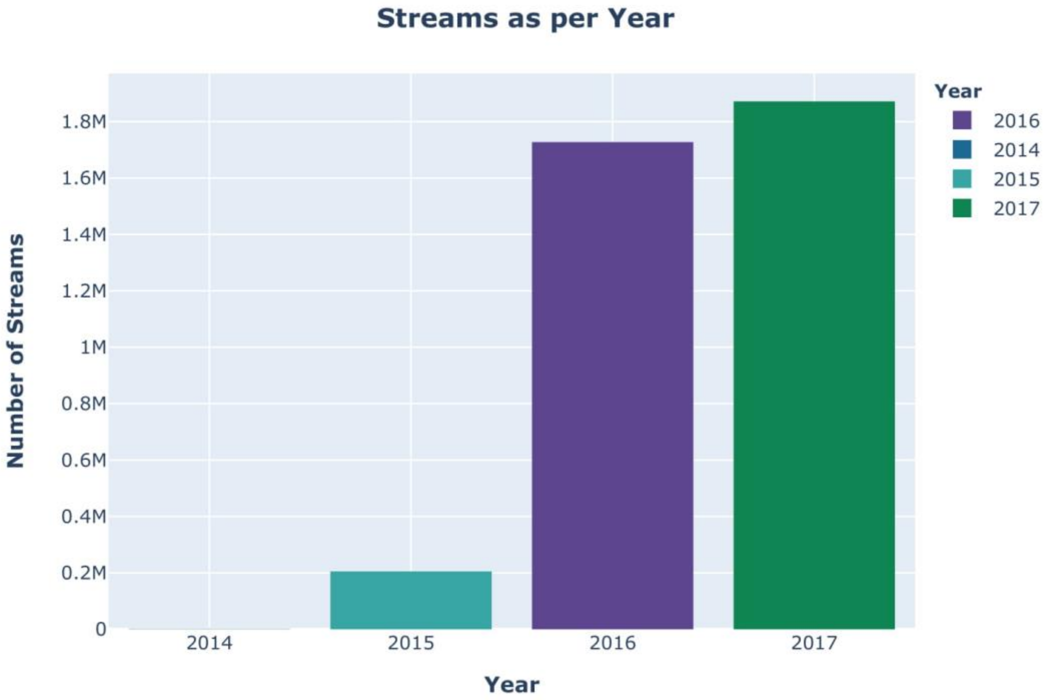

Figure 26: Stream counts as per year in the dataset provided.

Bibliography

Acquisti, A. &. (2006). Imagined communities: Awareness, information sharing, privacy on the Facebook. *Privacy Enhancing Technologies, Lecture notes*, 36-58.

Chan-Olmsted, S. W. (2020). Millennials' Adoption of Radio Station Apps: The Roles of Functionality, Technology, Media, and Brand Factors. *Journalism & Mass Communication Quarterly*, 9.