CONTENTS

INTRODUCTION

This report focuses on artists' performance and user engagement on Trackd. A brief overview of the user demographics and behaviours will be discussed, with data visualisation of important metrics on Trackd's usage such as signup and sign-in activities, artist collaborations and revenue from Chipin subscription. We will also be utilising principal component analysis, clustering, and modelling to gain deeper insights on artists' performance. The report will conclude by shedding light on the limitations of our study and provide strategic recommendation for Trackd's platform.

DATA PREPARATION FOR DATA VISUALISATION

Firstly, we extracted the important tables for the purpose of our analysis from the original Trackd dataset such as users, Chipin, playlist, studio song tables etc. Next, we joined the tables through unique keys to prepare a separate dataset for visualisation.

GENERAL OVERVIEW

Users are mostly concentrated between 20-40 years old (Figure 1). Consumers of Trackd are predominantly male (34.38% men, 19.58% women and the rest unknown) (Figure 2); and consist of more listeners than artists (Figure 3).

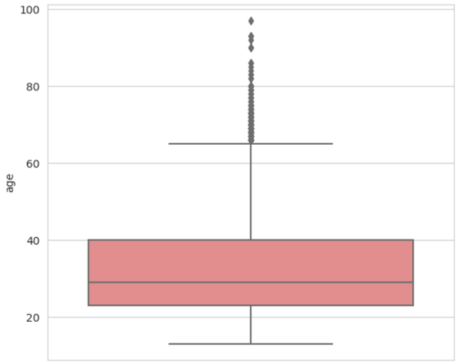


Figure 1: User Age Range

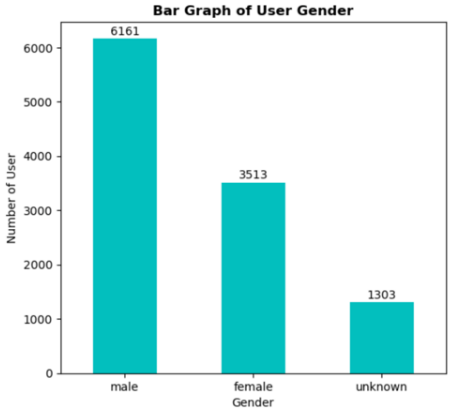


Figure 2: User Gender Distribution

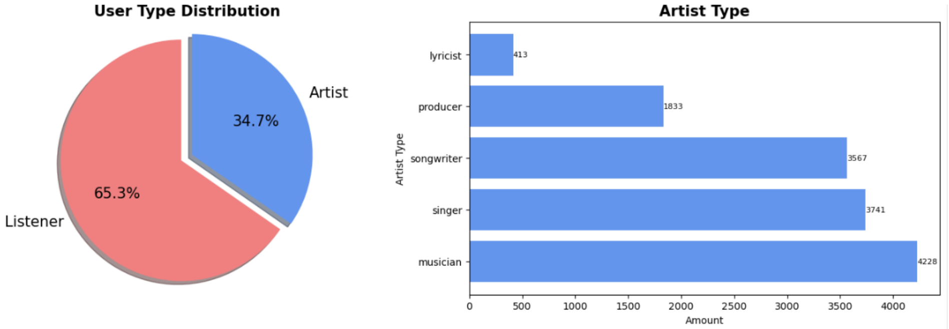
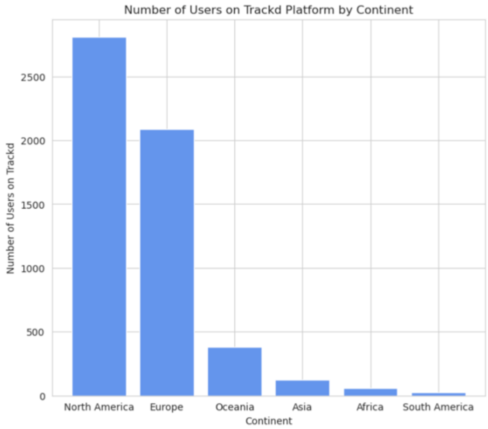


Figure 3: Listeners and Artists

Trackd is most popular in North America and Europe, with most users being based in the US and UK (Figure 4) and (Figure 5).



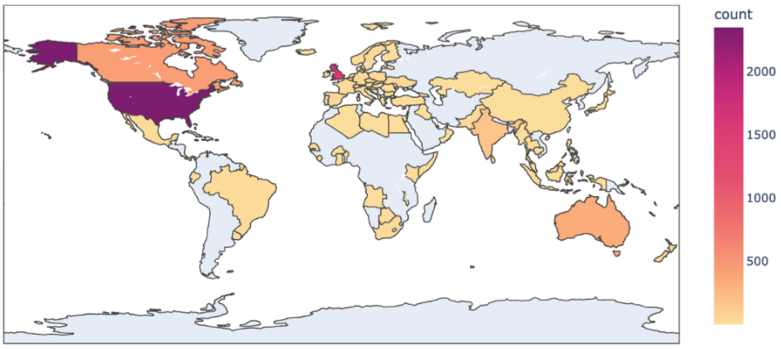


Figure 4: User Continental Distribution

Figure 5: Usage by Country

To study user preference, we extracted the 30 most played and downloaded songs to identify the genre of songs that fell into both categories to find that Vocals, Guitar and Keys are most popular (Figure 6).

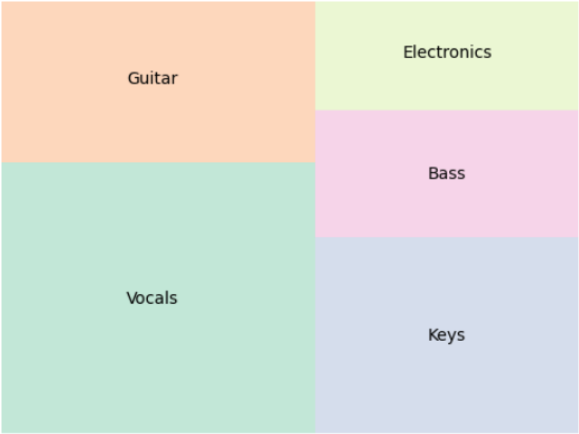


Figure 6: Popular Genres

There are only 230 users who utilise the Chipin/Chipin+ subscription plan, which is about 3.7% of all artists- a significantly low use of the service considering it is the only source of income for the platform. Majority of the Trackd users have renewed their subscription (38.4%), while 16.3% have cancelled, which amounts to 2928 customers lost. (Figure 7).

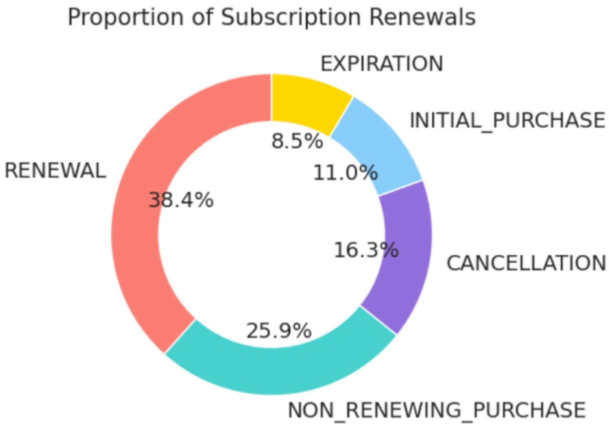


Figure 7: Proportion of Renewals

EXPLORATORY DATA ANALYSIS

To study user behaviour, the sign-up and sign-in activity was analysed, findings show that people have a much higher preference for accessing Trackd through their app rather than the website; this could be due mobile usage increasingly overtaking desktops due to its convenience and accessibility as current statistics show that 60% of all search volumes consist of mobile searches (Search, 2023). Nevertheless, a continuous decrease in sign-ups is seen on both platforms since May 2021 (Figure 8). Similarly, the number of sign-in via iOS experienced a sharp drop during September 2021(Figure 9), and engagement has been decreasing since.

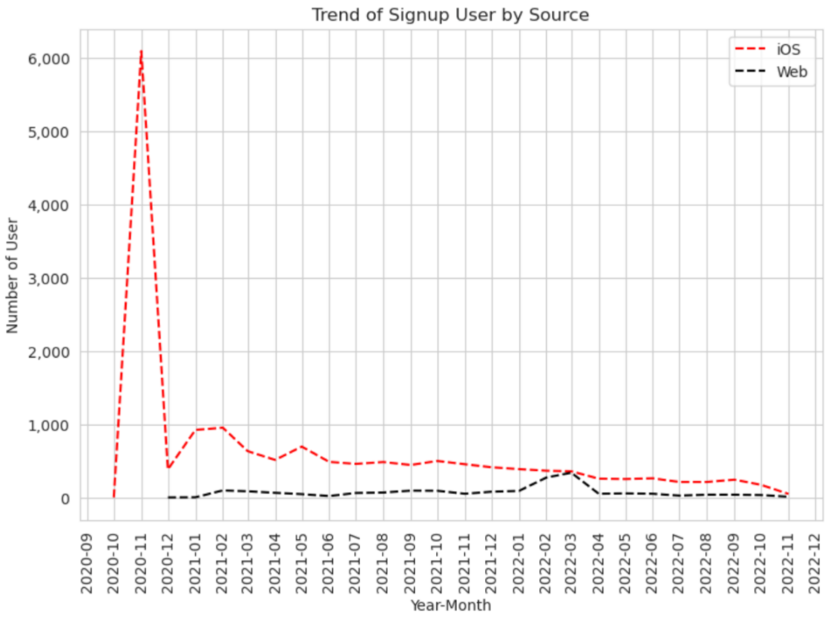


Figure 8: Trend of Sign-up Users by Source

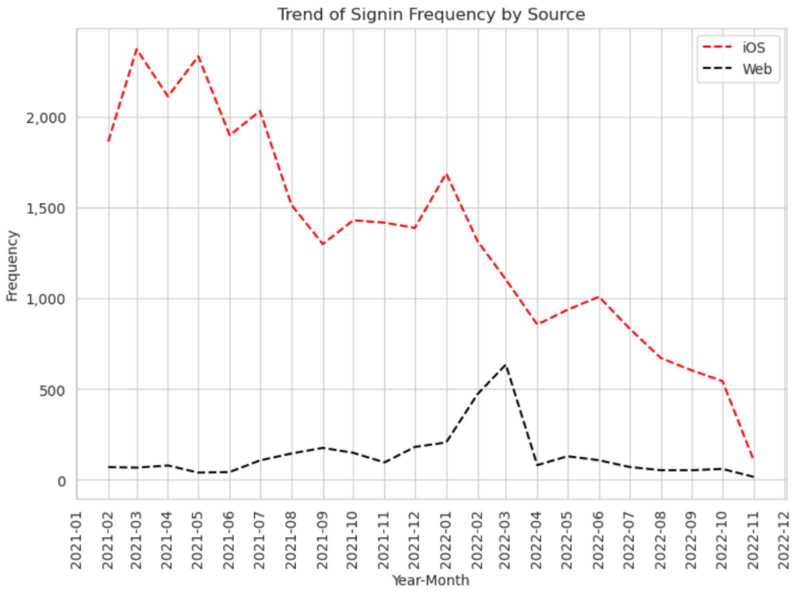


Figure 9: Trend of Sign-in Frequency by Source

Another consumer behaviour analysis is the Chipin/Chipin+ usage. While Trackd have a similar number of Chipin (110) and Chipin+ subscribers (120), people with Chipin are more likely to profit from their subscription- 41% of these consumers get paid while only 7.5% of Chipin+ users generate money (Figure 10). We believe that this is because Chipin subscribers can profit from 95% of the direct donations from fans. However, Chipin+ subscribers have to set a specific number for the subscription plan and fans have to comply with the high prices while only receiving 85% of the money *(Music, 2023)*.

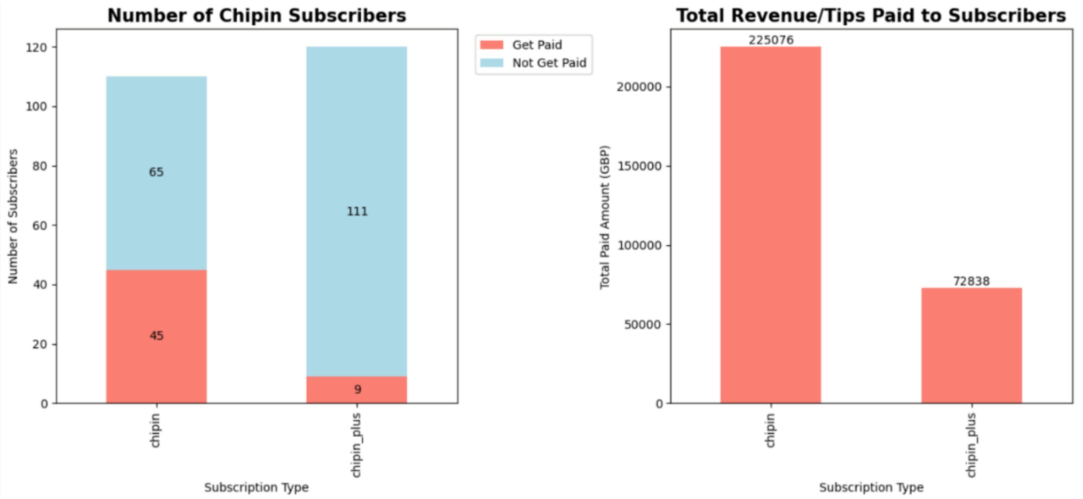


Figure 10: Chipin Subscribers and Total Revenue

To understand the determining factors of artists' revenue, we studied the tips earned from Chipin against variables such as total songs, plays and likes (Figure 11). Our findings illustrate that while more liked songs have a slightly better advantage of reaping the Chipin benefits, high revenue is only enjoyed by a select few artists, regardless of these factors- potentially those who are more recognised or have spent longer building their portfolio. Nevertheless, when comparing revenue against plays vs downloads (Figure 12), it is evident that most played songs have received more money than those that have been downloaded the most. Our assumption is that number of plays is a better determining factor of popularity- a trending song will be played repeatedly but a downloaded song just indicates it being in people's playlists.

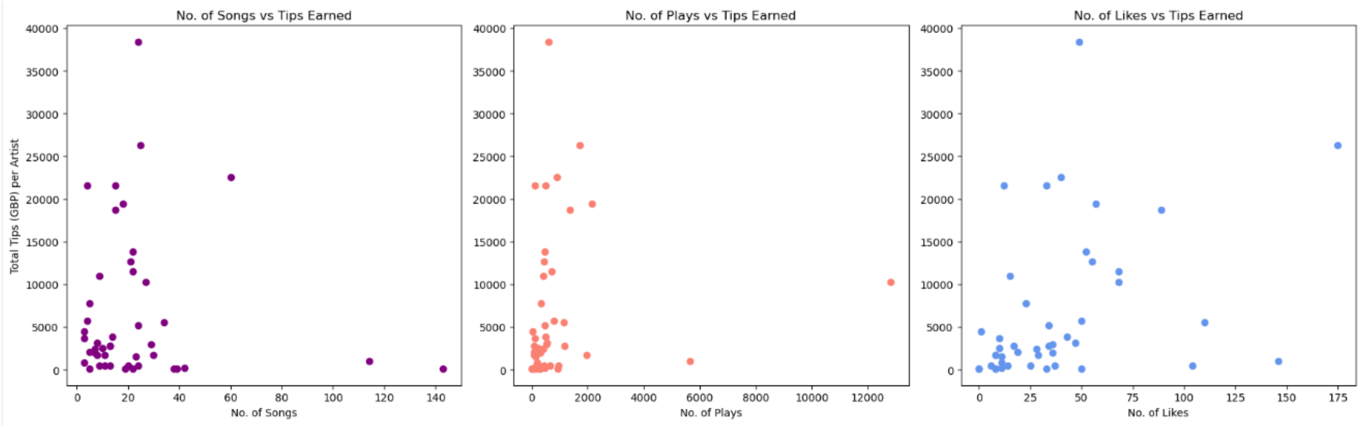


Figure 11: Chipin for Total Songs, Plays and Likes

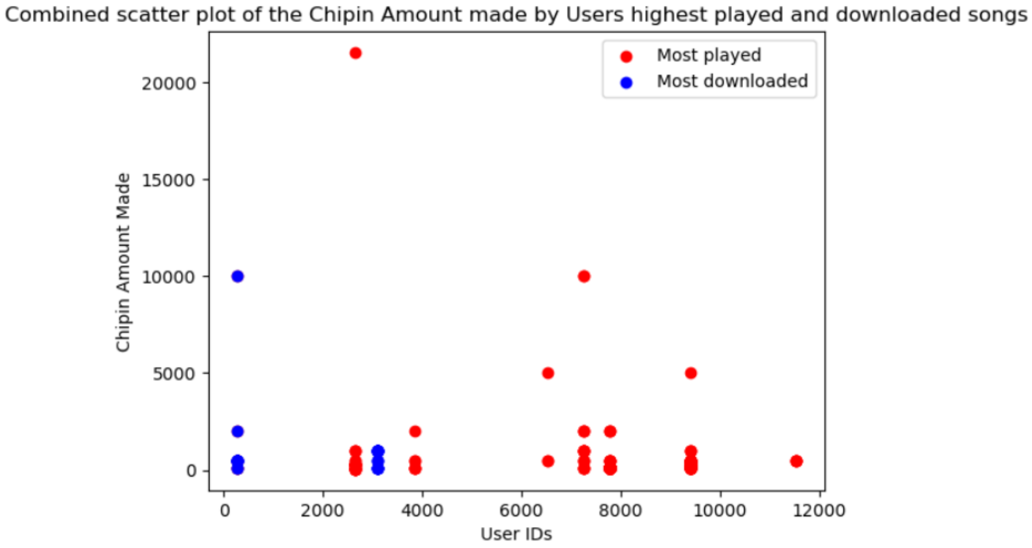


Figure 12: Chipin from most played and most downloaded songs

Next, we explored the impact of artist collaboration on engagement (Figure 13). Collaboration on Trackd is extremely low (0.9%), potentially due to artists collaborating ‘justforfun’ rather than to grow their portfolio. The grouped bar chart illustrates those songs where collaboration occurred received significantly lower average likes and absolutely no downloads (Figure 14); given our previous finding these songs are also less likely to be profitable.

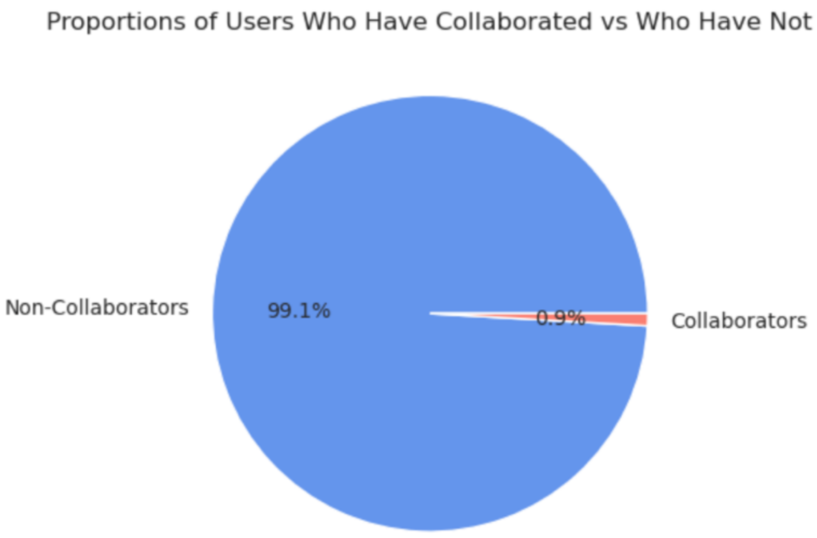


Figure 13: Proportion of Users Collaboration

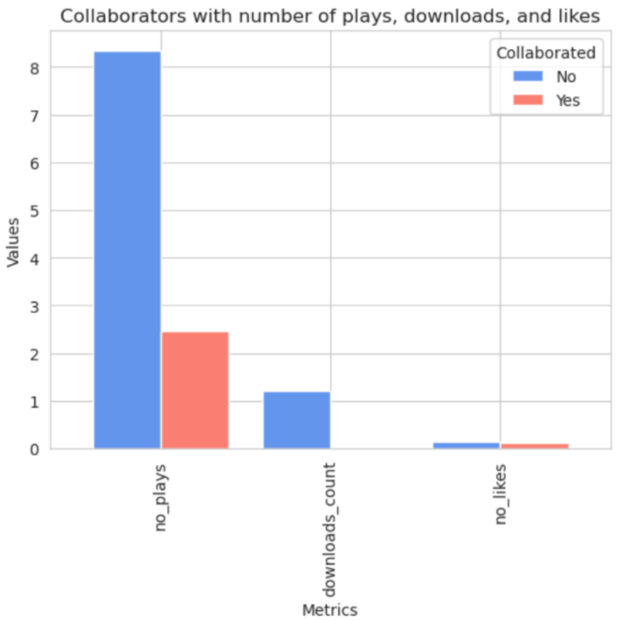


Figure 14: Collaborators Performance Based on average song plays, downloads, and likes

PRINCIPAL COMPONENT ANALYSIS

Doing PCA before clustering can help reduce dimensions, improve signal to noise ratio and cluster accuracy by removing redundant features. First, we performed PCA on the artist data set and choose two principal components based on the Scree plot and using the elbow method, as shown in Figure 15.

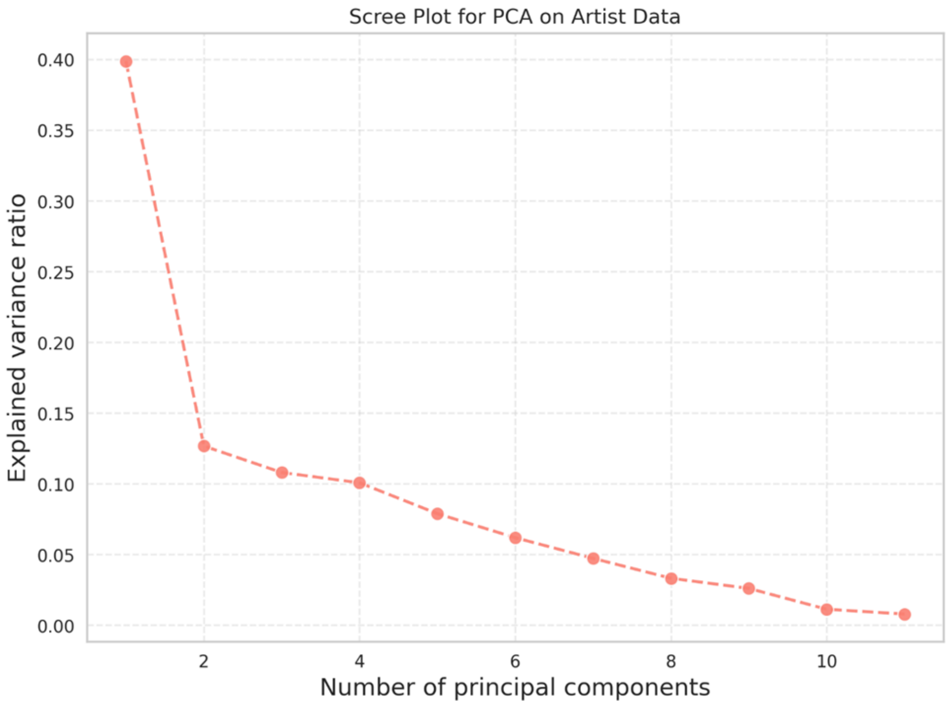


Figure 15: Scree Plot for PCA on Artist Data

The two principal components have an explained variance ratio of 0.4 and 0.13 respectively, as shown in Table 1.

Table 1: Explained Variance in Component 1 & 2

Component	Explained Variance
1	0.398731
2	0.126637

According to Table 2, the top 3 most important loading for each component is listed. Based on the top loadings of each component, we labelled Component 1 as "Artist Digital Footprint Metrics" and Component 2 as "Artist Fan Engagement Metrics", as shown in Table 2 and Table 3, respectively.

Table 2: Top 3 Loadings in Component 1

Component 1	
Artist Digital Footprint Metrics	
Variable	**Loading**
artist_number_of_songlikes	0.445628
artist_totalnumbersonginplaylist	0.424032
artist_numbersongprofileview	0.423366

Table 3: Top 3 Loadings in Component 2

Component 2	
Artist Fan Engagement Metrics	
Variable	**Loading**
artist_number_of_tags	0.550142
artist_number of profileclassifications	0.521732
artist_number_of_downloads	0.415299

Visualisation of the PCA data can also reveal the structure of the high-dimension and show how the data are clustered or spread out in a reduce dimension space, as depicted in Figure 16.

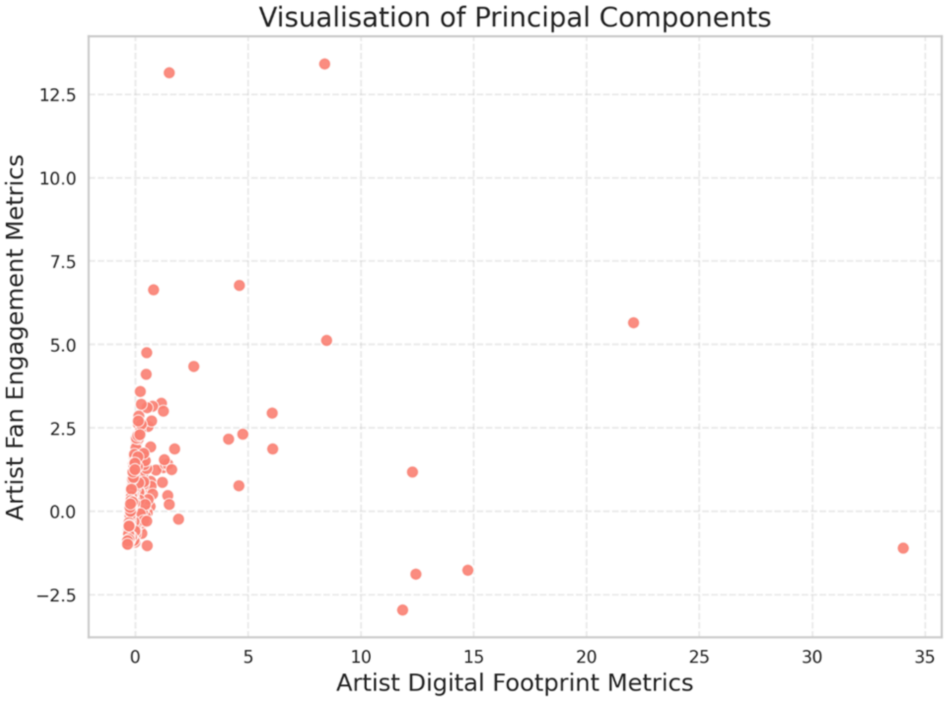


Figure 16: Visualisation of Principal Components

CLUSTERING

Next, based on the 2 Principal Components selected, ‘Artist Digital Footprint Metrics’ and ‘Artist Fan Engagement Metrics’, Cluster Analysis was performed to aid Trackd to better understand its artists, segmenting them to determine groups with shared characteristics.

All Cluster Analysis conditions were satisfied, including the sample size, no correlation (factor scores) (Figure 17), and measurement levels (scales) (Frades&Matthiesen, 2009).

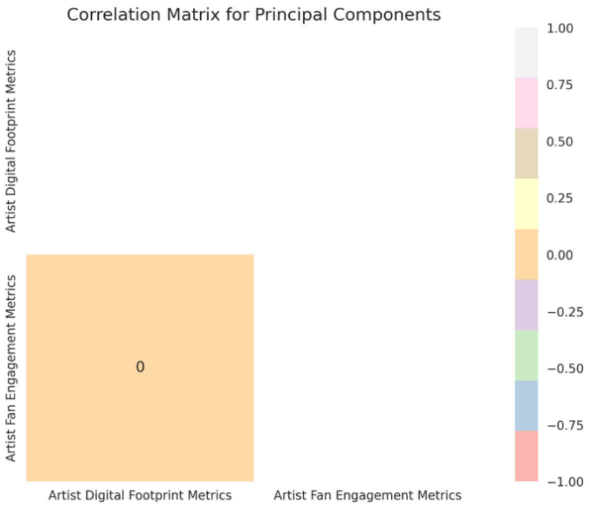


Figure 17: Correlation Matrix for Principal Components

The K-Means Clustering (Partitional Algorithm) was selected as a primary method to determine the cluster memberships due to its computational efficiency and interpretability (Wu&Lin,2005).

The Silhouette Score and Scree Plot (Inertia/Elbow) method were employed to determine the optimal number of clusters. Both methods showed 2 clusters as being optimal (Table 4 and Figure 18).

Table 4: Silhouette Scores for each number of clusters (2-10 Clusters)

Number of clusters	Silhouette score
2	0.964
3	0.888
4	0.580
5	0.586
6	0.586

7	0.590
8	0.599
9	0.578
10	0.581

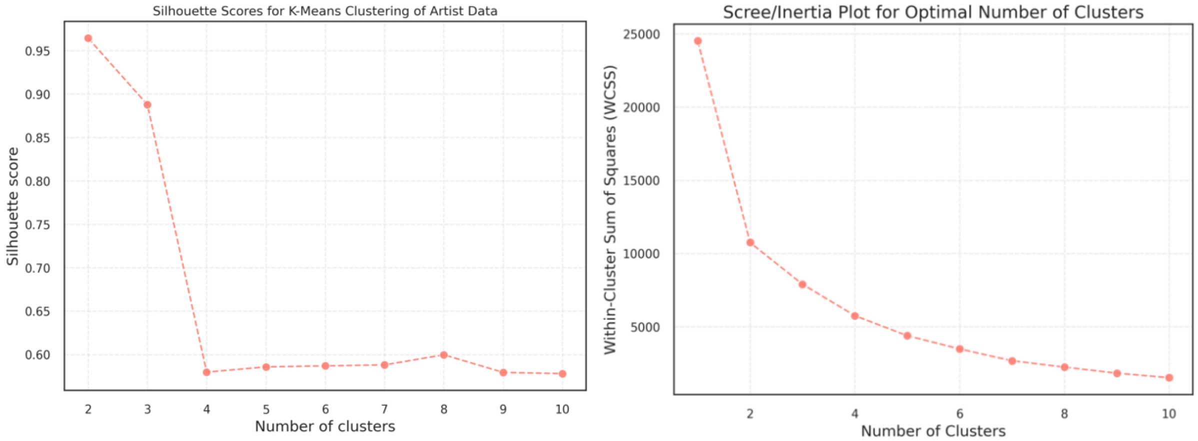


Figure 18: Silhouette Score and Inertia Methods for determining the Optimal Number of Clusters

However, since dividing the whole customer database into solely 2 clusters might be unreasonable, the 3 cluster solution was also evaluated (Figure 19 and Figure 20).

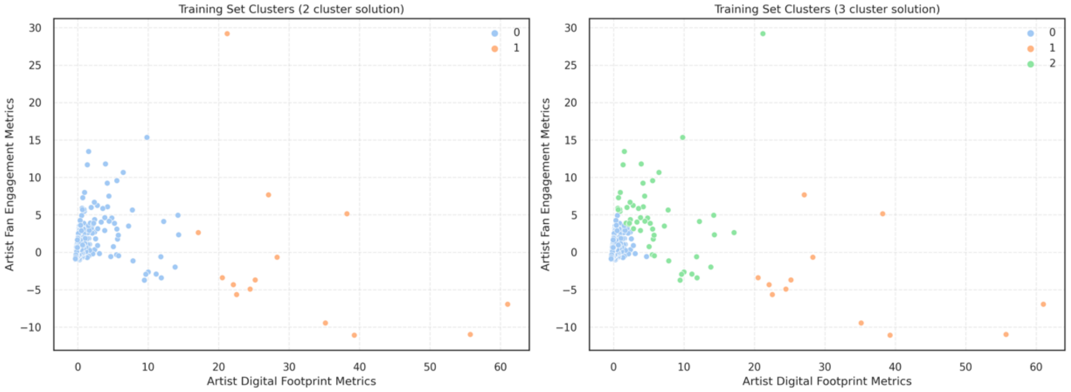


Figure 19: Scatterplot representing the distribution and relationship of the data points in a 2D space (for the 2 and 3 Cluster Solution) on the Training Set

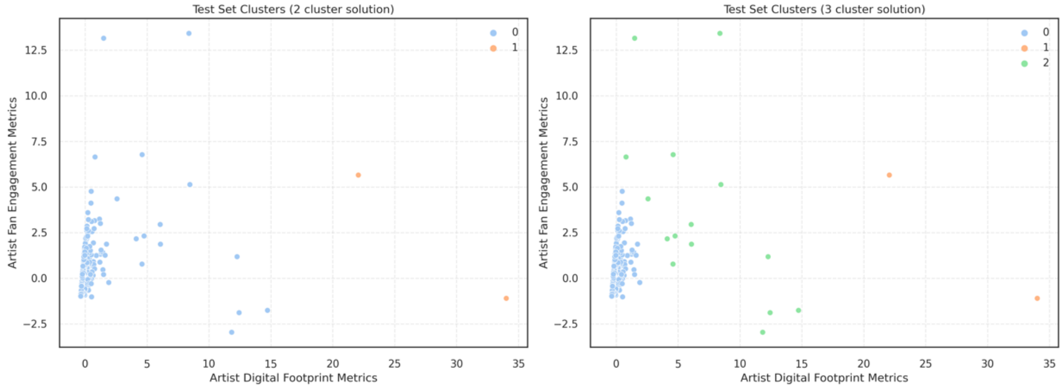


Figure 20: Scatterplot for the 2 and 3 Cluster Solution for the Test Set

Ultimately, due to too few cases pertaining to cluster 1 and 2, the 2 cluster solution was retained and artists' cluster memberships were then examined for this solution (Figure 21 and Figure 22).

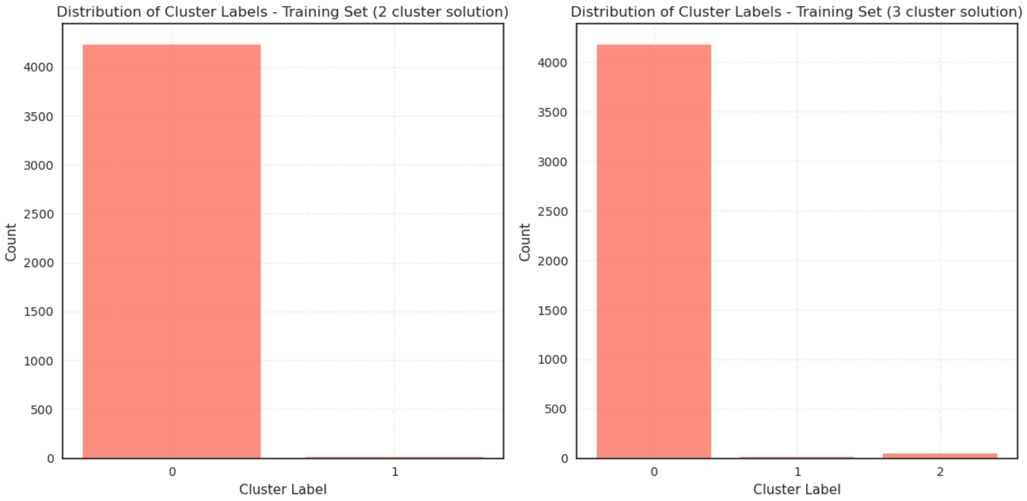


Figure 21: Distribution of Cluster Labels on the Training Set (2 and 3 Cluster Solution)

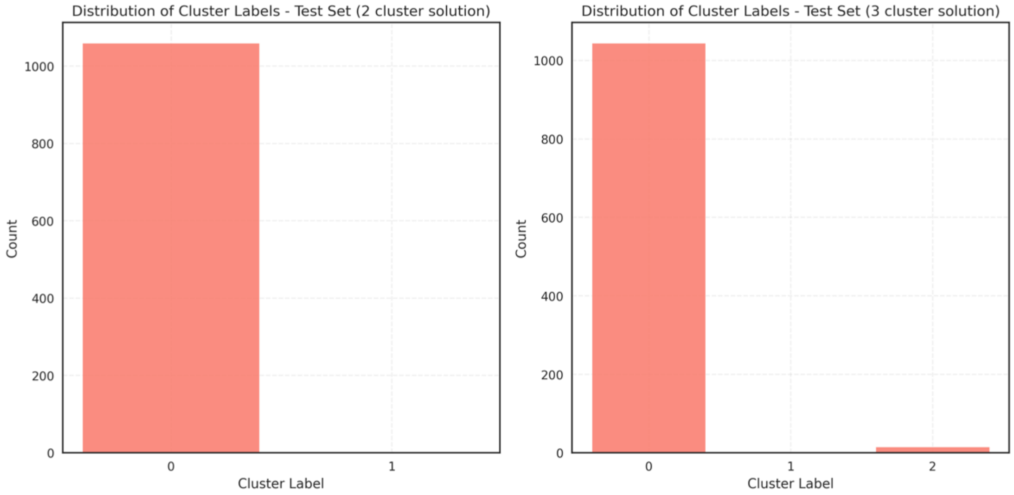


Figure 22: Distribution of Cluster Labels on the Test Set (2 and 3 Cluster Solution)

According to Figure 23, the distribution of mean PCA scores demonstrated that Cluster 2 (Label 1) scored high in both components, having a high digital footprint and fan engagement. Thus, these artists possessed solely 14 and 2 cases in the training and test sets, respectively. The dominant cluster (Label 0), however, had a mean of 0, meaning that most artists on the platform have low customer engagement and digital footprint. Analysing the separate variables for this cluster, the highest variables were number of songs' likes and number of downloads (Figure 24). Therefore, Trackd should motivate artists to enhance their digital footprints to increase the popularity among the users, which would, potentially, enhance the platform's popularity.

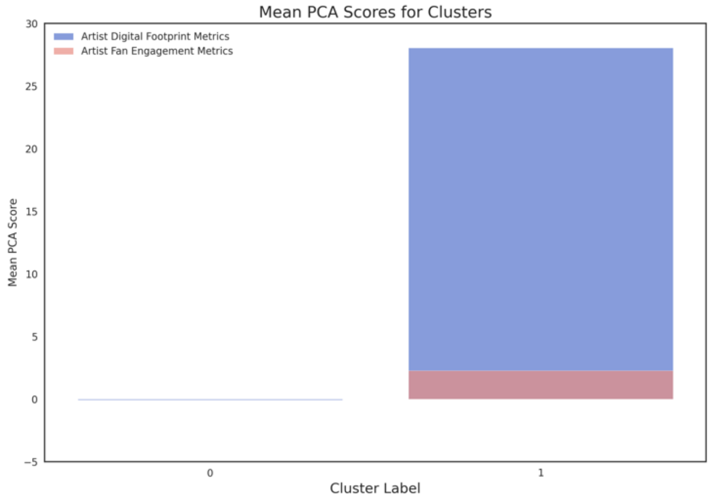


Figure 23: Mean PCA Scores for the 2 Cluster Solution (on the Test Set)

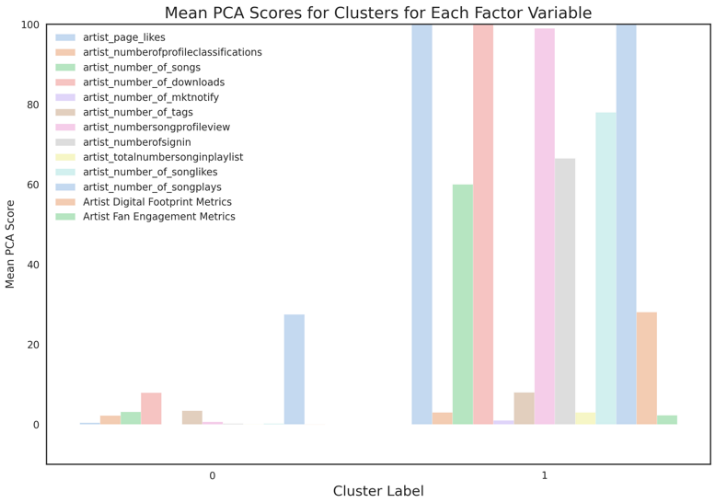


Figure 24: Mean PCA Scores for each Factor Variable for the 2 Custer Solution (Test Set)

The cluster solution validity was assessed with different clustering algorithms, including Hierarchical clustering (agglomerative) and DBSCAN (partitional) (Frades&Matthiesen, 2009). The dendrogram in the Agglomerative Hierarchical Cluster Analysis (with Ward method to calculate cluster distance) suggested a similar solution to K-Means, with 2 dominant clusters and the most cases being attributed to one of them, as shown in Figure 25. For the DBSCAN method, a minimum of 5 data points were selected to form a dense region (cluster). Selecting epsilon of 0.5 given to capture the cluster structure given the spread-out data structure, allowing for larger neighbourhoods and, potentially, larger clusters(ibid.). However, the algorithm demonstrated that most cases assigned to cluster 2 and 3 in K-Means 3-cluster solution or merged to cluster 1 in the 2-cluster solution were 'noise' present in the data, comprising 71 data points, according to Figure 26. Hence, meaningful interpretations could solely be drawn from Cluster 0 mean values.

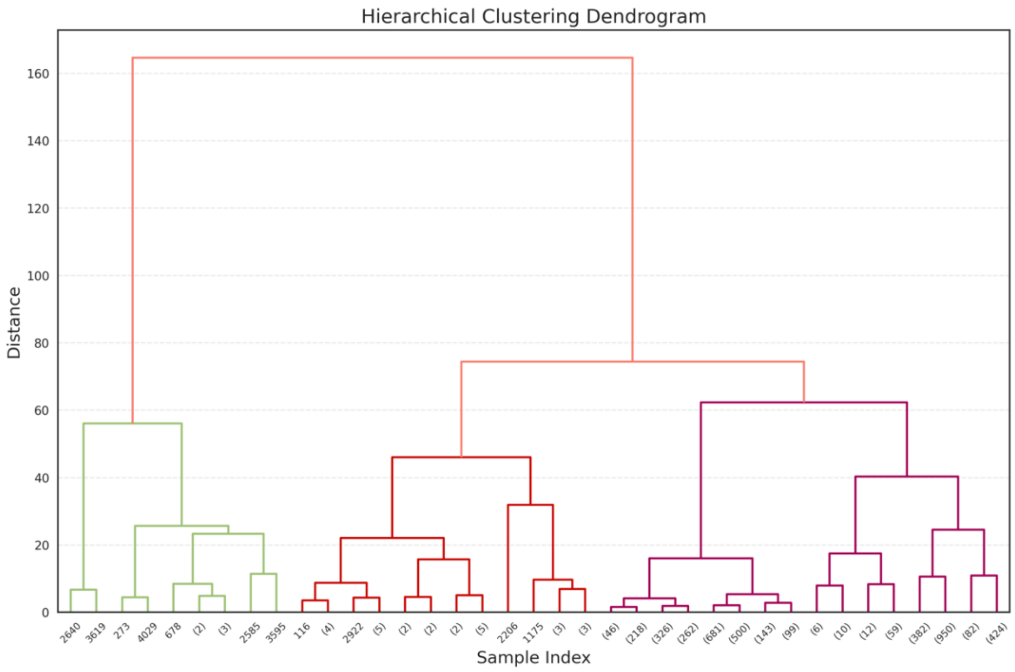


Figure 25: Validating the 2-Cluster Solution with the Hierarchical Clustering Algorithm

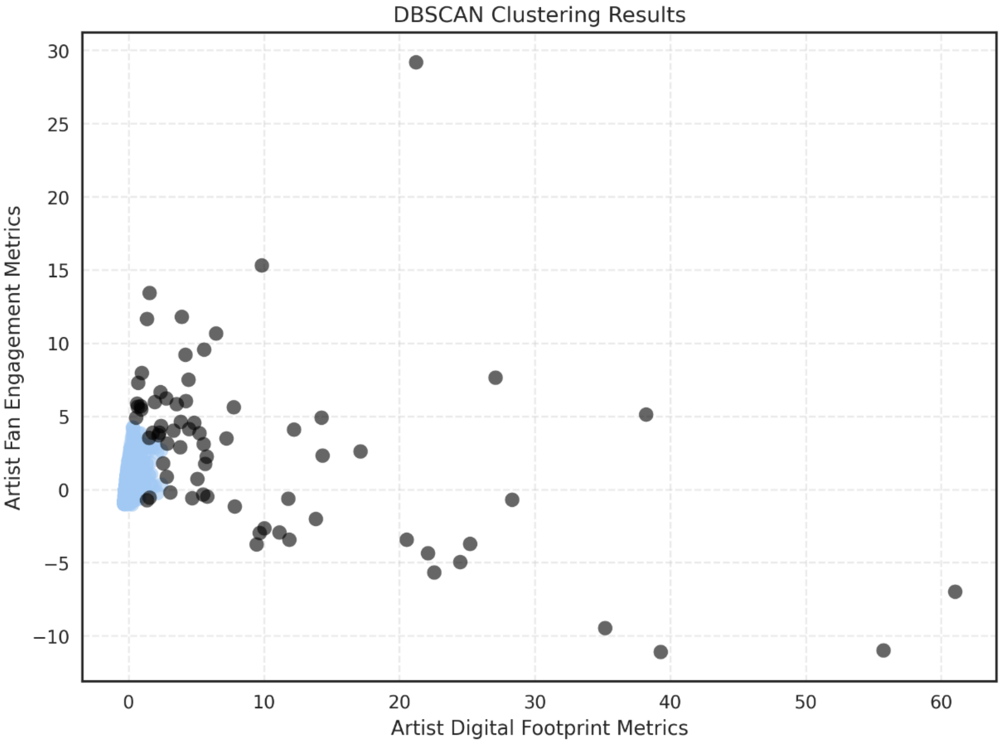


Figure 26: Validating the 2-Cluster Solution with the DBSCAN Algorithm

Due to having solely 1 dominant cluster, predictive models were conducted on the whole dataset instead of performing them separately for each cluster. Overall, a larger sample size would allow to draw more meaningful conclusion from the cluster analysis.

PREDICTIVE MODELING

The data utilised in the predictive modeling section is extracted from Trackd SQL database dump file consisting of a rich amount of artist level characteristics and their performances on the platform, such as total number of like, streams, profile views, etc. All the variables, including 18 predictors and a target variable, are described in Table 5. There were a total of 1267 observations after aggregating the meaningful metrics and dropping most missing values.

The target variable is "**subscribed_chipin**" which indicate if artists are subscribed to the Trackd platform. By predicting if artists are likely to pay a premium to be a Chipin member, Trackd can allocate resources such as marketing promotion effectively to increase platform revenue.

Table 5: Data Dictionary for Predictive Modeling Part

Predictors (X)	Data Type	Description
artist_page_likes	Numeric (integer)	Artist's Total Page Like
artist_numberofprofileclassifications	Numeric (integer)	Artist's Number of Profile Classifications
artist_number_of_songs	Numeric (integer)	Artist's Total Number of Songs
artist_number_of_downloads	Numeric (integer)	Artist's Total Number of Downloads
artist_number_of_mktnotify	Numeric (integer)	Artist's Total Number of Marketing Notifications
artist_number_of_tags	Numeric (integer)	Artist's Total Number of Tags
artist_numbersongprofileview	Numeric (integer)	Artist's Total Number of Song Profiles' Views
artist_numberofsignin	Numeric (integer)	Artist's Total Number of Sign-ins
artist_totalnumbersonginplaylist	Numeric (integer)	Artist's Total Number of Songs Added to Playlists
artist_number_of_songlikes	Numeric (integer)	Artist's Total Number of Songs' Like
artist_number_of_songplays	Numeric (integer)	Artist's Total Number of Songs' Plays
allow_collaboration	Boolean	Indicates whether the particular artist is opened for collaboration on Trackd platform or not (Yes = 1, No = 0)
allow_comments	Boolean	Indicates whether the particular artist is opened for comments to be public on Trackd platform or not (Yes = 1, No = 0)

show_skills	Boolean	Indicates whether the particular artist is showing the skills being specified on Trackd platform or not (Yes = 1, No = 0)
hide_profile	Boolean	Indicates whether the particular artist is willing to hide the profile on Trackd platform or not (Yes = 1, No = 0)
email_verified	Boolean	Indicates whether the particular artist has verified his- or herself through email channel or not (Yes = 1, No = 0)
is_male	Boolean	Indicates whether the particular artist is a male artist or not (Yes = 1, No = 0)
is_female	Boolean	Indicates whether the particular artist is a female artist or not (Yes = 1, No = 0)
Target Variable (Y)		**Description**
subscribed_chipin (Binary Response)	Boolean	Indicates whether Trackd should target the particular artist or not • Yes (denoted by class 1) o If the artist has subscribed to Chipin, he or she is considered as likely to generate revenues for Trackd so it is worth for Trackd to target this artist. • No (denoted by class 0) o If the artist has not subscribed to Chipin, he or she is considered as likely not to generate revenues for Trackd so it is not worth for Trackd to target this artist.

The objective of the task is to target artists that are likely to have a Chipin subscription, thus increasing revenue for the company. Therefore, we focus on precision score which prioritise artist that are indeed subscribing to the platform (True Positives).

Initially, Logistic Regression and Random Forest were used to train the model. However, due to the class imbalance in the dataset, Balanced Random Forest and an oversampling technique (SMOTE), were introduced as there were significantly fewer artists that have subscribed to both Chipin and Chipin plus (33 observations), as compared to artists that did not subscribe to any of Trackd's premiums (1,234 observations).

After tuning hyperparameters with GridSearch CV for all models, we find that the tuned random forest has the highest precision on the test set at 75% and an accuracy of 98.81% The confusion matrix of the tuned random forest model is shown in Figure 27.

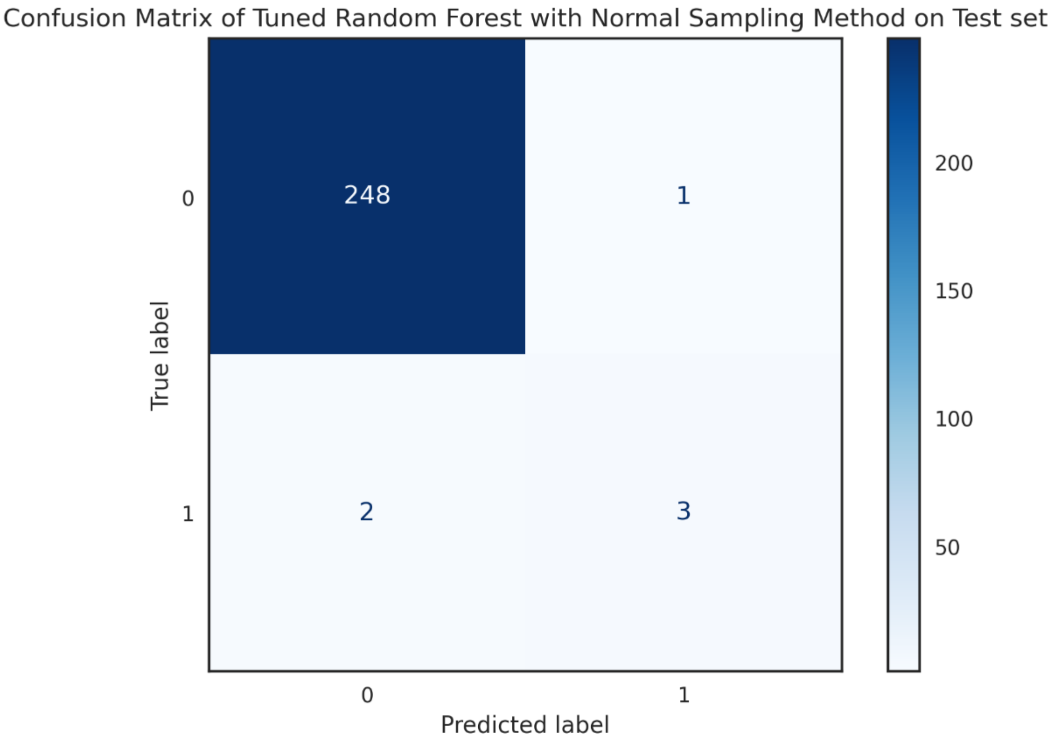


Figure 27: Confusion Matrix of Tuned Random Forest with Normal Sampling Method on Test set

The feature importance score is represented in a bar chart, as shown in Figure 28. The number of all song profiles' views for each artist has the highest feature importance score, following by number of times that the artist had signed-in to the platform, and the number of songs' likes.

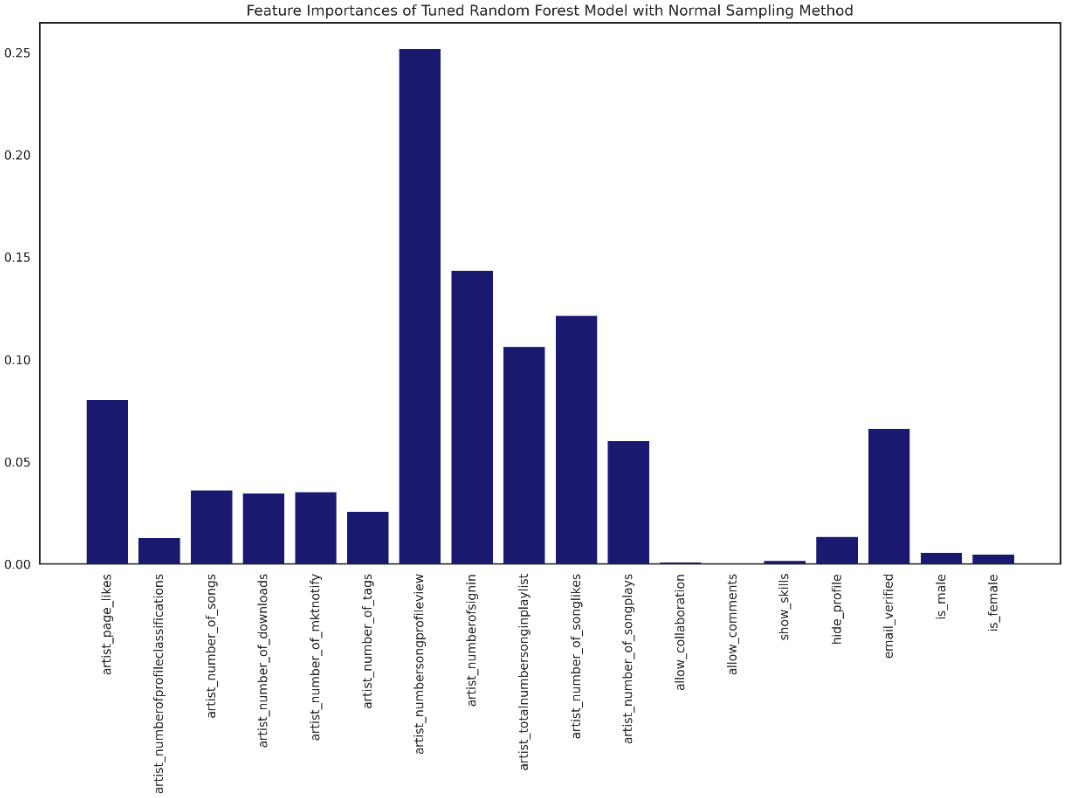


Figure 28: Feature Importance of Tuned Random Forest with Normal Sampling Method

LIMITATION AND FUTURE WORK

Due to the imbalanced dataset where majority of the data points falls under class 0 for our target variable "**subscribed_chipin",** we chose to focus on Precision where we prioritise the prediction of artists that are likely to subscribe for Chipin. This suggests that the models chosen for our study might not be the best model to be used for general accuracy. In addition, we were only able to use 1267 observations out of 5303 as originally intended due to missing values in the dataset for clustering and modelling. The clusters and prediction made would be more generalizable to all Trackd users if there were more complete set of data available for analysis.

RECOMMENDATION

Given the uneven distribution of listeners to artists, Trackd can focus on acquiring social celebrities such as influencers to increase artists on the platform while onboarding their already large fan-base. Moreover, implementing media marketing strategies on Instagram and Facebook can improve Trackd's customer acquisition rate since a large volume of their users are young adults (Figure 1) - studies show that 18-29 year olds accounts for 88% of the population of these popular platforms (Shaffer, 2023).

Since users prefer accessing Trackd through iOS (Figure 8 & Figure 9), the company should make the app accessible to android users to increase engagement- Android dominates the current mobile industry by maintaining 72.2% of the global market share (Wallen, 2021); hence, allowing these consumers to access Trackd through their phones will greatly increase user engagement. Currently, users must sign-in every time they open the app; providing consumers the option to stay signed in can simplify user-experience and result in more engagement. Trackd can further invest in their UX by allowing users to personalise their profile to match their music theme, push notification features to keep people informed about new music, and rewarding/incentivising engaged users.

Users that invest in Chipin+ do not profit as much as Chipin subscribers, even though they pay more. We recommend that Trackd implement similar incentives that Chipin subscribers indulge in- allowing listeners to donate an amount that is comfortable for them will encourage more fans to contribute and in turn, generate more money for the Chipin+ subscribers. An alternative is to also allow Chipin+ subscribers to benefit from 95% of the fan's monthly subscription plan, since Chipin subscribers receive 95% of their fan's willing contributions.

Trackd being a social music platform, value collaboration yet, our findings illustrate that less than 1% of current users collaborate (Figure 13). A method of increasing engagement between artists is a forum implemented on Trackd's platform, where artists can promote themselves and communicate the type of collaborator they want to engage with. This will encourage discussions between artists who have similar visions for their musical content and bring together people who can help each other grow. These approaches will help Trackd's value proposition and establish its niche within the market.

REFERENCES

Music, T. (2023). *Features and Pricing*. Retrieved 03 01, 2023, from https://trackdmusic.com/pricing

Shaffer, N. (2023). *Social Media Demographics: What Marketers Need to Know In 2023*. Retrieved 03 03, 2023, from https://nealschaffer.com/social-media-demographics/#:~:text=Studies%20show%20that%20young%20adults,Facebook%2C%20Instagram%2C%20and%20Twitter

Wallen, J. (2021). *Why is Android more popular globally, while iOS rules the US?* Retrieved 03 03, 2023, from TechRepublic: https://www.techrepublic.com/article/why-is-android-more-popular-globally-while-ios-rules-the-us/#:~:text=The%20primary%20reason%20why%20Android,on%20a%20phone%20is%20cost.

Search, B. (2023). *Mobile Vs Desktop Internet Usage*. Retrieved 03 01, 2023, from https://www.broadbandsearch.net/blog/mobile-desktop-internet-usage-statistics#post-navigation-0

Matthiesen, I. F. (2009, Januaray 1). *Overview on Techniques in Cluster Analysis*. Retrieved 03 2023, from https://link.springer.com/protocol/10.1007/978-1-60327-194-3_5

Lin, J. W. (2005, August 15). *Research on customer segmentation model by clustering*. Retrieved March 2023, from https://dl.acm.org/doi/abs/10.1145/1089551.1089610