

Nanyang Technological University

Academic Year 2022/2023 Semester 2

**BC2407: Analytics II: Advanced Predictive Techniques**

**Seminar Group 01: Group 04**

**Project Title : Finding New Talents for Universal Music Group**

Date of Submission : 2 April 2023

| **Name:** | **Matriculation Number:** |
| --- | --- |
| Advait Bharat Deshpande | U2122619C |
| Ng Wee Kiat | U2122504J |
| Foo Shi Jian | U2110071E |
| Neha Bhagat | U2110235D |
| Ho Hui Yu | U2110926J |

### 

**Table of Contents**

[**Executive Summary 2**](#_eexwg1co99iy)

[**1. Introduction 3**](#_wzebtj8uvygs)

[1.1 Background of Case 3](#_moghcqnzphpl)

[1.2 Business Problem 4](#_2ixsxxe5qap6)

[1.3 Our Approach 6](#_afp5qa9rcood)

[1.4 Business Outcome Measures & Desired Targets 7](#_ncejzsmnoij3)

[**2. Secondary Research 8**](#_g5lo1sw2c8c5)

[2.1 Origin of Dataset 8](#_ipvqgvkj3053)

[2.2 Web Scraping 9](#_7igxqmpi17hl)

[2.3 Data Handling 9](#_9zlc057svup5)

[2.3.1 Data Cleaning 9](#_tph8sio4kayq)

[2.3.2 Data Exploration 10](#_el3c4v98ms2l)

[2.4 Data Preparation 13](#_69ezl545ckff)

[2.4.1 Data Scaling 13](#_4uw1kyh9kku0)

[2.4.2 Train-Test Split 13](#_rwobc1gmrphe)

[2.4.3 Feature Selection 13](#_38xqxiabs7co)

[2.4.4 Balancing Class in Training Set 14](#_aeyi03x9m08a)

[**3. Data Modelling 15**](#_pm41o3pvfgx1)

[3.1 Logistic Regression 15](#_7l833odr9o8j)

[3.2 CART 15](#_jqusbytc1lpx)

[3.3 Random Forest 16](#_8vgcshzf9jgg)

[3.4 MARS 16](#_pelnp5pj5ui3)

[3.6 Analytics Performance Measures and Targets 17](#_wlya6nvjnyas)

[3.7 Further Improvements of Chosen Model 18](#_fytkup9n3c0h)

[**4. Proposed Solution 18**](#_wwy26ss3e9cv)

[4.1 BeatScout 18](#_r2kx4dvw754q)

[4.1.1 How it works 18](#_s47mfskr9rbz)

[4.1.2 How UMG can implement this tool 19](#_kfpquoloyg1b)

[4.1.3 Impact on Business Outcomes 19](#_uz7lwje3d05d)

[4.2 How UMG can extend / improve on BeatScout 20](#_xz6grmesqjqd)

[**5. Conclusion 20**](#_selgd8xhi2su)

[**References 21**](#_z3svjt43nof3)

[**Appendices 23**](#_7leb656g53ix)

[Appendix A—Data Dictionary 23](#_f40e1in85o0k)

[Appendix B— Data Exploration Plots 25](#_s116xtj0axx1)

[Appendix C - Data Modelling 26](#_695whpysd82j)

# **Executive Summary**

Although UMG boasts a star-studded roster of artists and a market leader position, there is still a need for UMG to review and improve on their talent acquisition process. The rise of streaming and social media platforms have levelled the playing field between labels and it is increasingly difficult for UMG to identify and sign artists with potential accurately.

As a result, record labels including UMG has been playing the numbers game by signing as many artists as possible to increase their scope of finding the next up-and-coming viral artist. However, this process is not sustainable as it is resource-intensive in terms of evaluating thousands of potential new artists and becoming less cost-effective due to the inflated talent costs while rate of returns are low. UMG’s competitors have begun to adapt their strategies. To protect UMG’s market position and secure future growth, UMG should adopt a measured and analytical approach to confidently sign future stars, be cost-effective and resource-efficient. Hence, our group proposes for the implementation of BeatScout.

BeatScout is an all-in-one software which offers convenience and analysis for A&R employees who struggle to identify and sign the next up-and-coming viral artist in an abundance of data. BeatScout is created in 3 stages: Data Collection, Exploration & Preparation, Data Modelling, and Synthesis & Business Implementation.

In the first stage, we built a robust and relevant dataset through combining Kaggle’s popular TikTok songs dataset from 2019 to 2022 and webscrapped an additional variable ‘Videos\_created’ from TikTok as a mean to evaluate the virality of a track on Tiktok. Next, data cleaning followed by data exploration is carried out to gain a better understanding of our dataset and brief insights into significant variables. Subsequently, data preparation processes including data scaling, train-test split, feature selection and balancing of data are required for better model performance created in data modelling.

In the data modelling stage, we explored models including Logistic Regression, CART, Random Forest, MARS and Neural Network models and selected Random Forest as the best predictive model for BeatScout taking into account metrics such as accuracy, false negative rate, false positive rate, F1 and F1 Gain.

In the final stage, we package our analytics solution into a feasible business implementation. BeatScout will deploy our trained Random Forest model and generate an interactive dashboard. It will provide relevant information extracted from the Spotify API and accurate analysis of the narrowed selection of artists who have the potential to go viral on TikTok. Marketing team can also leverage on BeatScout to improve UMG’s ability to market both new and current artists effectively based on BeatScout’s predictions. This will significantly increase convenience and reduce resource wastage by both A&R and Marketing teams.

BeatScout is a sustainable and progressive solution for UMG as it is continuously retrained with updated musical and non-music data to stay relevant in the ever-changing music industry while improving its accuracy. By leveraging the power of BeatScout, UMG is able to see the increase in profit margins with efficient resource allocation and gain a competitive edge over their competitors by signing potential up-and-coming artists more accurately.

# **1. Introduction**

## **1.1 Background of Case**

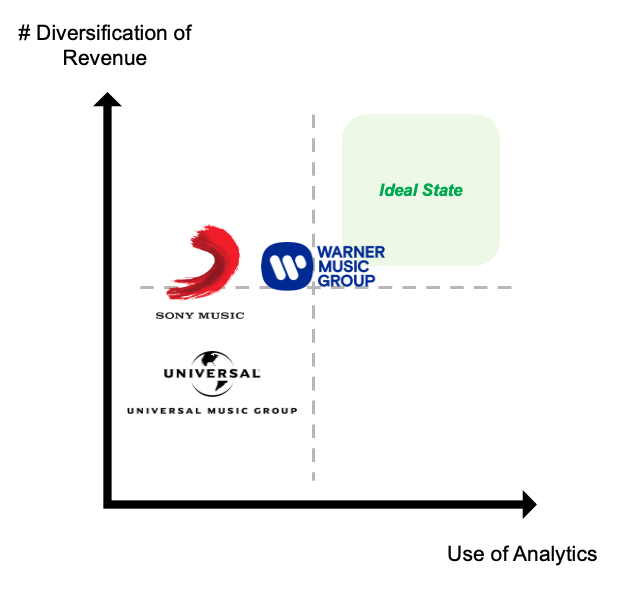
A record label is defined as a company in the music industry that creates and sells music recordings such as records, CDs and other media. The three biggest record labels are Universal Music Group (UMG), Sony Music Group (SMG) and Warner Music Group (WMG). These companies primarily discover and sign talented singers and musicians to exclusive contracts, and then release and distribute their music to generate revenue for both the label and the artists.

Previously, record labels had the unwavering power to decide which artists become superstars, as they had a monopoly over promotions and distribution platforms. For instance, if an executive from a label liked your gig at a bar or saw potential in a demo you submitted to the label, they could fast-track you to success. The competition between labels hinged primarily on their connections in the industry or sheer luck to seek out hidden gems. (Producer Jerry Beltran, 2022).

Today, the landscape is very different, due to the following trends:

1. The rise of social media and streaming services reduces artists’ reliance on the major labels to build a career with independents accounting for 31% of the market (Runcie, 2021)
2. Every label has access to and is using publicly available data to find their next artists on TikTok, SoundCloud, etc (Miller, 2019)

There has been a power shift in favour of the artists, and this has led to the big three labels going on a signing spree, as they now must play a numbers game to find their next hit (Billboard, 2023).



*Figure 1.1: Comparison of strategy between the 3 Major Labels*

UMG’s competitors have already begun to adjust their strategies (Figure 1):

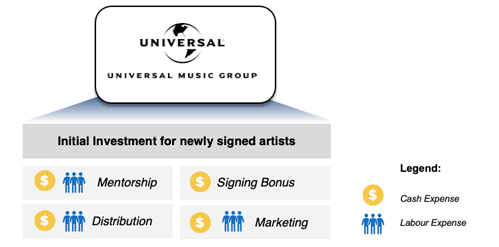
1. WMG is reducing their reliance on superstars and looking for less seasoned artists
2. SMG is focusing on investing in indie music and streaming services

This leaves UMG in a precarious position. Despite being a market leader, it is also the most reliant on superstars than the other two (Runcie, 2021). If its competitors' bets pay off, it risks losing its market position and bargaining power over its current roster of artists.

Ideally, UMG would want to move to the ideal state as depicted in Figure 1.1.

## **1.2 Business Problem**

While Universal Music Group **could simply continue to rapidly sign new artists** in search of the next big hit, industry trends do not support this approach. The number of first-timers who claimed the top 10 positions of Billboard’s Top 100 each year has fallen from 30 to just 12 in 2022 (Billboard, 2023).



*Figure 1.2.1: UMGs initial investment into newly signed artists*

Clearly, the surge in new signings has not translated into a surge of new superstars. Furthermore, signing unsuccessful artists increases opportunity costs for UMG:

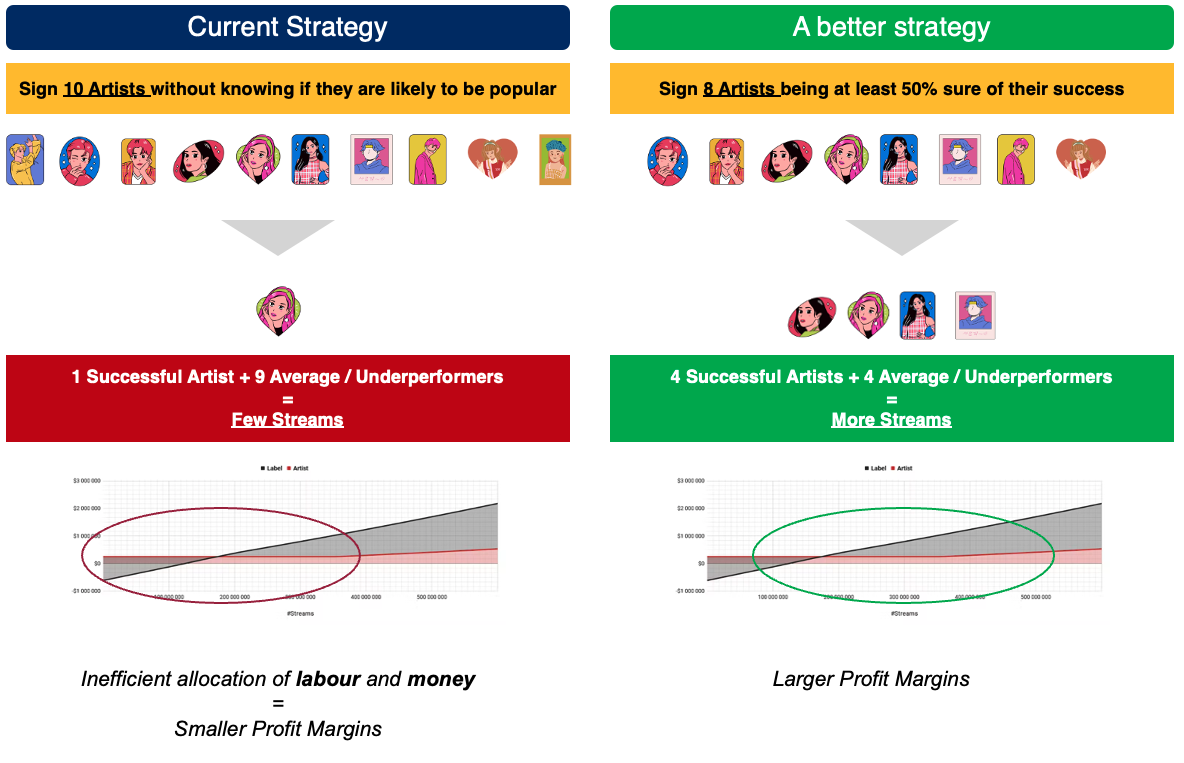
**Tangible Effects**

1. UMG’s spending on new musicians and songwriters have increased in the past two years due to the competitive environment which inflates talent costs (BSIC, 2021)
2. According to Morgan Stanley, A&R costs for introducing a new artist can reach $2 million in each major market, meaning UMG may spend up to $10 million to $15 million globally per artist (Langis, 2022).

**Intangible Effects**

1. The label’s staff growth is unable to keep up with roster growth, leading to difficulties in trying to market all the newly signed artists. This results in high levels of fatigue experienced by the participating employees (Leight, 2023).
2. Without a model to select the right artists, marketing talents is a highly time-intensive process and proves to be a challenge to even the most qualified marketers (Leight, 2023).

**This calls for an alternative strategy, one that will reduce the cost borne by the company through better deployment of manpower and resources, thereby increasing the profits earned through the success of new artists.**

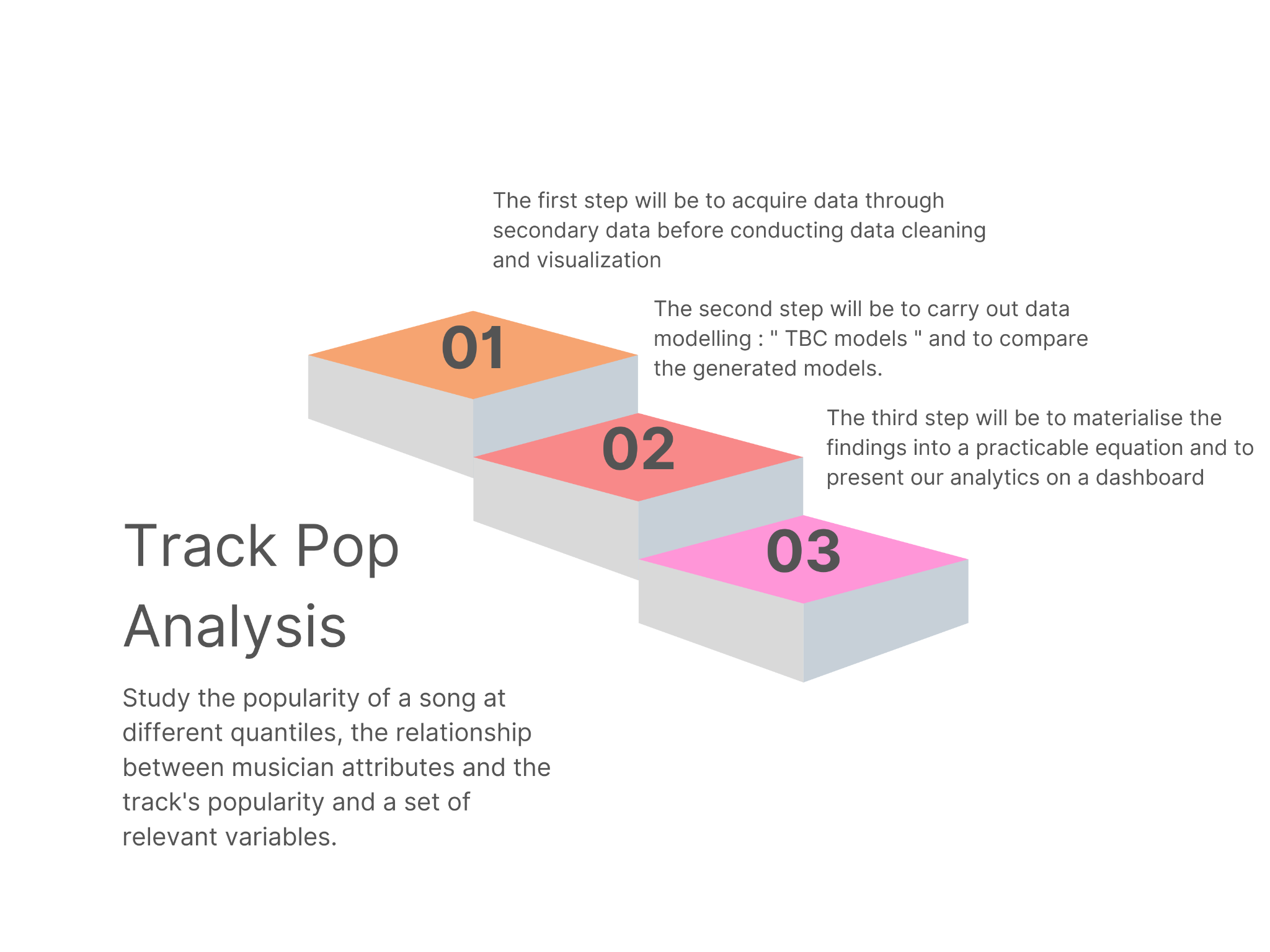


*Figure 1.2.2: Comparison of Strategies by Profit Margins (Pastukhov, 2020)*

To address this, UMG can apply machine learning to create an algorithm that will determine whether a song has the potential to be successful in the music industry. Success will be quantified by the assurance of the track being ranked in the Billboard Hot 100. UMG will then sign on artists who have produced songs with great potential in order to reap the financial benefits of distributing the artists’ work. Through this alternative strategy, UMG will not need to sign on artists excessively but instead, can limit their signings to artists’ whose song demo is predicted to be successful. This will reduce the costs borne by UMG when the label engages with unsuccessful artists and increase the profits earned by the company.

In conclusion, our goal is to exploit this opportunity for UMG to stay competitive with SMG and WMG by devising an alternate strategy **to sign profitable artists** and to, secondarily, **draw meaningful insights** that will boost the traction gained from their music.

## **1.3 Our Approach**



*Figure 1.3: 3 Stage Analysis*

Our team proposes the following 3-step analysis:

| **Stage** | | **Key Processes** | **Objective** |
| --- | --- | --- | --- |
| 1 | Data Collection, Exploration & Preparation | 1. Obtaining relevant dataset 2. Supplement dataset with additional data through web-scraping 3. Data cleaning, scaling and feature selection | 1. Build a robust and relevant dataset 2. Visualise trends to provide early insight into significant variables 3. Prepare data for modelling |
| 2 | Data Modelling | 1. Explore various models:    1. Logistic Regression    2. CART    3. Random Forest    4. MARS    5. Neural Network 2. Assess model performance 3. Optimise and tune chosen model | Maximise model performance for UMG in terms of :   1. F1 Score 2. F1 Gain |
| 3 | Synthesis & Business Implementation | 1. Package Analytics Solution into a feasible business implementation | 1. Make it easier for UMG to assess artist potential 2. Streamline artist recruitment process    1. Saving time    2. Reducing opportunity cost |

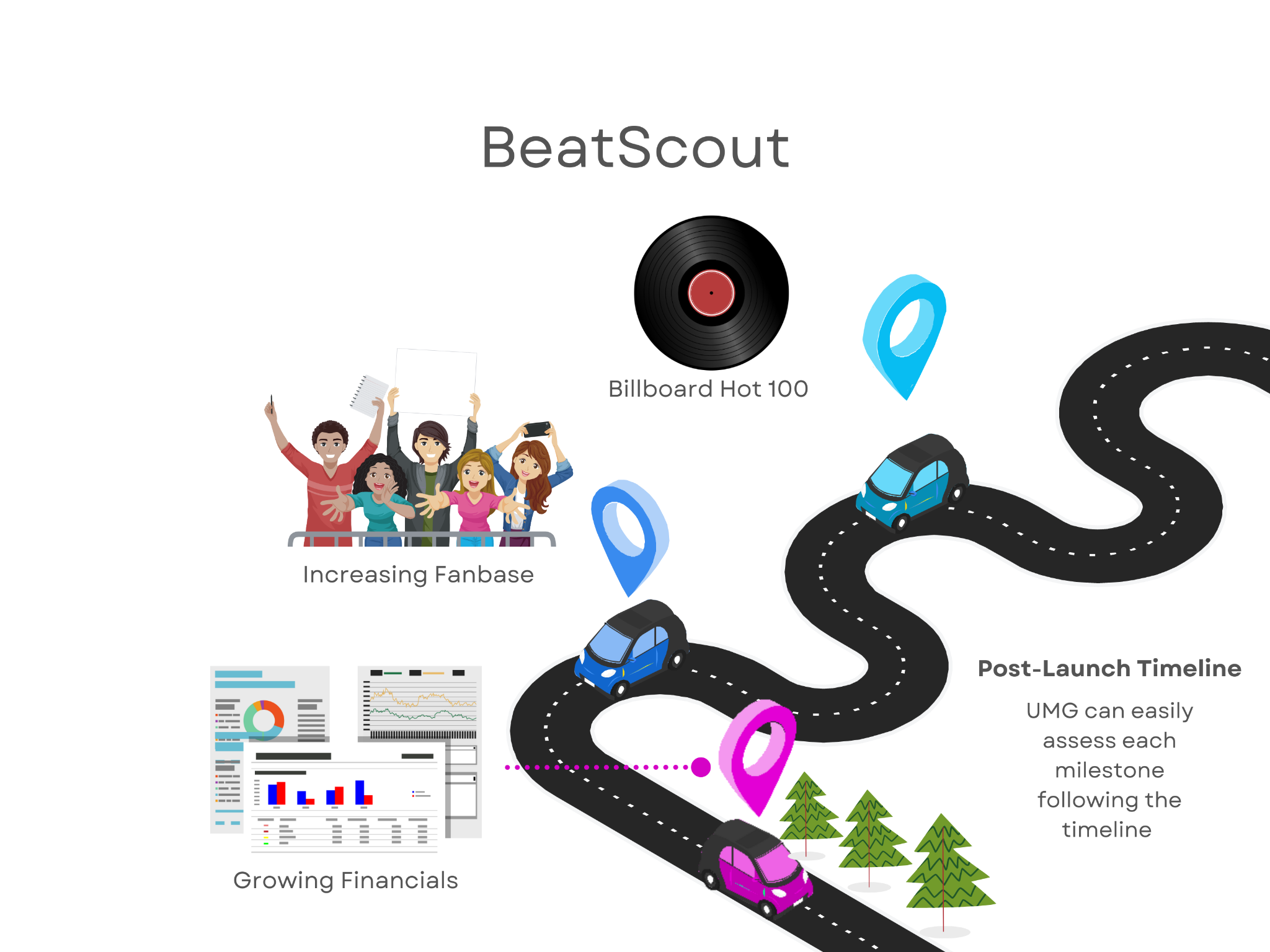
## **1.4 Business Outcome Measures & Desired Targets**

Labels use a varying combination of key performance indicators (KPIs) to quantify the success of their signed artists and, subsequently, the label as a whole.

In the short-run, labels like UMG mainly look at streaming data from services such as Spotify and Apple Music to track an artist’s immediate success upon the release of a song (Medium, 2023). This data is then reflected on charts such as Billboard Hot 100, Spotify’s Top 50 - Global and Apple’s Top 100 - Global, among which the most reputable chart is Billboard’s. To appear on Billboard Hot 100, the song would need to have generated prominent sales, radio airplay and streaming activity before being ranked against other music of comparable calibre (Billboard, 2013).

In the long-run, labels like UMG also take into account the popularity of artists as a means to evaluate their performance and also potentially determine if labels would offer a recontract with the artists. The growth rate of the fanbase is one of the key determinants of artists’ popularity (Musiio, 2021).

In terms of financial performance, UMG utilises three main KPIs in their end-of-year financial reports. The first KPI is Recorded Music Revenue - the sum of subscription and streaming revenue, other digital revenue, physical revenue and licence and other revenue (UMG, 2021). The second KPI is music publishing revenue - revenues benefitted from the continued growth in subscription and streaming (UMG, 2021). The last KPI is merchandising and other revenue, contributed largely by retail revenues (UMG, 2021).



*Figure 1.4: Timeline and Milestones*

To monitor the progress of BeatScout, the various KPIs mentioned will be assessed following the timeline above, the three main milestones are:

1. The Billboard Hot 100 chart ranks the top artists and tracks popularity on a weekly basis (Billboard, 2023). UMG should aim to see a weekly increase in the number of their newly signed artists and tracks on the Billboard Hot 100 chart.
2. UMG can gauge the artists’ popularity by monitoring the growth rate of their fanbase. The label can track the artist’s Spotify’s Monthly Listens and the artist’s followers on social media platforms such as Instagram. UMG should aim to see gradual growth of their artist’s fanbase as the artist gains popularity with viral songs (Musiio, 2021).
3. UMG’s financial report is generated every 6 months to analyse their business performance which is an appropriate period for UMG to review the performance and improvements brought by BeatScout. UMG should aim to see a higher rate of return.

# **2. Secondary Research**

In our secondary research, we searched for multiple datasets comprising factors and attributes of popular tracks before deciding on the one to be used in order to build our analytical model.

## **2.1 Origin of Dataset**

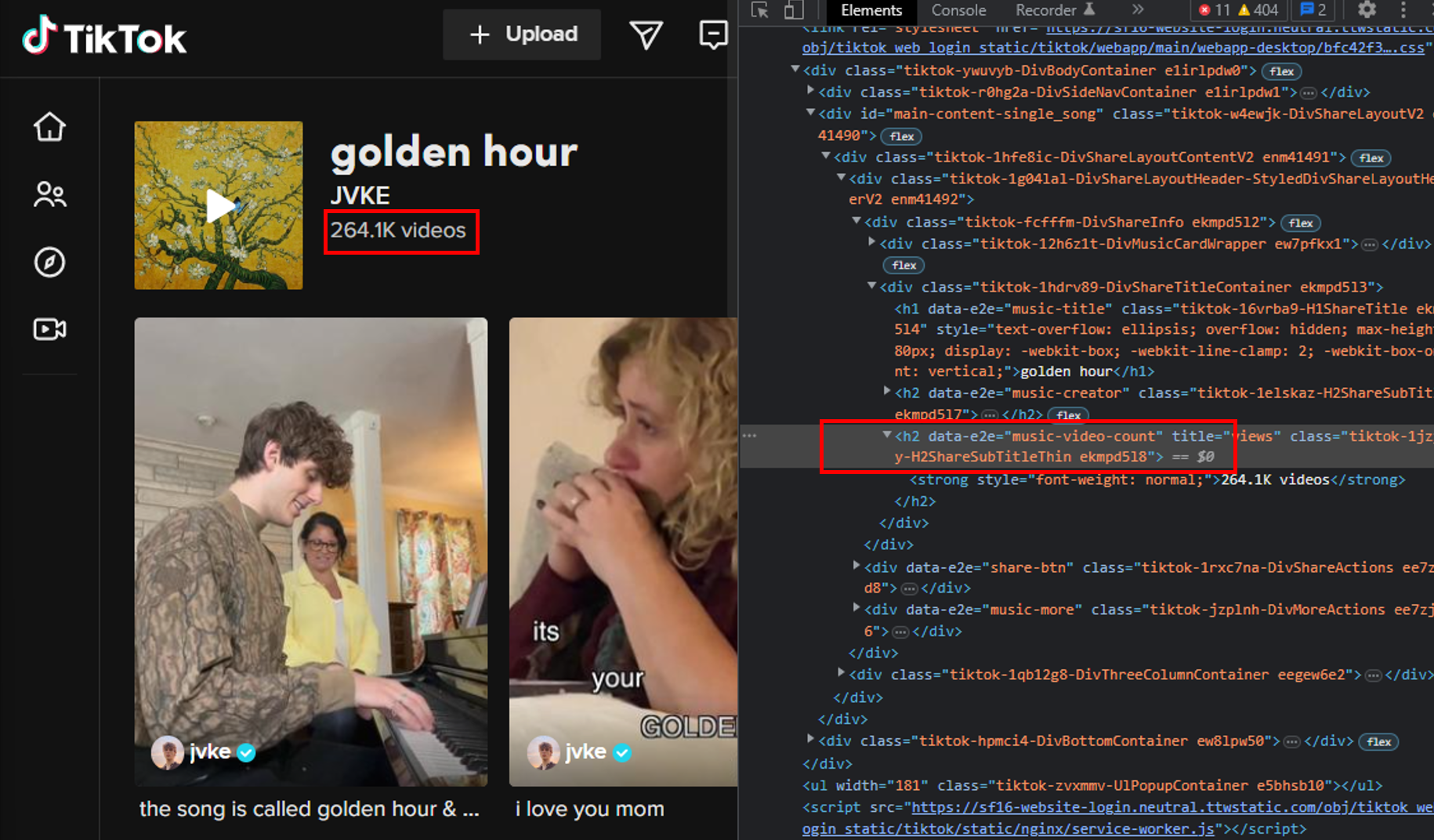
The dataset we are utilising is named “Tiktok \_songs\_2022.csv ”. This file, collated by an anonymous netizen, is a collection of all the popular songs in Tiktok from 2019 to 2022 via Spotify API and web scraping using BeautifulSSoup.

We have strategically chosen a dataset that represents analytical insights from TikTok specifically. According to TikTok’s end-of-year analysis, 13 out of the 14 Billboard Hot 100 No.1 songs in 2022 “were driven by significant viral trends on TikTok '' (Music Business Worldwide, 2022). Similarly in the UK, 10 of the 12 No.1s on the UK Official Singles Chart were driven by virality on TikTok (Music Business Worldwide, 2022). These statistics broadly prove the direct positive relationship between virality on TikTok and profitable success in the music industry.  Hence, the algorithm to be created can constitute variables that contribute to success on TikTok as these variables do, subsequently, contribute to success on Billboard. Furthermore, UMG can utilise the algorithm to identify significant variables that lead to musical success. The company can then use these insights to better market their artists.

## 

## **2.2 Web Scraping**

To expand upon our dataset, we webscraped an additional variable “Videos\_created”. This variable is necessary for an evaluation of the virality of a track on Tiktok.



*Figure 2.2.1: Tiktok WebElements*

Web Scraping was done in Python using the Selenium package and Chromedriver, with Pandas used to write the data into a spreadsheet. After the HTML content is received from the page request, its contents are parsed and the number of videos created under the Tik Tok sound is extracted. This process is repeated for all songs in our dataset.

## **2.3 Data Handling**

### ***2.3.1 Data Cleaning***

Our raw dataset constitutes 968 rows and 24 different columns - “Year\_of release”, “Videos\_created”, “track\_name”, “artist\_name”, “artist\_pop”, “album”, “track\_pop”, “danceability”, “energy”, “loudness”, “mode”, “key, “speechiness”, “acousticness”, “instrumentalness”, “liveness”, “valence”, “tempo”, “time\_signature”, “duration\_ms”, “time\_signature.2”, “duration\_ms.2”, “time\_signature.1” and “duration\_ms.1”.

This dataset is void of any NA values but has 1 duplicate row that was subsequently omitted, resulting in 967 rows.

The webscrapped column, “Videos\_created” is first pre-processed to contain only the extracted numeric values. A for-loop is created such that the words “video” and “videos” were dropped from all the rows. Next, rows which had the character ‘M’, representing millions, were dropped and multiplied by 1,00,000. Similarly, rows which had the character “K”, representing thousands were dropped and multiplied by 1,000. The class of the resulting “Videos\_created” column switched from character to numeric data type. Additionally, this process generated 3 NA values that were subsequently omitted.

Secondly, we created a new binary column named “virality” that will act as our target Y variable which is based on a threshold that will determine whether the song will be viral or not. This threshold is equivalent to the mean value of the “Videos\_created column”, 849396.219. If the “Videos\_created” value in the row is greater than or equal to the threshold, it will correspond to a “Yes” in the “virality” column. Else, a “No” will be inputted under the “virality” column. Essentially, we are reasoning that tracks will go viral if the number of corresponding videos created is above the mean number.

Thirdly, we factored the categorical X variables - “key”, “mode” and “time\_signature”. We then dropped all irrelevant columns with respect to our predictive outcome - “Year\_of\_release”, “track\_name”, “artist\_name”, “album”, “track\_pop”, “artist\_pop”, “time\_signature.1”, “time\_signature.2”, “duration\_ms.1” and duration\_ms.2”. “track\_pop” and “artist\_pop” are among the variables we have chosen to drop as the model is used to predict the virality of songs made by musicians that have yet to gain popularity in our context. Hence, these variables are not relevant to our predictive outcome. The “Videos\_created” column is also dropped as inclusion of the column as a predictor variable will result in multicollinearity and, subsequently, contribute to overfitting and instability of future models created.

Our final dataset has 964 rows and 14 columns.

### ***2.3.2 Data Exploration***

**Treating Outliers**

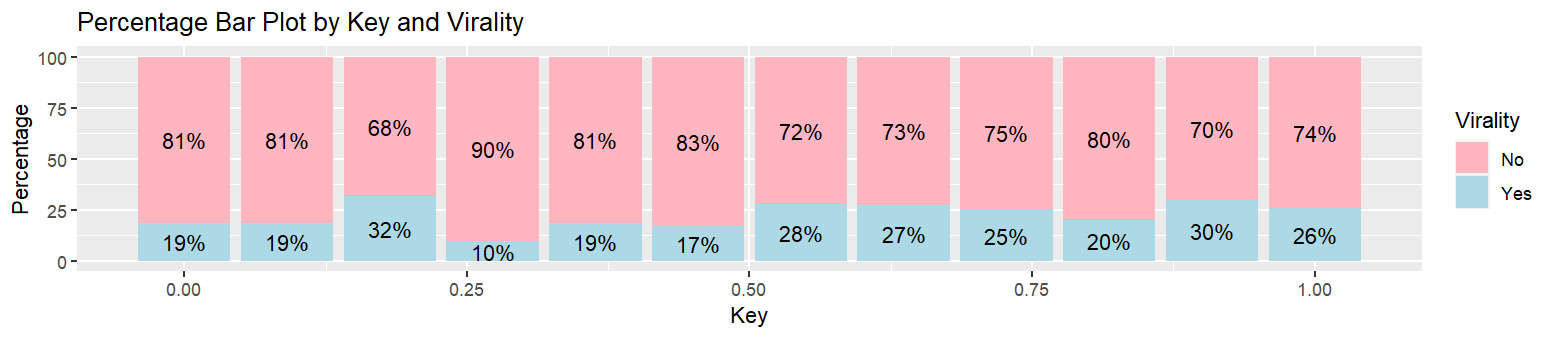
Outliers were replaced with the median value of the corresponding “danceability”, “energy”, “loudness”, “acousticness”, “tempo” and “duration\_ms” as the outliers in these variables do not contribute any significant value to the predictive outcome and are replaced with the median value to prevent data leakage. For the remaining variables - “speechiness”, “instrumentalness” and “liveness”, the outliers were not cleaned as these values hold significant value. Refer to Appendix B for boxplots of the variables.

1. The variable “duration\_ms” represents the duration of the track in milliseconds. We have decided to remove the outliers in this variable for two reasons. Firstly, it is not logical to have tracks that are 0 milliseconds long. Additionally, the outliers range from 300,000 to 400,000 milliseconds (5 to 6 minutes) which is an unusually long duration for songs.
2. Outliers for the variable “liveness” have been kept as the range of outliers starts from 0.50 and ends close to 1.0 which represent tracks with a likelihood of being performed live. Values closer to 0 represent tracks that are not performed live while values between 0.8 and 1.0 represent a strong likelihood that the track is live.

**Data Visualisation**

In order to gain a better understanding of the dataset and to explore the underlying relationships, we examined each variable in the dataset deeply through the use of graphical representations.

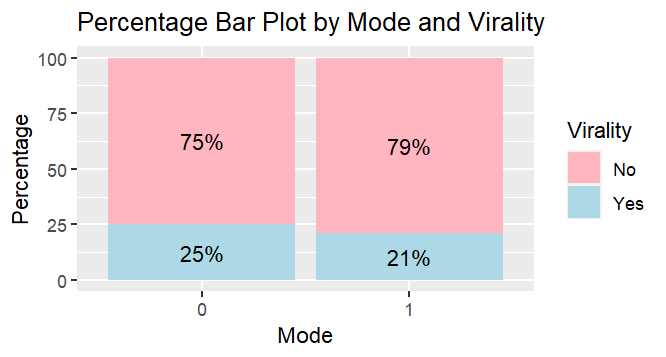
Key and Virality



*Fig 2.3.2.1 : Percentage Bar Plot on Key and Virality*

Key represents the key the track is in where integers map to pitches using standard Pitch Class notation. For instance, 0 corresponds to C, 1 corresponds to C♯/D♭, 2 corresponds to D, and so on. From the plots, it can be observed that values 2, 10, 6, and 7, corresponding to D, F♯, B♭ and G respectively, are the top 4 keys present in viral songs. According to Captain Chords (2023), the top 4 keys are C Major, G Major, A Minor and E Minor. Both sources identify G Major as a definite pitch in popular songs.

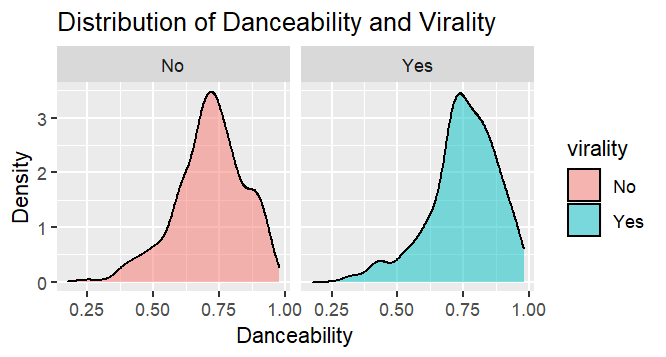
Mode and Virality



*Fig 2.3.2.2 : Percentage Bar Plot on Mode and Virality*

Mode indicates the modality (major or minor) of a track, major is represented by 1 and minor is 0. From the plots, it can be observed that tracks with a minor modality have a higher proportion, 25%, of viral songs as compared to tracks with a major modality, 21%. This is in contrast to research conducted that states that only 36% of viral TikTok songs are in minor modality, leaving the other 64% to be in major modality (Medium 2020).

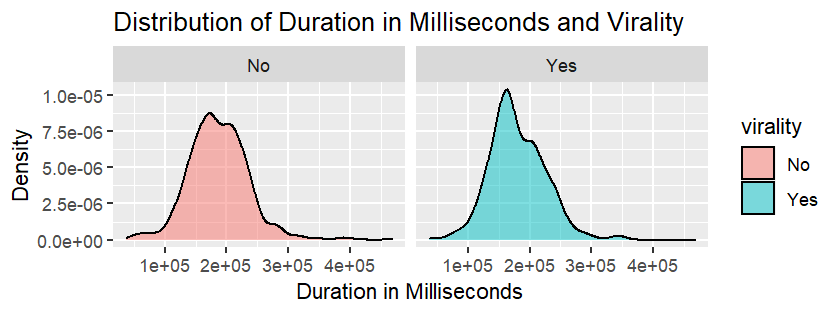
Danceability and Virality



*Fig 2.3.2.3 : Density Plot on Danceability and Virality*

Danceability describes how suitable a track is for dancing based on a combination of musical elements. A value of 0.0 is least danceable and 1.0 is most danceable. From the plots above, it can be observed that the distribution of danceability is relatively more right-skewed for viral songs than it is for non-viral songs. Hence, it can be inferred that the majority of the viral songs have a high danceability index, ranging from 0.7 to 0.9. This is supported by a study that proved that all the 10 top songs tested on the generated model displayed high levels of danceability of over 60% against a scale of 0 to 100% (Medium, 2018).

Duration in Milliseconds and Virality



*Fig 2.3.2.4 : Density Plot on Duration in Milliseconds and Virality*

It can be observed that the plot from viral songs has a higher peak and a steeper downslope as compared to the plot from non-viral songs. This shows that viral tracks have, predominantly, a song duration of 150000 to 200000 milliseconds (2.50 to 3.33 minutes). This is supported by an analysis conducted in 2020 that finds the average duration of popular songs to be 3.28 minutes (Statistica, 2020).

## **2.4 Data Preparation**

### ***2.4.1 Data Scaling***

Our final dataset has continuous variables with different scales, units and ranges. Specifically, values under “loudness” are in decibels, ranging from -23.928 to -1.609, values under “tempo” are in beats per minute, ranging from 62.62 to 210.86 and “duration\_ms” are in milliseconds, ranging from 37632 to 467587. The remaining variables are measured on a scale of 0.0 to 1.0. Hence, scaling is necessary to bring all the features to the same scale to improve model accuracy. We performed MinMax scaling to rescale the data such that all the features are normalised and put on the same scale, between 0.0 and 1.0.

### ***2.4.2 Train-Test Split***

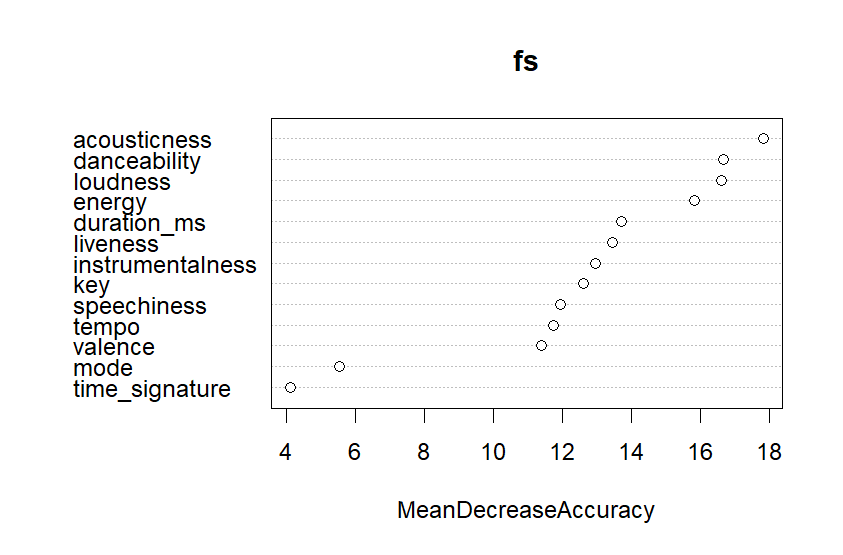
To evaluate the accuracy of the models, we carried out a 70:30 train-test split. To ensure consistency in the random split we have set seed to 2. This allowed us to train the models with 675 observations before testing them to determine the predictive accuracy of models.

### ***2.4.3 Feature Selection***

Feature Selection is performed in two phases - Random Forest and Logistic Regression.

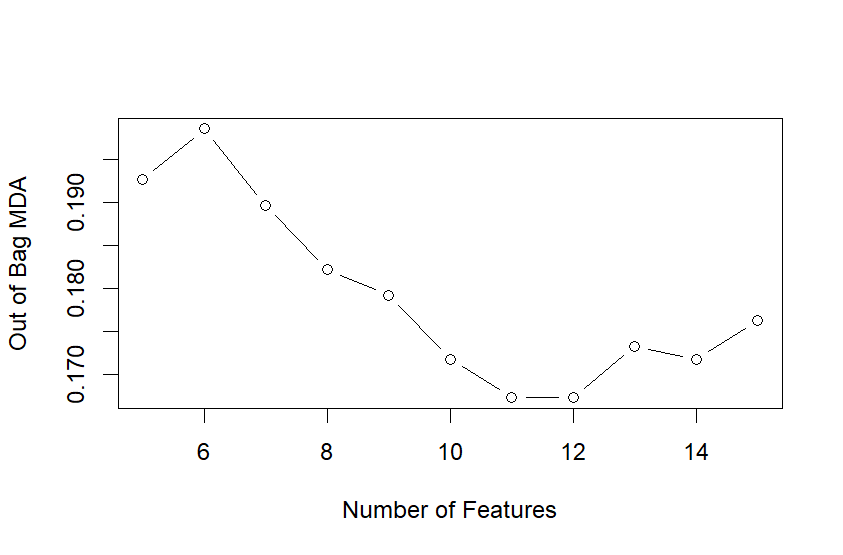
**Random Forest**

We trained a Random Forest model using all 14 variables in the training set and calculated the importance of each variable using Mean Decrease Accuracy (varImpPlot). The graph below shows the ranks of the variables by their mean decrease in accuracy and Gini Impurity which indicates the variable’s importance.



*Fig 2.4.3.1 : Random Forest Variable Importance*

Next, we created subsets of the training set on which we trained a Random Forest model and plotted the OOB error rate for each subset of features to determine the optimal number of features for the model.



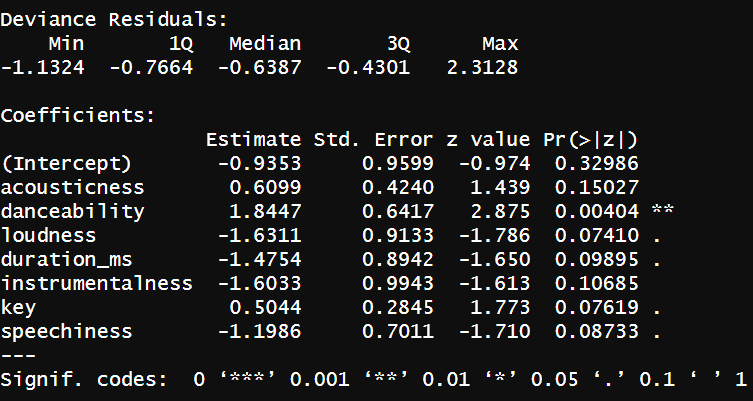
*Fig 2.4.3.2 : OOB Error Rate*

The plot shows the optimal number of features to be 11. The training set is then updated to include only the top 11 most important features, F1.

*F1 ={ acousticness, danceability, loudness, energy, duration\_ms, liveness, instrumentalness, key, speechiness, tempo, valence}*

**Logistic Regression**

We conducted a backward stepwise selection method to select the best set of features using a logistic regression model. The training set utilised is that from the Random Forest analysis performed before.



*Fig 2.4.3.3 : Features from Logistic Regression*

Running a Logistic Regression has further minimised the set of features selected by excluding those with high p-values i.e. insignificant variables. The final training and testing set are both updated to only the final set of features F2 .

*F2 ={ acousticness, danceability, loudness, duration\_ms, instrumentalness, key, speechiness}*

### ***2.4.4 Balancing Class in Training Set***

The dataset suffers from a class imbalance problem, 155 instances in the minority class (Yes) and 520 instances in the majority class (No). The minority class is upsampled to match the majority dataset count to resolve this. The final dataset contains 1040 instances. This helps prevent the models from blindly predicting the majority class.

# **3. Data Modelling**

This section covers the models created to predict the virality of tracks. Logistic Regression, CART, Random Forest, MARS and Neural Network models are run on our final training and testing set to accurately predict the chance of a song going viral based on the track’s musical attributes.

The following performance metrics will be used to evaluate the robustness of the 5 models : Accuracy, False Positive Rate (FPR), False Negative Rate (FNR), Precision and Recall, F1 Score and F1 Gain. The F1 score is a metric that combines Precision and Recall using their harmonic means such that a higher F1 score denotes a better quality classifier. F1 Gain is the improvement in F1 Score achieved by a model as compared to a baseline model. These two measures have been chosen in our study as they are popular metrics for unbalanced classifications, akin to our testing dataset (MachineLearningMastery, 2021).

To accommodate for the consideration of the FPR (sensitivity) and FNR (specificity) of our models, we relied on the optimal probability thresholds determined by the models’ ROC curve for use in the calculation of the models’ performance metrics. This is done through identifying the point on the curve that maximises the area under the curve (AUC).

## **3.1 Logistic Regression**

A logistic regression model is created through the use of stepwise backward regression. The eventual function of the logistic regression model is as follows:

**virality ~ acousticness + danceability + loudness + duration\_ms + instrumentalness + key + speechiness**

The final logistic regression model can be seen in Appendix C. The optimal threshold determined by the model’s ROC curve is -0.2971487. A negative threshold is indicative of a model’s predictive probability being less extreme such that a lower threshold value will maximise both sensitivity and specificity. This threshold is used for subsequent performance evaluation. The model’s accuracy, FPR and FNR are 0.439, 0.691 and 0.121 respectively.

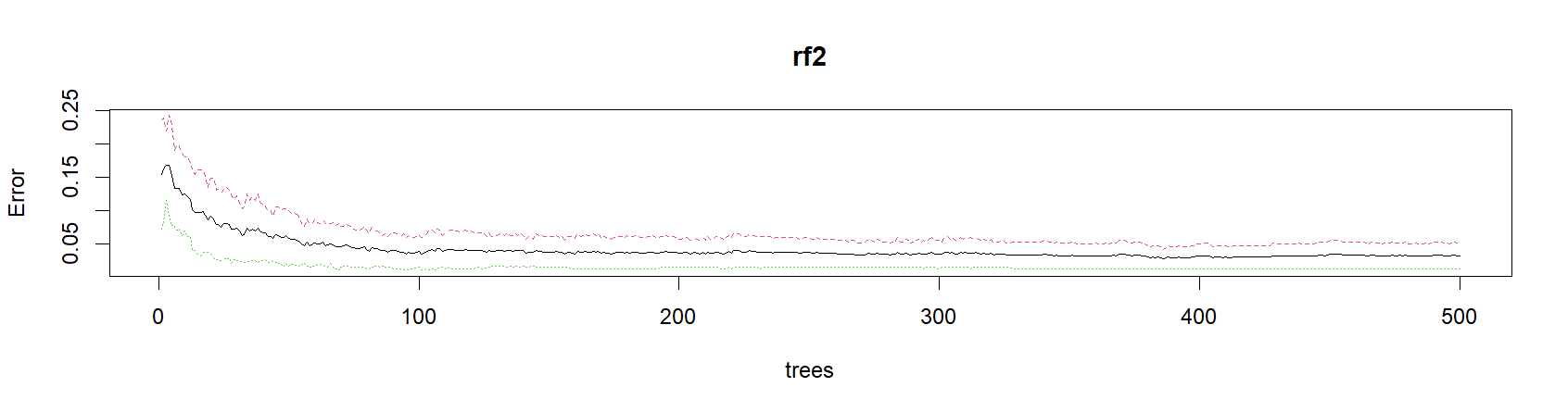
## **3.2 CART**

A CART model is grown with parameters such as a minimum split of 2 observations in a node and a complexity parameter (CP) of 0. Based on the 1 Standard Error (SE) rule, the optimal CP value is computed, 0.001570186, and used to prune the original tree. The final CP table and pruned tree can be found in the Appendix C.

The optimal threshold determined by the model’s ROC curve is 0.9505882 and is used for subsequent performance evaluation. The model’s accuracy, FPR and FNR are 0.740, 0.184 and 0.515 respectively.

## **3.3** **Random Forest**

Using the ‘randomForest’ package in R, a random forest classification model was trained and tuned to derive the following optimal set of parameters based on the model’s Out-of-Bag Error : Number of Trees = 500 and RSF size = 1 (Appendix C). A plot of the model revealed that MSE stabilised before 500 trees, confirming the validity of these parameters.

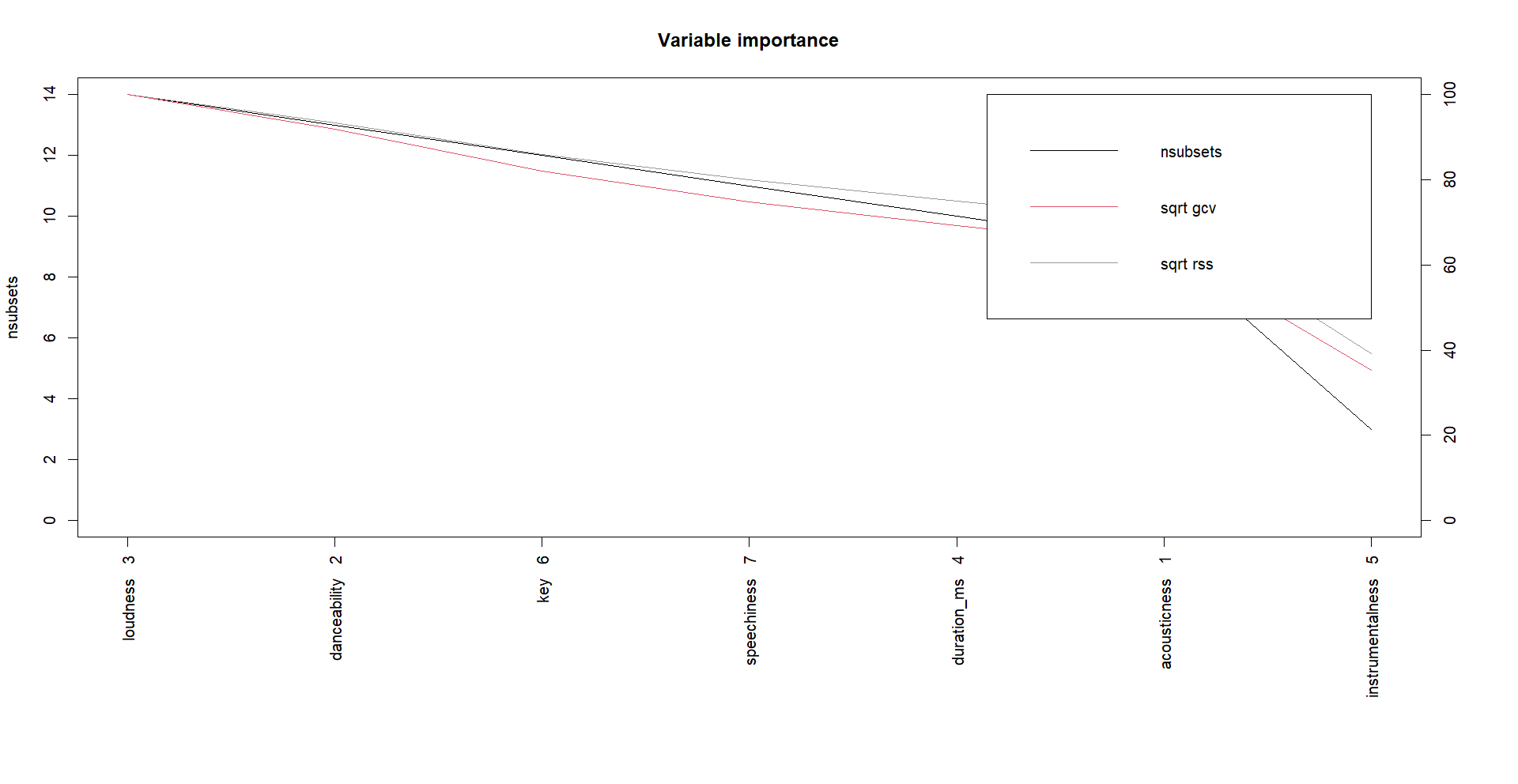


*Fig 3.3 : OOB Error Plot for Overall, ‘No’ Class and ‘Yes’ Class*

The optimal threshold determined by the model’s ROC curve is 0.402 and is used for subsequent performance evaluation. The model’s accuracy, FPR and FNR are 0.796, 0.112, and 0.515 respectively.

## **3.4 MARS**

Using the “earth” package in R, a MARS model was trained and tuned to parameters such as a degree of one for all linear terms. The final MARS model can be viewed in (Appendix C). Using the “evimp” function, the variable importance of each predictor variable in the model was also calculated and plotted.

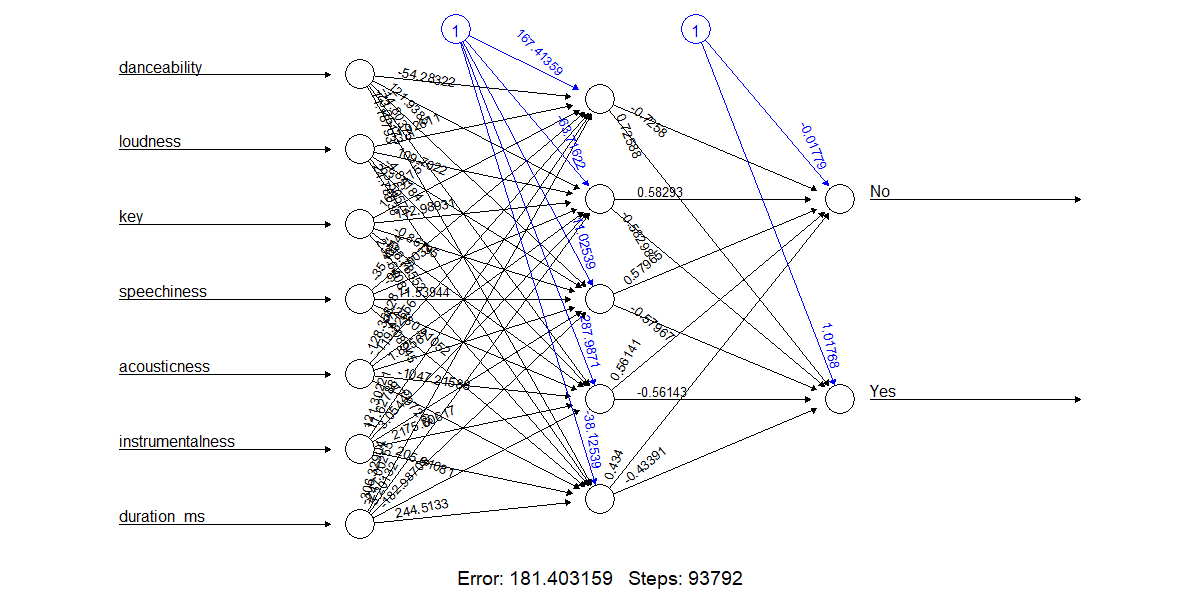


*Fig 3.4 : MARS Variable Importance Plot*

The optimal threshold determined by the model’s ROC curve is 0.4173628 and is used for subsequent performance evaluation. The model’s accuracy, FPR and FNR are 0.481, 0.574, and 0.333 respectively.

**3.5 Neural Network**

Using the NeuralNet library in R, a neural network was made with the following parameters: one hidden layer with 5 nodes and a maximum number of iterations (Stepmax) set to 1 million.



*Fig 3.5: Neural Network structure for Neural Network Model*

The predicted probabilities output by the neural Network Model are then converted into class labels based on the highest probability. A vector of predicted class labels is hence obtained, allowing us to evaluate the performance of the neural network model on our test set.

The optimal threshold determined by the model’s ROC curve is 0.5480015 and is used for subsequent performance evaluation. The model’s accuracy, FPR and FNR are 0.595, 0.381, and 0.485 respectively.

## **3.6 Analytics Performance Measures and Targets**

The performance measures of all 5 models are presented in the table below. After comparing all the metrics, we have chosen the Random Forest (RF) as our best predictive model.

We notice that despite the model having the best performance scores, the FNR is still at an undesirable level even after using the model’s optimal threshold. However, in our context, we prioritise capturing more true positives i.e. having a low FPR. This is because having a high FPR will lead to investing resources in promoting a track that does not have commercial potential, resulting in financial losses. This is a more severe consequence to UMG as compared to the consequence of a high FNR - missing out on the signing of an artist with a high profitability potential. Hence, it is appropriate for us to compromise on the FNR to ensure low FPR and, subsequently, decide on the RF model as the best predictive model.

| **Classification Model** | **Accuracy** | **FPR** | **FNR** | **F1 Score** | **F1 Gain** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 43.9% | 69.1% | 12.1% | 0.417 | 0.587 |
| CART | 74.0% | 18.4% | 51.5% | 0.460 | 0.653 |
| Random Forest | 79.6% | 11.2% | 51.5% | 0.520 | 0.727 |
| MARS | 48.1% | 57.4% | 33.3% | 0.370 | 0.496 |
| Neural Network | 59.5% | 38.1% | 48.5% | 0.368 | 0.491 |

*Fig 3.6 : Table Comparing Models with Metrics*

## **3.7 Further Improvements of Chosen Model**

To improve on the Random Forest model, we could decrease the probability threshold used in calculating the performance metrics. Decreasing the threshold from 0.402 to 0.30 will decrease the FNR of the model from 51.5% to 34.8%, but at the expense of the model’s accuracy, illustrated by the decrease in the model’s accuracy from to 79.6% to 66.1%. Additionally, the F1 and F1 Gain scores also face a decrease from 0.520 to 0.467 and 0.727 to 0.663 respectively.

| **Classification Model** | **Accuracy** | **FPR** | **FNR** | **F1** | **F1 Gain** |
| --- | --- | --- | --- | --- | --- |
| Random Forest | 79.6% | 11.2% | 51.5% | 0.520 | 0.727 |
| Random Forest Tuned | 66.1% | 33.6% | 34.8% | 0.467 | 0.663 |

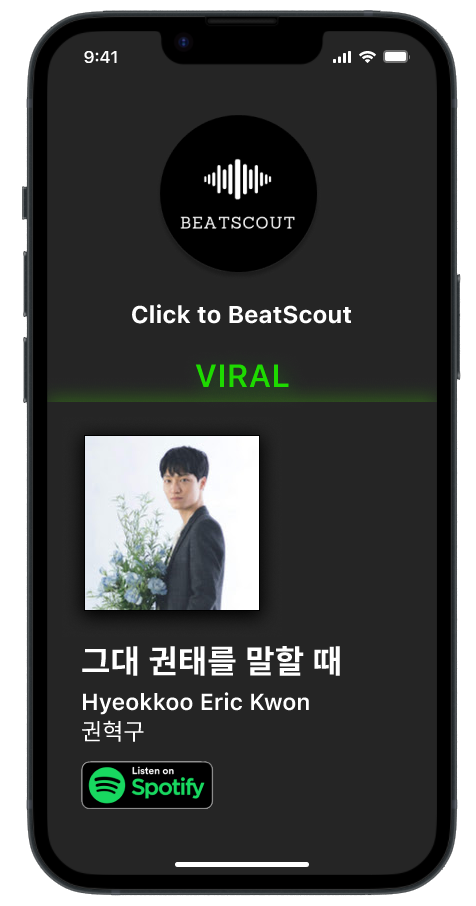
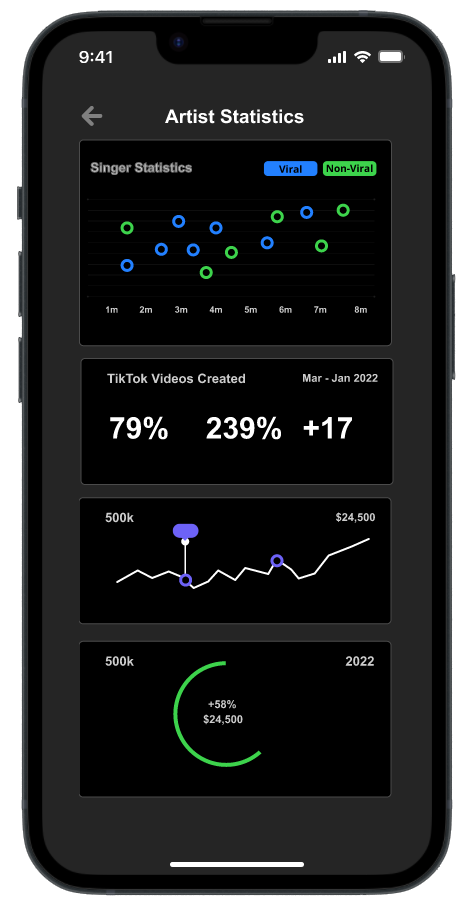
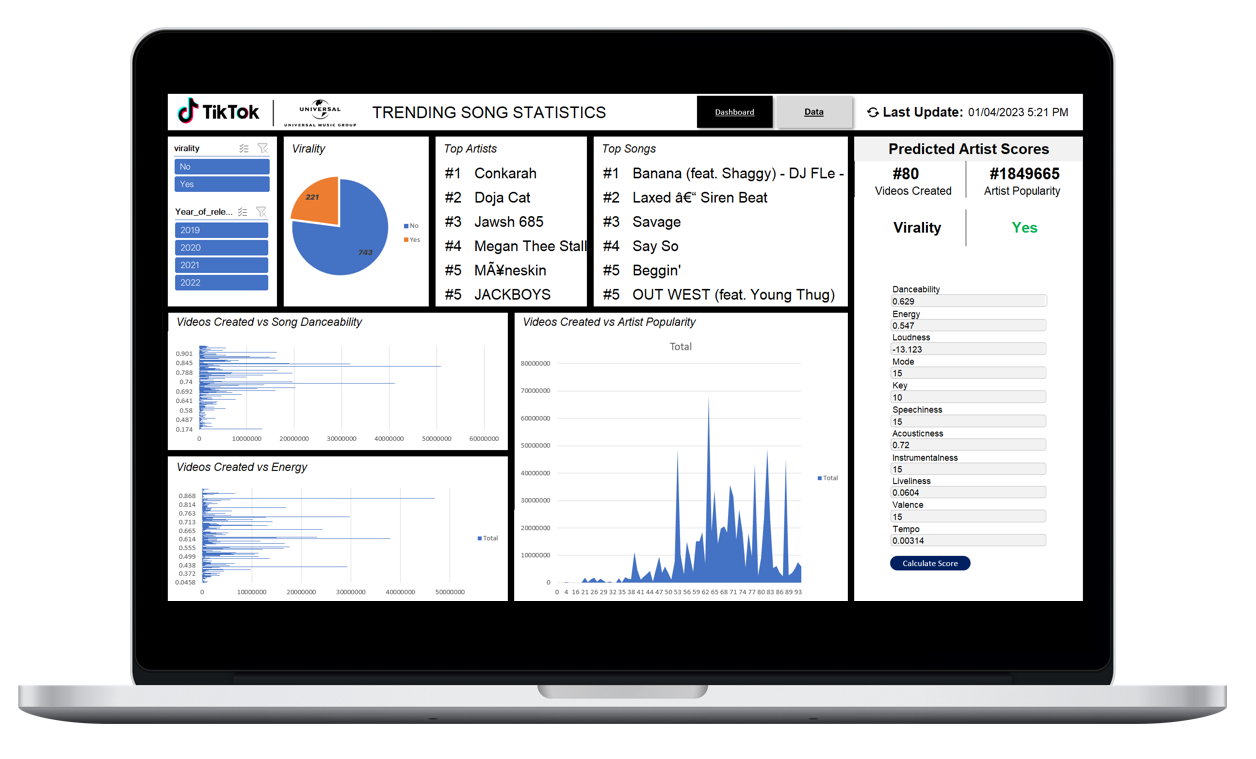
*Fig 3.7 : Table Comparing Random Forest Models with Metrics*

# **4. Proposed Solution**

## **4.1 BeatScout**

## ***4.1.1 How it works***

BeatScout is an all-in-one talent assessment software. It is targeted at Artist & Repertoire (A&R) personnel, who are often in-charge of finding new artists for a record label to sign. As illustrated earlier, A&R personnel today have the monumental challenge of having to sieve through a mountain of information from various social media platforms, streaming services, demo submissions, etc. to find an artist with the potential to shine. BeatScout eliminates this tedious process, and gives them a narrowed selection of artists who have the potential to go viral on TikTok.



*Fig 4.0 : BeatScout Dashboard Mockups*

The BeatScout dashboard deploys our Random Forest Model trained in earlier sections, and connects to the Spotify API. A&R personnel can simply pull up a potential artist's spotify profile, select songs, and get an evaluation of whether the song has the potential to go viral on TikTok.

We understand that different labels may have different standards for what they consider a ‘viral’ song. Hence, we will also allow users to retrain the BeatScout algorithm by adjusting the threshold for virality (Section 2.3.1).

### ***4.1.2 How UMG can implement this tool***

We foresee that UMG can best use BeatScout in the following ways:

1. **Discovery:** A&R Teams who are actively searching for artists may use BeatScout to quickly assess hundreds or thousands of artists at once.
2. **Auditioning:** UMG is the biggest label, receiving millions of demos, A&R teams can streamline their judging process with BeatScout and save crucial man-hours.
3. **Marketing:** For artists currently on their roster, UMG can see whether the music they produce has the potential to go viral. Based on BeatScouts’ prediction, the marketing team can decide how much money to invest into promoting each song.

### ***4.1.3 Impact on Business Outcomes***

BeatScout aims to achieve several positive business outcomes:

1. **More Efficient Talent Acquisition**: UMG can more accurately predict and identify up-and-coming artists. This will help UMG reduce the time and costs invested into simply signing more new artists, in hopes for one to succeed.
2. **Competitive Advantage**: UMG can gain a competitive advantage over their competitors as BeatScout allows UMG to reach out quicker and to a wider pool of talents. By doing so, UMG can outperform their competitors and acquire a larger share of the market.
3. **Better Financial Performance**: UMG can tailor their marketing efforts to capture the right audience through the right channels, further boosting the success of a hit song. With better investment of their resources, UMG can generate more revenue and achieve overall higher returns.

## **4.2 How UMG can extend / improve on BeatScout**

There are 3 possible areas of improvement to explore.

1. **Expand the Data Sources** : While BeatScout is currently only drawing data from Spotify API, UMG could benefit from insights provided by alternative streaming platforms such as Apple Music, Tidal and YouTube. This will provide a more comprehensive view of an artist’s potential and increase the accuracy of the predictions
2. **Include Non-Music Data** : In addition to analysing solely musical attributes, BeatScout can analyse non-music data such as an artist’s social media following, engagement rates and online presence. This can help UMG make a more comprehensive perspective of an artist and subsequently, make more informed decisions.
3. **Continuously Retrain the Model** : As the music industry and trends constantly evolve, it is important to continuously retrain the BeatScout algorithm to ensure that it remains relevant and accurate. This can be done through regular updating of the model with new data and adjusting the threshold for virality when necessary.

# **5. Conclusion**

In the competitive environment UMG is in, it is increasingly difficult for UMG's A&R personnel to identify and sign up-and-rising artists accurately which brings about inflated talent costs and opportunity costs. Hence, our proposed analytics solution, BeatScout, can and aims to address these issues by providing a streamline artist-recruitment process, making assessments of artists easier, time-efficient and cost-effective.

Powered by a Random Forest Model and a connection to the Spotify API, BeatScout is an all-in-one software which makes it convenient for A&R teams to assess hundreds or thousands of artists at once. Additionally, BeatScout eliminates the tedious process of sifting through the overwhelming amount of data by providing an accurate and narrowed selection of artists so that A&R teams are able to make a more informed and reasoned decision. Marketing teams can also tap on BeatScout to improve UMG’s ability to market both new and current artists effectively based on BeatScout’s predictions.

BeatScout is a sustainable and progressive solution for UMG. It is able to be continuously adapted and retrained with updated musical and non-music data such that it is increasingly relevant in the ever-evolving music industry. At the same time, retraining of the Random Forest Model in BeatScout will improve its accuracy, thereby strengthening the benefits brought by BeatScout. By leveraging the power of BeatScout, UMG is able to increase their revenue with better resource allocation and gain a competitive edge over their competitors by signing potential up-and-coming artists more accurately.

# References

1. Brownlee, J. (2021, April 30). *Tour of evaluation metrics for imbalanced classification*. MachineLearningMastery.com. Retrieved April 2, 2023, from <https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>
2. BSIC. (2021, October 3). *Universal Music Group out into the wild: The story behind its* *IPO*. BSIC. Retrieved February 25, 2023, from [https://bsic.it/universal-music-group-out into-the-wild/](https://bsic.it/universal-music-group-out-into-the-wild/)
3. Caswell, E. (2022, May 31). *We tracked what happens after Tiktok Songs Go Viral*. Vox. Retrieved February 26, 2023, from [https://www.vox.com/videos/23148752/viral-tiktok musicians-songs-data-investigation](https://www.vox.com/videos/23148752/viral-tiktok%20musicians-songs-data-investigation)
4. Common chord progressions - pop music. Mixed In Key. (2020, June 5). Retrieved April 1, 2023, from <https://mixedinkey.com/captain-plugins/wiki/common-chord-progressions-pop-music/#:~:text=C%20major%20and%20G%20major,the%20Major%20scale%20is%20best>
5. Horsburgh, A. (2021, November 8). *23 metrics that matter*. Medium. Retrieved March 31, 2023, from <https://amberhorsburgh.medium.com/23-metrics-that-matter-229cfba81440>
6. Fixmer, A. (2021, October 27). *Universal Music Group N.V. Reports financial results for the third quarter ended September 30, 2021*. UMG. Retrieved March 31, 2023, from <https://www.universalmusic.com/universal-music-group-n-v-reports-financial-results-for-the-third-quarter-ended-september-30-2021/>
7. *Frequently asked questions*. Billboard. (2022, July 5). Retrieved March 31, 2023, from [https://www.billboard.com/frequently-asked-questions](https://www.billboard.com/frequently-asked-questions/)/
8. Götting, M. C. (2022, November 22). *Global recorded music market worldwide by label 2021*. Statista. Retrieved February 26, 2023, from<https://www.statista.com/statistics/947107/recorded-music-market-worldwide-label/>
9. Greenwald, M. (2022, November 9). *Audience, algorithm and Virality: Why TikTok will continue to shape culture in 2021*. Forbes. Retrieved February 20, 2023, from<https://www.forbes.com/sites/michellegreenwald/2021/04/01/audience-algorithm-and-virality-why-tiktok-will-continue-to-shape-culture-in-2021/?sh=682548f12af7>
10. Kytka, M. (2022, September 27). *Universal Music Group Stock: A royalty on global music*  *consumption (OTCMKTS:UMGNF)*. Seeking Alpha. Retrieved February 26, 2023, from <https://seekingalpha.com/article/4543455-universal-music-group-a-royalty-on-global-music-consumption>
11. Langis, B. (2022, September 15). *Universal Music Group N.V. Stock: The music never stops (OTCMKTS:UMGNF).* Seeking Alpha. Retrieved February 25, 2023, from<https://seekingalpha.com/article/4476258-universal-music-group-nv-the-music-never-stops>
12. Leight, E. (2023, February 3). *More signings fail to produce more stars*. Billboard. Retrieved February 20, 2023, from<https://www.billboard.com/pro/more-artist-signings-record-labels-doesnt-mean-more-stars/>
13. Miller, L. S. (2019, January 10). *Menu*. New Report Illustrates How Modern Record Labels Remade Themselves in the Streaming Era |. Retrieved February 19, 2023, from<https://musonomics.org/modernlabelreport>
14. Pastukhov, D. (2020, February 10). *Market intelligence for the music industry.* Soundcharts. Retrieved February 26, 2023, from<https://soundcharts.com/blog/splits-and-profits-record-deals-analysis>
15. Runcie, D. (2021, August 4). *Inside universal, Sony, and Warner's arms race for your attention*. Trapital. Retrieved February 19, 2023, from [https://trapital.co/2021/06/21/inside-universal-sony-and-warners-arms-race-for-your-attention/](https://trapital.co/2021/06/21/inside-universal-sony-and-warners-arms-race-for-your-attention/%E2%80%AF)
16. Stassen, M. (2022, December 16). *13 out of the 14 no.1 songs in the US in 2022 were driven by viral trends on TikTok*. Music Business Worldwide. Retrieved February 20, 2023, from<https://www.musicbusinessworldwide.com/13-out-of-the-14-no-1-songs-in-the-us-in-2022-were-driven-by-viral-trends-on-tiktok/>
17. Trust, G. (2014, September 8). *Ask billboard: How does the hot 100 work?* Billboard. Retrieved March 31, 2023, from <https://www.billboard.com/pro/ask-billboard-how-does-the-hot-100-work/>
18. Viner, J. (2021, March 13). *How to make a viral TikTok song*. Medium. Retrieved April 1, 2023, from <https://medium.com/the-dopamine-effect/how-to-make-a-viral-tiktok-song-ece56a838a30/>
19. Virojsirasak, S. S. (2021, October 15). *What are factors that make song in Spotify popular?* Medium. Retrieved April 1, 2023, from <https://medium.com/@sunsunvirojsirasak/what-are-factors-that-make-song-in-spotify-popular-3cdcb3fb3a10>
20. Xiao'an, L. (2023, January 11). *Predicting song streams and artist popularity with Artificial Intelligence.* Blog @ Musiio. Retrieved April 1, 2023, from <https://blog.musiio.com/2021/09/23/predicting-song-streams-and-artist-popularity-with-artificial-intelligence/>
21. Zandt, F., & Richter, F. (2022, January 7). *Infographic: The shorter the song, the sweeter the stream?* Statista Infographics. Retrieved April 1, 2023, from <https://www.statista.com/chart/26546/mean-song-duration-of-currently-streamable-songs-by-year-of-release/>

# Appendices

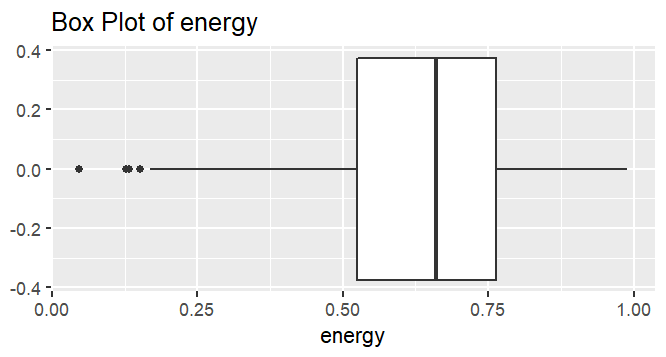
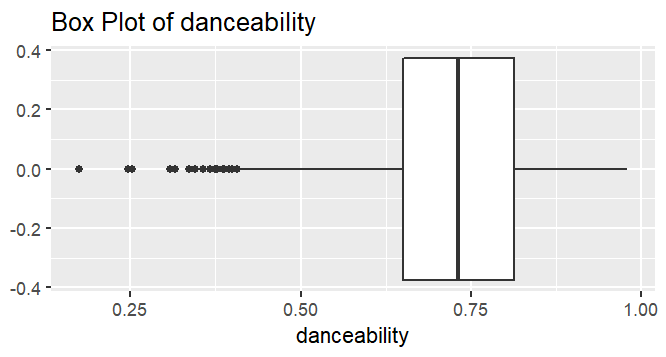
## **Appendix A—Data Dictionary**

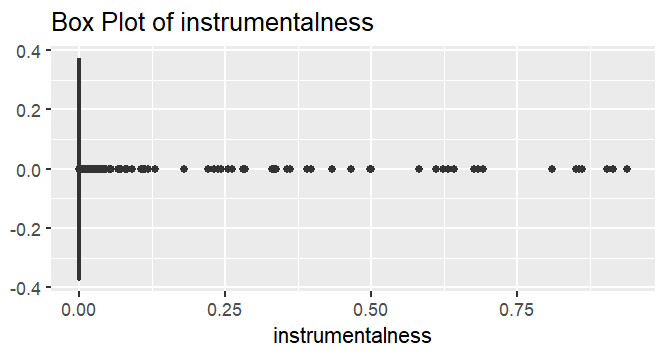
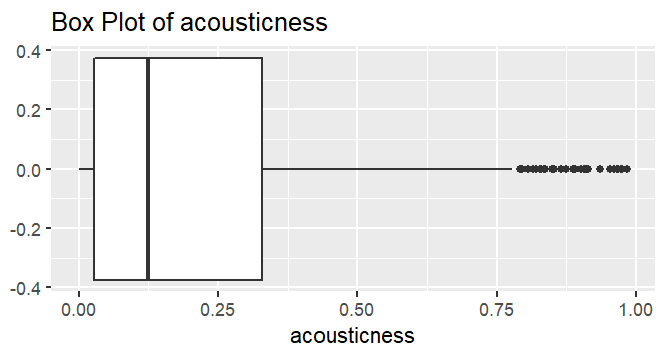
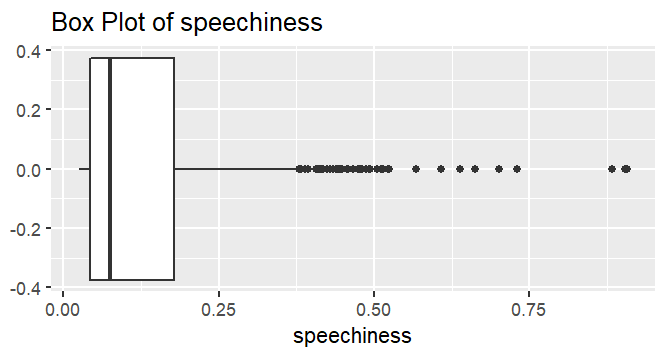
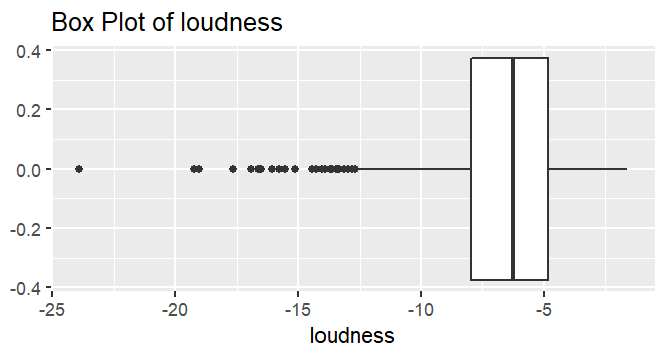
| **S/N** | **ATTRIBUTE** | **DESCRIPTION** |
| --- | --- | --- |
| 1 | virality | Target Y variable where the song is viral or not. |
| 2 | danceability | Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable. |
| 3 | energy | Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. |
| 4 | loudness | The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db. |
| 5 | mode | Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0. |
| 6 | key | The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1. |
| 7 | speechiness | Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks. |
| 8 | acousticness | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| 9 | instrumentalness | Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. |
| 10 | liveness | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. |
| 11 | valence | A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry). |
| 12 | tempo | The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration. |
| 13 | time\_signature | An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4". |
| 14 | duration\_ms | The duration of the track in milliseconds. |
| **Columns Removed** | | |
| 1 | track\_name | Name of track |
| 2 | artist\_name | Name of Artist |
| 3 | artist\_pop | Popularity of the artist |
| 4 | album | Name of the album |
| 5 | track\_pop | Popularity of the track |
| 6 | videos\_created | Number of videos created with the song on TikTok |
| 7 | Year\_of\_release | Year when the song is released |

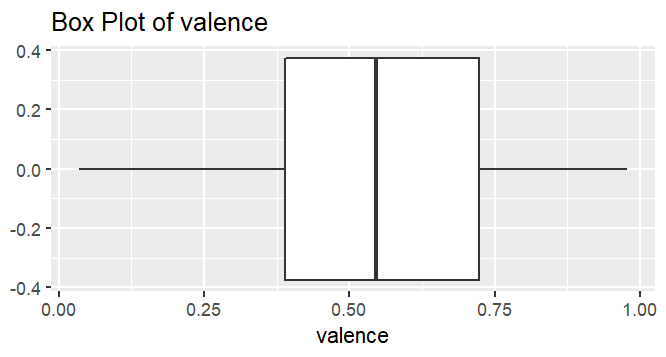
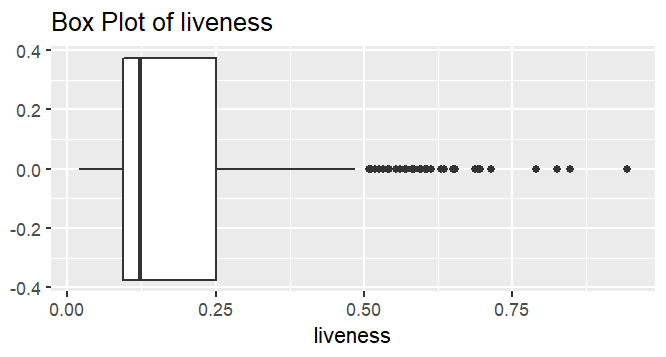
## 

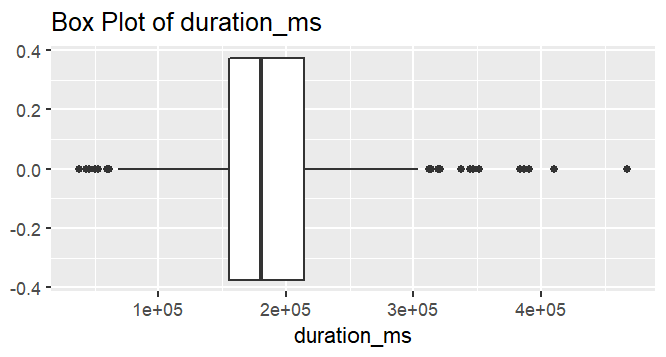
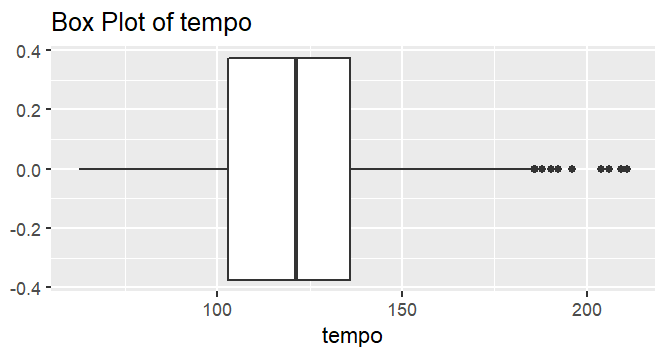
## 

## **Appendix B— Data Exploration Plots**

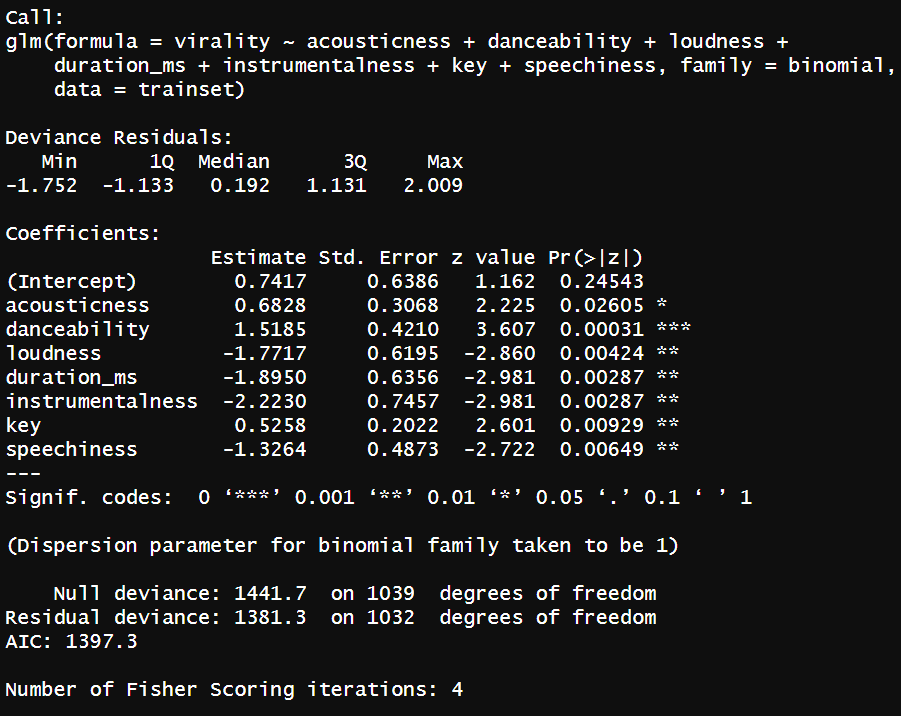




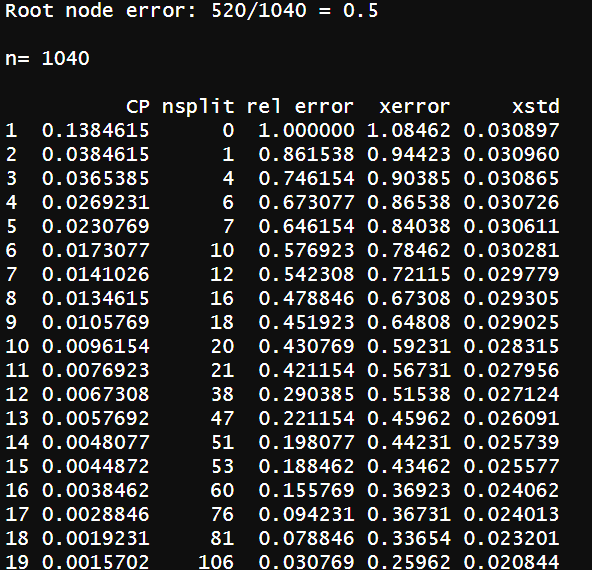




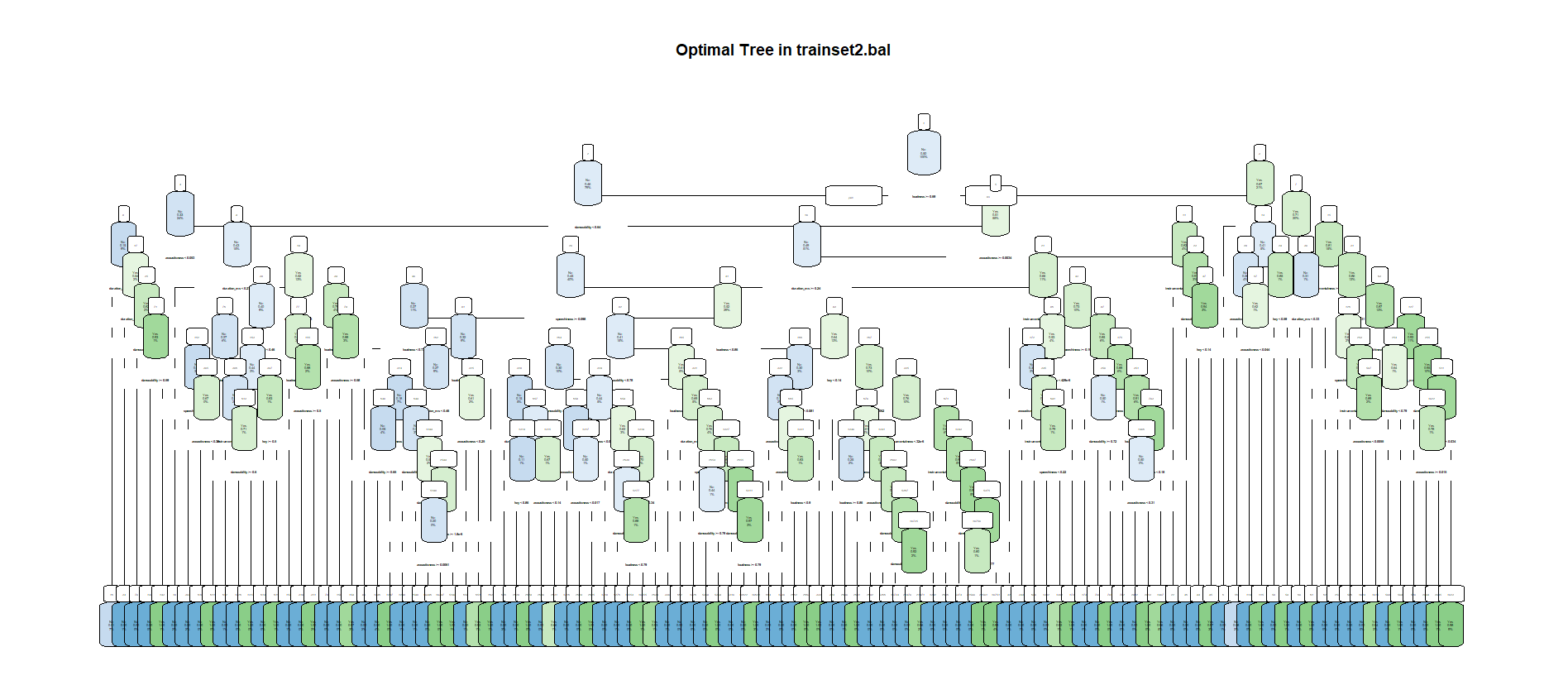
## **Appendix C - Data Modelling**



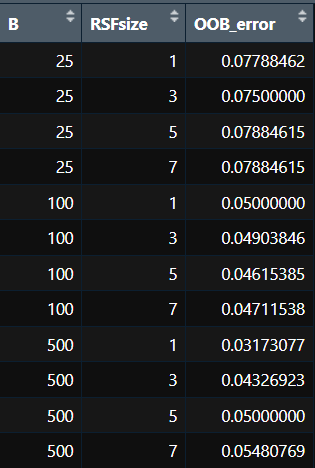
*Figure 1: Summary of Final Logistic Regression Model*



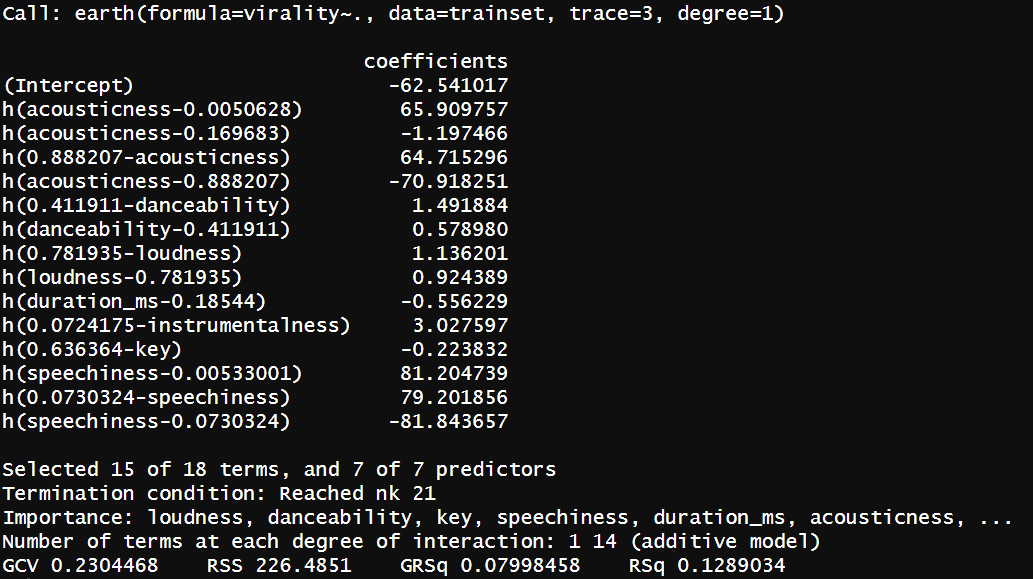
*Figure 2: Final CP Table for CART*

**

*Figure 3: Optimal Tree in CART*



*Figure 4: Table Comparing OOB Error in Random Forest*



*Figure 5: Summary of Final MARS Model*