MSc Business Analytics
MSIN0097 Predictive Analytics 2021-2022

Title of Project: Prediction of Artists' Success on Spotify

Team Name/Letter: Group 21

Word Count: 1933

Disclaimer:

We hereby declare that this dissertation is my individual work and to the best of my knowledge and confidence, it has not already been accepted in substance for the award of any other degree and is not concurrently submitted in candidature for any degree. It is the end product of my own independent study except where other acknowledgement has been stated in the text.

General marking guidelines

85+	Outstanding work of publishable standard.
70-84	Excellent work showing mastery of the subject matter and excellent analytical skills.
60-69	Very good work. Interesting analysis with original insights. Some minor errors.
50-59	Good work which only covers a basic analysis. Some problems but no major omissions.
40-49	Inadequate work. Not sufficiently analytical. Some major omissions.
39-	Work is seriously flawed. Lack of clarity and argumentation. Too descriptive.

Mark: _______

UCL SCHOOL OF MANAGEMENT

Prediction of Artists' Success on Spotify

MSIN0097 Predictive Analytics - Group Assignment

Table of Contents

1. Introduction

1.1 Framing Business Problem

In recent years, online digital music streaming services have brought people the chance to listen to music without buying CDs at traditional offline shops. Such a new way of streaming music has generated a large amount of information regarding listeners, artists, and playlists. Therefore, such a phenomenon also requires a data-driven approach to analyse artists' success and popularity on online streaming platforms. By applying such analysis, record labels like Warner Music will be more likely to identify high potential artists as early as possible, and sign them so that these labels can allocate their resources and effort, in hopes of increasing their return on investment of signing the artists. Artists themselves can also adopt potential improvements based on listeners' preference accordingly.

One of the databases Warner Music Group owns as one of the biggest global music groups belongs to Spotify, which is the research platform target in this report. The dataset would be used to predict the success of artists on Spotify.

Generally, Spotify deals with nearly 1 billion daily streams (Dredge, 2015). The primary dataset used here is a sample of records of every stream from 2015 - 2017. It is found that some playlists (i.e. songs can be grouped under a list which can be saved, accessed and renamed by users at any time) have a large influence on the stream count, popularity as well as the future of songs, as listeners may rely on specific playlists to discover their potential favourite songs. This is also the reason why artists always seek to achieve top rankings in some key playlists, which means they can attract more attention from the public.

1.2 Using Data to Solve the Problem

Hence, to measure an artist's success on Spotify, a binary dependent variable is created (1 = streamed song is featured on at least one of the 4 key playlists - Hot Hits UK, Massive Dance Hits, The Indie List and New Music Friday, otherwise 0). Thus the problem would be a supervised classification task.

2. Data Understanding & Visualisation

2.1 Data Understanding

In the dataset, there are over 3.8 million observations in total and each row represents a recorded stream, including information regarding listeners (gender, age, region), streams (length, source, device, time) and songs (the name of artist, track, album, playlist). There are a lot of missing values within several variables (e.g. gender, age, playlist).

2.2 Data Visualisation

Data visualisation reveals the details of the source where listeners find the streamed music, the stream pattern during a day, a week, and a year, their streaming devices and so on.

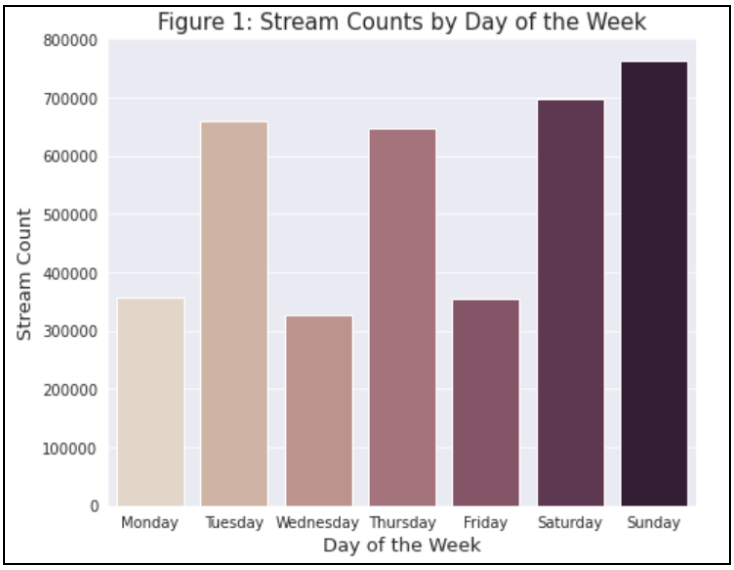


Figure 1: Stream Counts by Day of the Week

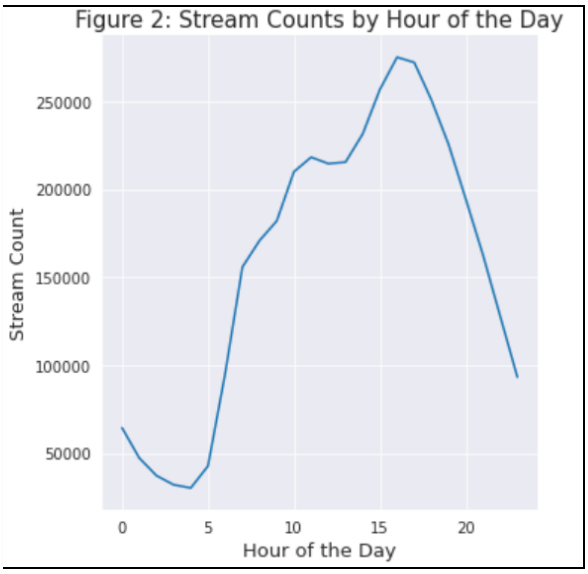


Figure 2: Stream Counts by Hour of the Day

According to Figure 1, stream counts are highest during the weekend. Figure 2 reveals the stream count throughout the day. It reaches a peak at around 4PM and is lowest at around 4AM.

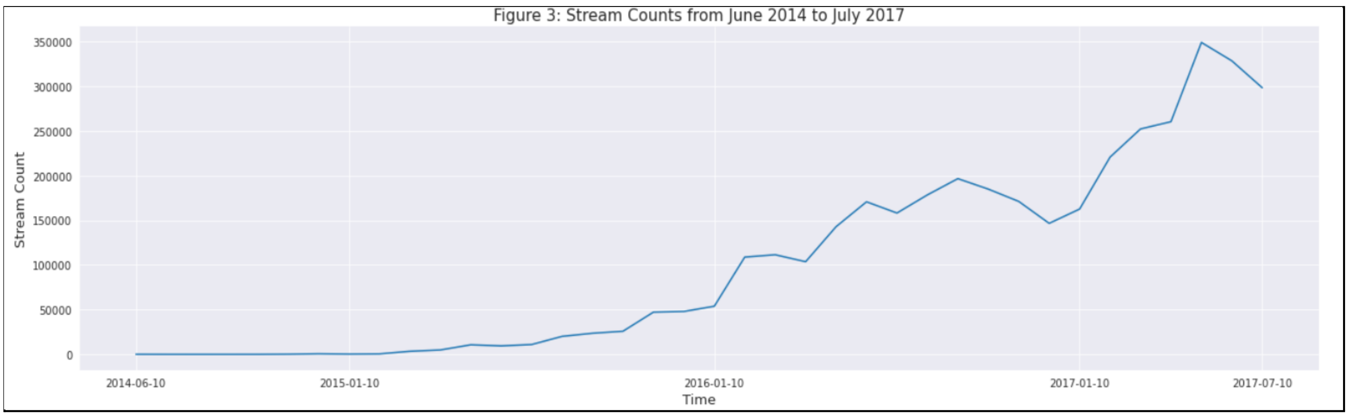


Figure 3: Stream Counts from June 2014 to July 2017

Based on Figure 3, across 2014-06-10 to 2017-07-10, the stream count started to gradually increase (from less than 50,000 daily stream) from the second half of 2015 onwards, hitting a crescendo of around 350,000 streams later in 2017, which reflects the rapid development of Spotify stream service.

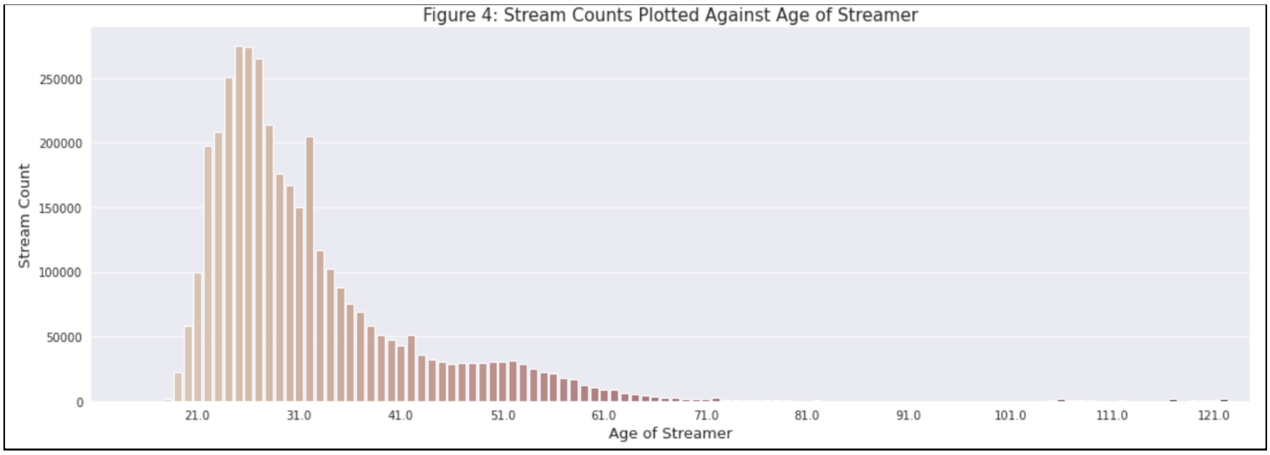
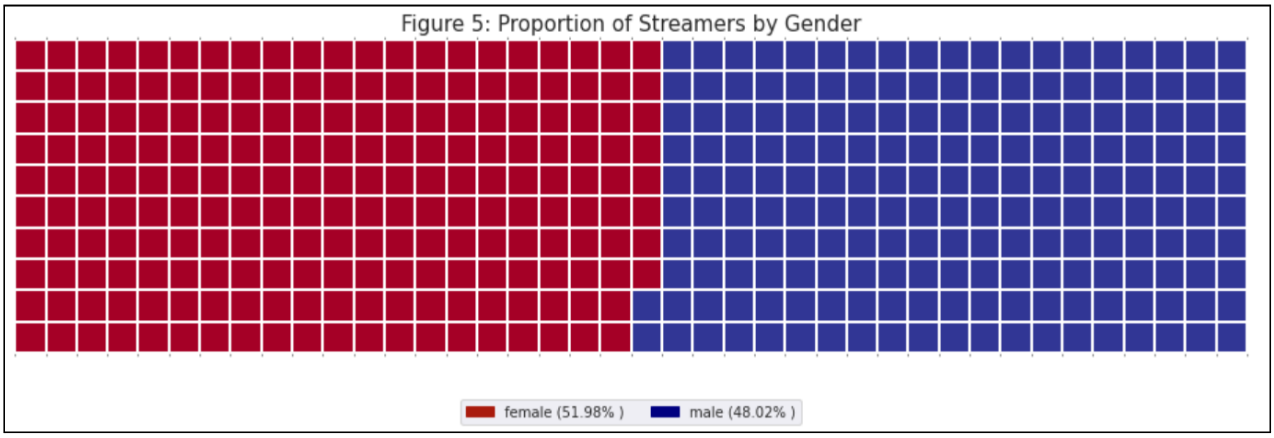
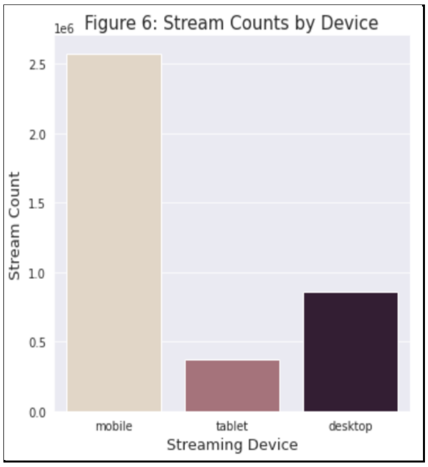
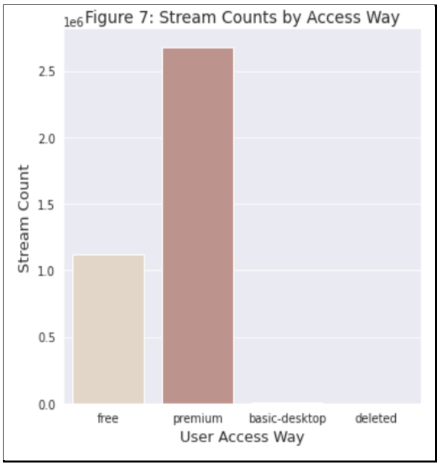


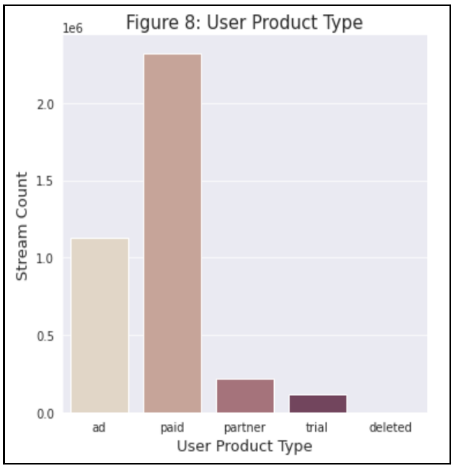
Figure 4 shows the stream count among ages of listeners, which mainly concentrates between 20 to 30 years old. Simultaneously, there are some age outliers as they extend way beyond 100, which is unreasonable and provides few insights about the listeners, they are removed during later stages.



Besides, according to Figure 5, the gender proportion for all listeners is fairly balanced (51.98% female and 48.02% male).

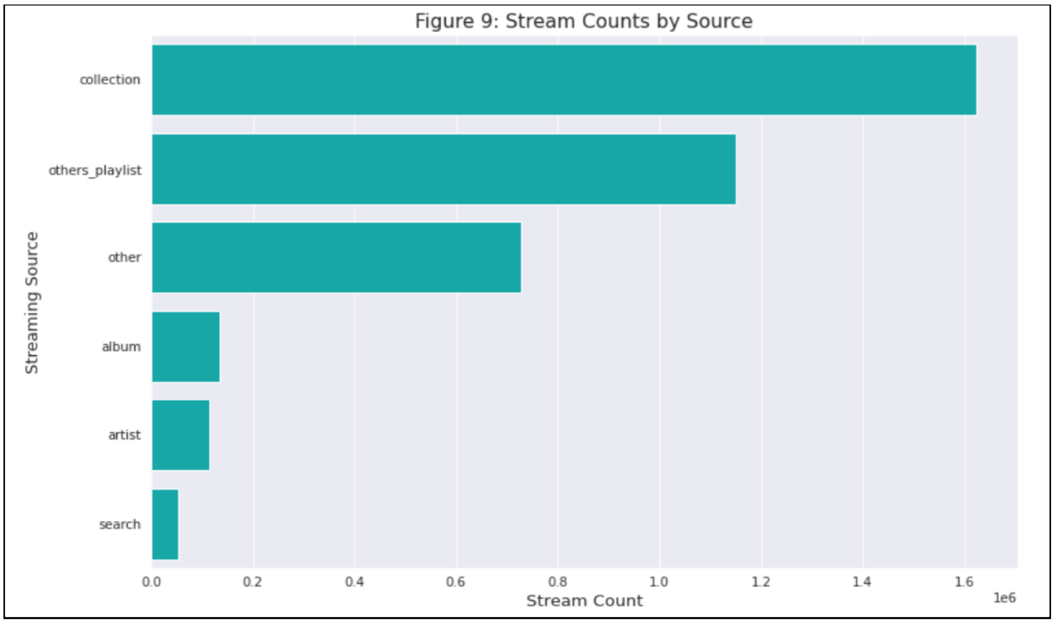






From Figure 6, other stream relevant information has been plotted. Most songs are streamed on mobile devices. Furthermore, according to this dataset, premium subscribers stream the most (Figure 7), which is also reflected within paid users in Figure 8.

When it comes to the stream source (Figure 9), it can be found that most streams were accessed through listeners' local collection (playlists they created on their own) and others' playlists.



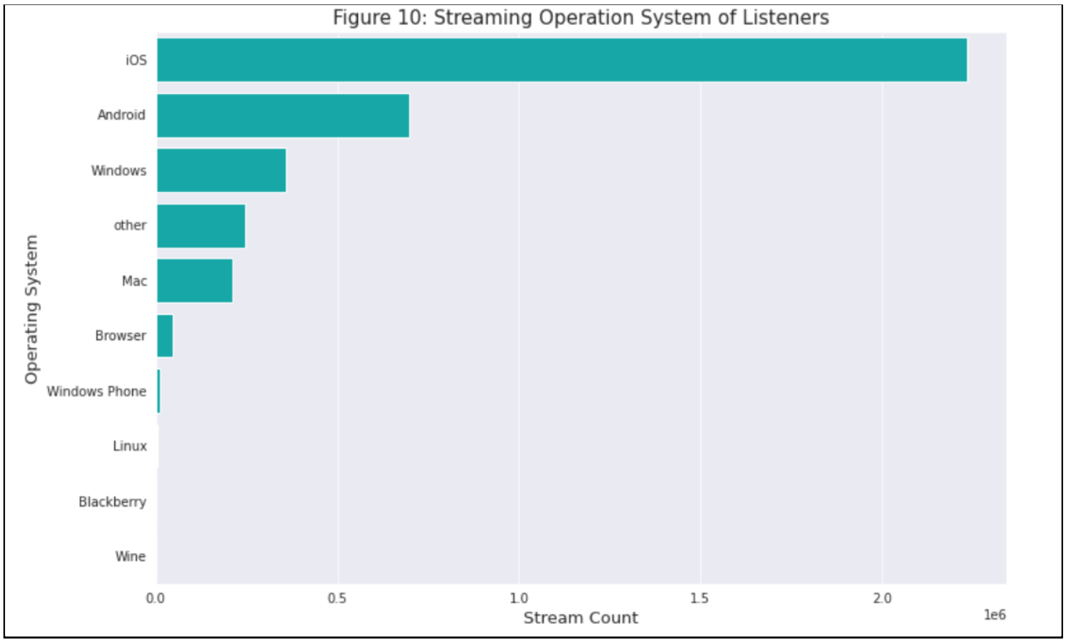
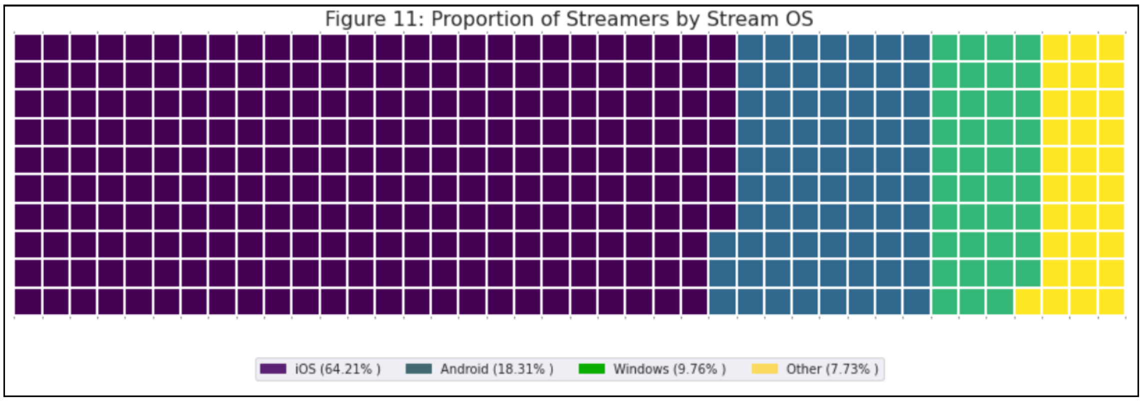


Figure 10 reveals that most streams were accessed through iOS, followed by Android and Windows. After further grouping, the Apple system still dominates (Figure 11).



3. Data Preparation and Feature Engineering

3.1 Feature Engineering

For data preparation, some artists were duplicated due to discrepancies in capitalised letters. This was handled by converting them to lowercase. For the age of the streamers, those with age above 100 were filtered, since it was unrealistic.

As a classification problem, the dependent variable assessed a particular artists' by determining whether or not they were featured in the four key playlists. Thereafter, feature engineering was performed on three dimensions:

- Artist features involved generating total stream counts, total unique customers and artist passion score.

	artist_name	stream_count	customer_id	passion_score
0	Charlie Puth	445222	364964	1.219907
1	Dua Lipa	314389	259731	1.210441
2	Lukas Graham	309379	246075	1.257255
3	Cheat Codes	254804	224761	1.133666
4	Anne-Marie	246783	219394	1.124839
...	...	...	...	...
633	Giuseppe Gibboni	1	1	1.000000
634	Frederik Leopold	1	1	1.000000
635	Rebecka Karlsson	1	1	1.000000
636	Sinfonia Varsovia Brass	1	1	1.000000
637	Nicolas Motet	1	1	1.000000

638 rows × 4 columns

(Figure 12: Artist features)

- Playlist features take each artist's prior playlists to generate prior playlist stream counts, unique users as well as playlist passion score. The net effect of the prior playlists is obtained by taking the mean of passion scores for each artist.

	artist_name	passion_score
0	#90S Update	1.000000
1	17 Memphis	1.000000
2	3Js	1.000000
3	99 Percent	1.000000
4	A Boogie Wit Da Hoodie	1.102682
...	...	...
462	Yvng Swag	1.000000
463	Zac Brown	1.000000
464	Zak Abel	1.015512
465	Zarcort	1.000000
466	Zion & Lennox	1.046235

467 rows × 2 columns

(Figure 13: Playlist features)

- User base features consisted of profiling each artist's user base through gender split as well as percentage breakdown of age groups for each artist.

	artist_name	female
0	#90S Update	0.437500
1	17 Memphis	0.666667
2	2D	0.000000
3	3Js	0.200000
4	99 Percent	0.677953
...	...	...
633	Zak Abel	0.529985
634	Zakopower	0.000000
635	Zarcort	0.200000
636	Zbigniew Kurtycz	0.000000
637	Zion & Lennox	0.537792

638 rows × 2 columns

(Figure 14: User-base features, gender breakdown)

artist_name	stream_count	customer_id	passion_score_artist	passion_score_playlist	female	% of Youth	% of Young Adults	% of Adults	% of Middle Age Adults	success
Charlie Puth	445222	364964	1.219907	1.037291	0.578163	0.207576	0.372672	0.288190	0.126399	1
Dua Lipa	314389	259731	1.210441	1.044375	0.594638	0.158290	0.379690	0.336884	0.122224	1
Lukas Graham	309379	246075	1.257255	1.040604	0.480843	0.193636	0.377125	0.299011	0.125991	1
Cheat Codes	254804	224761	1.133666	1.029984	0.547597	0.204498	0.439692	0.266716	0.086832	1
Anne-Marie	246783	219394	1.124839	1.016233	0.602962	0.195595	0.387664	0.303646	0.110103	1

(Figure 15: DataFrame head with Artist, Playlist and User-base Features)

3.2 PCA

After creating the artist, playlist and user-based features, the location of the listeners was considered, and how they could correlate with the success of the artist. Due to the large number of regions in GB, PCA was performed in order to get 10 principal components that essentially give 10 features of linear combinations of the locations of the listeners.

```
Shape of dataset to perform PCA on is: (627, 514)
```

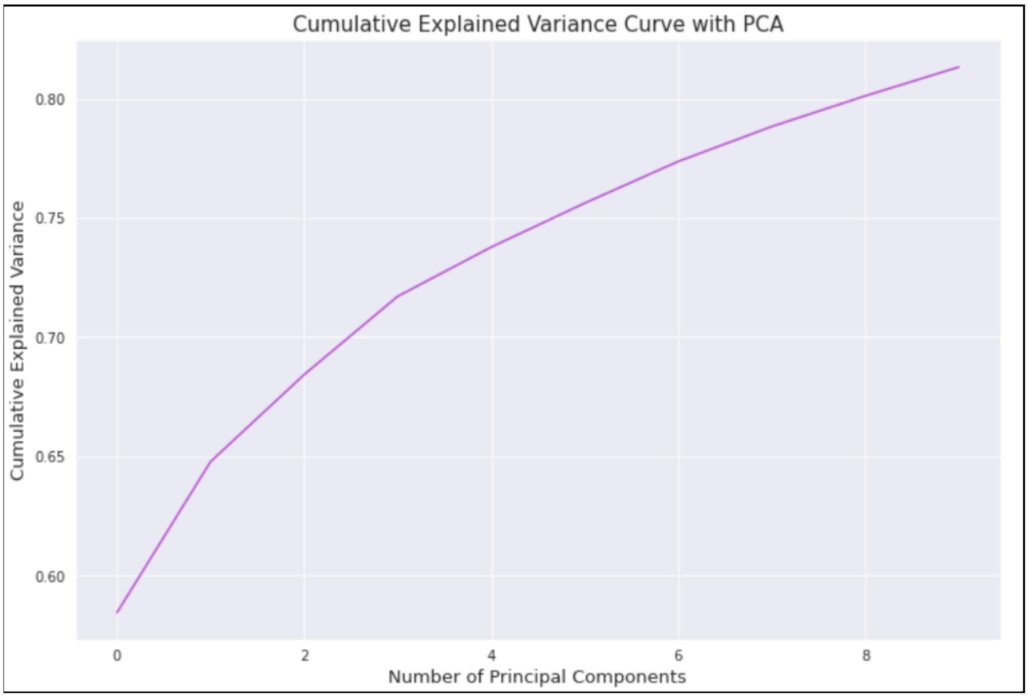
This was done by first splitting the current dataset on an 80:20 ratio for training-to-testing. The PCA was first performed on the training set and then the same transformation was applied on the testing set.

```
PCA(n_components=10)

After PCA, the regions train set is: (501, 10)
After PCA, the regions test set is: (126, 10)
```

One important note is that regional data was scaled to prevent any imbalances in listener population for specific regions (e.g. if London has significantly more listeners than other regions).

```
Total explained variance of the 10 principal components: 0.8132713463062121
```



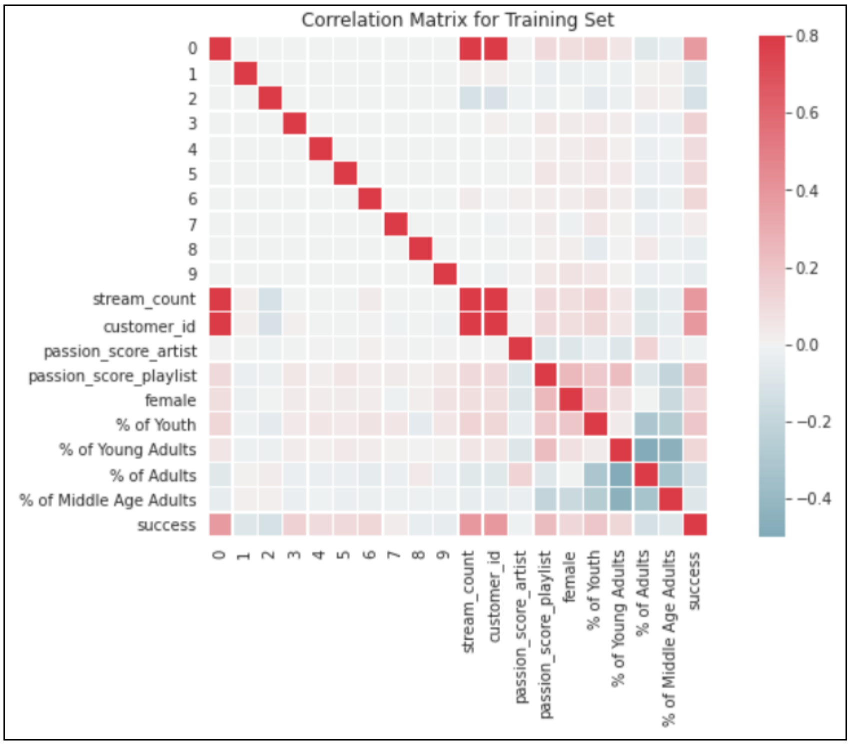
(Figure 16: Plot of cumulative explained variance curve)

The top 10 components have an explained variance ratio of 81.33%. Meaning, 81.33% of the variance in listener regional data is explained by the created 10 principal components.

3.3 Data Transformation

Finally, all features are merged to obtain the dataset with all variables of interest. Any remaining null values are filled, as well as standardising the data so that relative differences between variables are analysed rather than absolute differences.

3.4 Feature Selection



(Figure 17: Correlation Matrix Heatmap)

To avoid multicollinearity, a correlation matrix heatmap was created to uncover any highly correlated variables. Stream count turns out to be highly correlated with unique streamer ID and regional PCA 0. Hence, it was removed.

3.5 Class Balancing

Noticing that there are more failed artists, class balancing was done on the training set so that successful and failed artist classes are balanced to avoid bias toward predicting failures.

```
Before Class Balancing

0     436
1      65
Name: success, dtype: int64
```

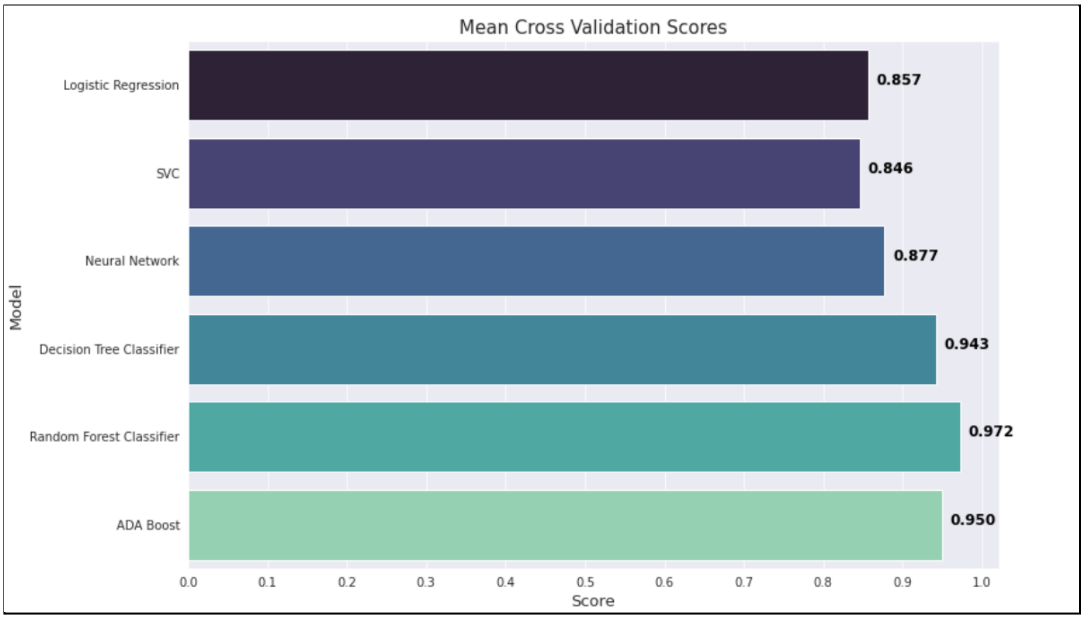
```
After Class Balancing

0     436
1     436
Name: success, dtype: int64
```

4. Model Selection

With the final dataset, we ran some classification models in order to see which model performs the best in terms of accuracy and ROC-AUC scores, as well as what the best model suggests the most important characteristics would be in forecasting an artist's success. The procedure in doing so was as follows:

1. Feed the training set into 6 classification models under their default hyperparameter settings
2. Calculate their cross validation score with 3 CV splits and compute the mean of the 3 scores



(Figure 18: Mean Cross Validation Scores of Chosen Classification Models)

3. Select the models with the 3 highest scoring mean CV-scores and feed them in a hard-voting classifier

```
Cross Validation Mean Score for Hard Voting Classifier is: 0.9701781411699648
```

4. Tune hyperparameters of the said 3 models

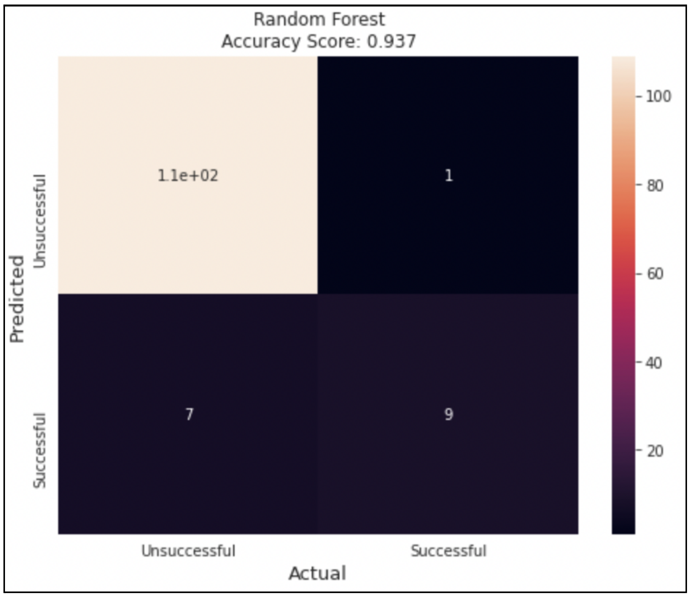
```
Best Score for Tuned Random Forest Classifier is: 0.9774425287356323
Best Hyperparameters for Tuned Random Forest are: {'max_depth': 10, 'n_estimators': 64}
********************************************************************************
Best Score for Tuned Decision Tree Classifier is: 0.9636929641239985
Best Hyperparameters for Tuned Decision Tree are: {'criterion': 'entropy', 'max_depth': 10}
********************************************************************************
Best Score for Tuned ADA Boost Classifier is: 0.9674982584465344
Best Hyperparameters for Tuned ADA Boost are: {'learning_rate': 0.701, 'n_estimators': 512}
```

5. Test these models on testing set
6. Evaluate accuracy and ROC-AUC scores and conduct feature importance analysis

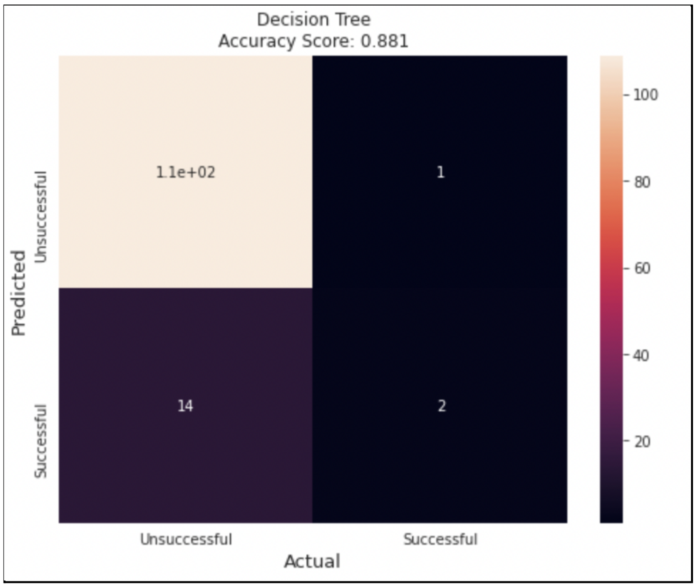
5. Evaluating Algorithms and Present Results

5.1 Confusion Matrix

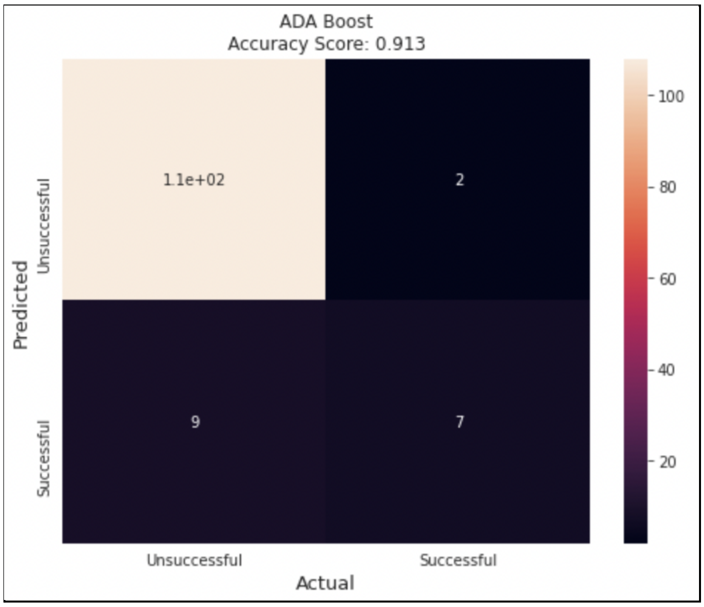
In order to evaluate the performance of the top three models, a confusion matrix accounts for actual and predicted values.



(Figure 19: Confusion Matrix of Random Forest)



(Figure 20: Confusion Matrix of Decision Tree)

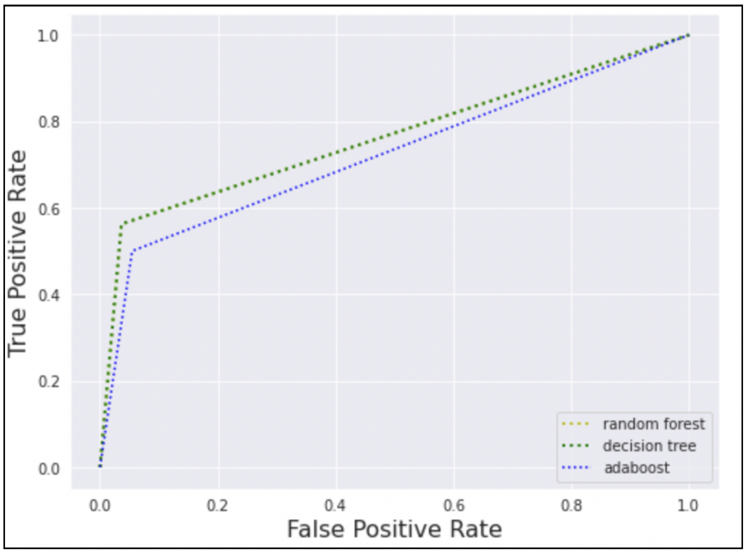


(Figure 21: Confusion Matrix of ADABoost)

The confusion matrices of Random Forest, Decision Tree and ADABoost show that all the three models have less false positives but a relatively high false negative.

5.2 ROC Curve

In order to show the ability of the models to classify subjects correctly across a range of decision thresholds, Receiver Operating Characteristic (ROC) is used. The ROC curves plot the true positive rate vs. false positive rate at every probability threshold. The closer the ROC curves follow the left and top border, the more they are indicative of the model's discriminative powers.



(Figure 22: ROC curves)

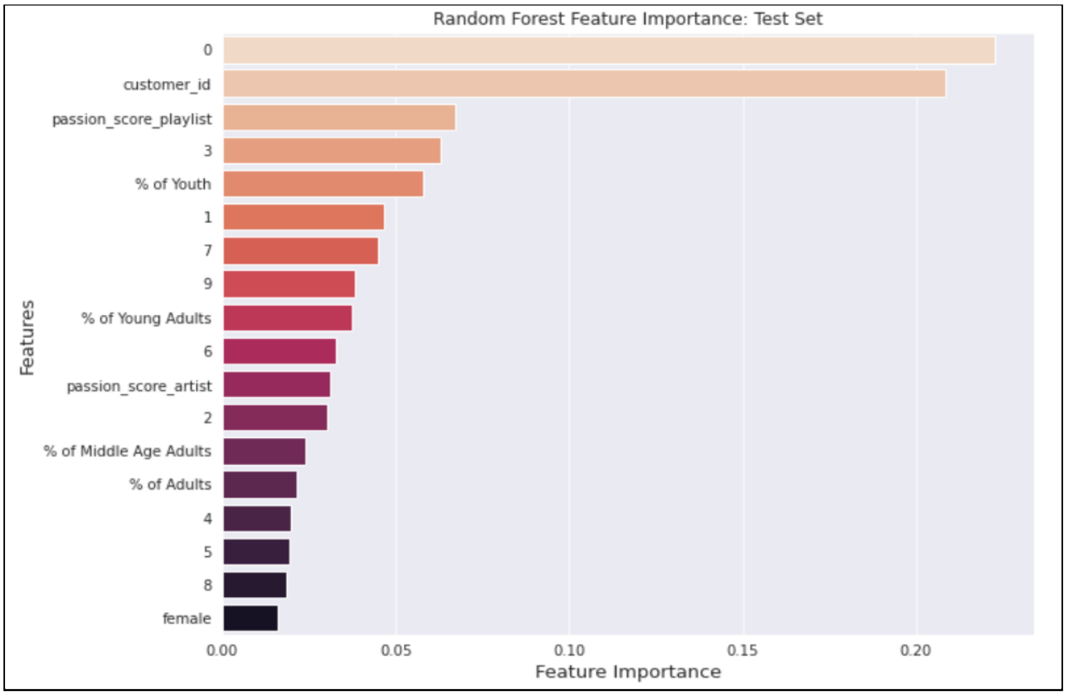
5.3 ROC-AUC Score and Accuracy Score

The following ROC-AUC scores are the summary of the results of ROC. They are the probability that a randomly chosen “success” example has a higher probability of being a success than a randomly chosen “failure” example. The random forest and decision tree get the highest ROC-AUC score among the three models. This means the random forest and decision tree model perform better on recognizing successful artists. The following accuracy scores show how many observations, both positive and negative, are correctly classified. Random forest has the highest accuracy score. Therefore, random forest performs the best.

	Model	Accuracy Score	ROC-AUC Score
0	Random Forest Classifier	0.936508	0.763068
1	Decision Tree Classifier	0.880952	0.763068
2	ADA Boost Classifier	0.912698	0.722727

(Figure 23: ROC-AUC Score and Accuracy Score)

5.4 Feature Importance



(Figure 24: Feature Importance of Variables in Testing Set)

6. Conclusion and Limitations

6.1 Conclusion

The accuracy score results confirm that the **random forest after tuning** performs the best at predicting successful artists, with an overall score of 0.937. Tuned random forest also performs the best via ROC AUC score, indicating that the model is better at assigning higher probabilities to randomly chosen positives (i.e. successful artists) than negatives (i.e. failed artists) on average.

Through this, we find that **unique streamer count** is one of the most important features. This makes sense, as an artist with more outreach to different listeners would imply more robust popularity, as opposed to the same group of listeners repeating streams of the artist's songs.

Also equally important is the prevalence of generalised characteristics (i.e. artist, playlist and/or user-based features), and that **passion score of each playlist** is the second most important factor. This implies that streams per listener is pivotal in forecasting artist success, which further suggests that the more popular the playlist, the more successful the artist would be if their songs landed on the playlist.

Moreover, the **percentage of youth listeners** is among the top 5 most important variables. This hints that in order for an artist to gain popularity on Spotify, he or she should produce songs that are able to be popular among a younger audience, as opposed to older age groups which do not stream on Spotify as much as their younger counterparts.

Furthermore, we discover the importance of **regional data** post-PCA as well under variable importance. This is apparent in the variable importance plot, where 2 PCA regions were among the top 5 most important variables.

6.2 Limitations

Due to the black-box nature of PCA, further exploration into the regional data to study streaming patterns is needed to group successful artists and perhaps better understand the reasons why some regions are more important. To better understand regional variation, we should consider using demographic and census data to resolve the issue.

To minimise omitted variable bias, we could further consider more features into our dataset. For instance, we could use natural language processing to look at how the artist's songs' lyrics correlate with their success. Perhaps the musical quality of their songs, such as rhythmic metrics, tempo and timbre and so on could be incorporated to predict an artist's success. Including these features allows us to understand song-based features associated with each artist, and how they relate to their success, with less concern with omitted variable bias.

Moreover, the issue of external validity may be another issue. If Warner Music's current issue is trying to predict artist success, a popular artist on Spotify might not necessarily mean they are popular in the grand scheme of things. For instance, the rapid rise in popularity of SoundCloud on youth groups should not be ignored, and the results obtained via the same data analysis performed in this project should be considered to further consolidate and enrich Warner's understanding of an artist's success.

Additionally, the analysis and results obtained above were essentially conducted on a relatively small number of artists and a class-biassed sample. Because of this, the implications and statistical significance of the evaluation of our models and associated results are somewhat questionable. More data collected over a longer period of time and a wider variety of artists should be considered to mitigate this issue.

Appendix A: Table of Variables

Variable Name	Description
day	day of stream
log_time	when streamers listen to the song
mobile	mobile or not
track_id	unique track identifier
isrc	unique track identifier
upc	universal product code
artist_name	name of primary artist
track_name	name of track
album_name	name of album
customer_id	unique customer identifier
postal_code	partial zip code
access	spotify account type: free / premium
country_code	2-character country code
gender	customer_id gender: male / female
birth_year	customer_id birth year
filename	data export file name
region_code	3-character region code
referral_code	
partner_name	
financial_product	
user_product_type	user type based on subscription method
offline_timestamp	
stream_length	length of each stream
stream_cached	
stream_source	

stream_source_uri	where on spotify the song was streams, for example, artist page, album page, playlist_id
stream_device	device used to stream
stream_os	OS of device on which song was stream
track_uri	Track URI
track_artists	names of all major performers on track
source	
DataTime	date and time in hr/min/sec format
hour	hour of DataTime
minute	minute of DataTime
week	week of the DataTime
month	month of DataTime
year	year of DataTime
date	date of DataTime
weekday	numerical name of each day. i.e Monday is 0, etc
weekday_name	name of each day
playlist_id	id of each playlist
playlist_name	name of each playlist

Bibliography

[1] Dredge, S., 2015. *Spotify has six years of mymusic data, but does it understand my tastes?*. [online] The Guardian. Available at: <https://www.theguardian.com/technology/2015/jan/06/spotify-music-streaming-taste-profile> [Accessed 17 March 2022].