

Spot the Hits!

Predicting The NEXT VIRAL song based on audio ATTRIBUTES: insights from the top songs on Spotify in 2024

Mariel Yap | Institute of Data - Capstone | August 2024

**Table of Contents**

[**Problem statement** 2](#_Toc174226331)

[Introduction 2](#_Toc174226332)

[Problem Investigation 2](#_Toc174226333)

[**Industry/domain** 3](#_Toc174226334)

[Value Chain 4](#_Toc174226335)

[Key Concepts 4](#_Toc174226336)

[**Stakeholders** 4](#_Toc174226337)

[**Business Question** 5](#_Toc174226338)

[Business Value 5](#_Toc174226339)

[Required Accuracy 5](#_Toc174226340)

[**Data Question** 6](#_Toc174226341)

[**Data** 6](#_Toc174226342)

[**Data Science Process** 7](#_Toc174226343)

[Data Analysis 7](#_Toc174226344)

[Data Wrangling Pipeline 7](#_Toc174226345)

[Exploratory Data Analysis (EDA) Highlights 8](#_Toc174226346)

[Reusability of Pipeline 15](#_Toc174226347)

[Intermediary Data Structures 15](#_Toc174226348)

[Modelling 15](#_Toc174226349)

[Outcomes 23](#_Toc174226350)

[Implementation 25](#_Toc174226351)

[**Data Answer** 25](#_Toc174226352)

[**Business Answer** 26](#_Toc174226353)

[**Response to Stakeholders** 26](#_Toc174226354)

[**End-to-end Solution** 27](#_Toc174226355)

[**References** 28](#_Toc174226356)

# **Problem statement**

## Introduction

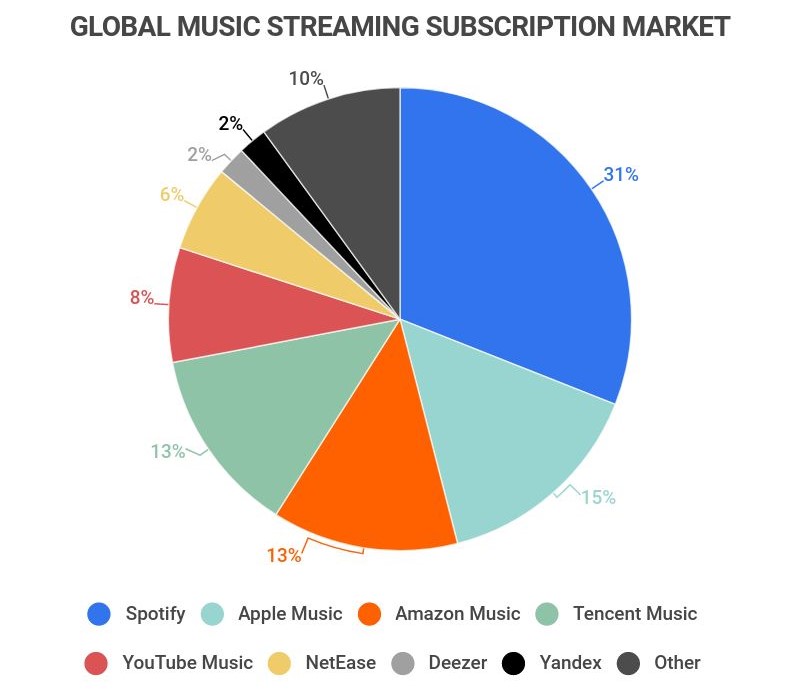
Music streaming is a competitive industry as users utilise these platforms to search for music they love and to access and engage with new and trending popular music. Understanding which elements contribute to a track’s popularity is essential for these music platforms as it impacts user engagement. Since the advent of music, the way that music has been accessed is constantly changing. In today’s music streaming landscape, services such as Spotify have made finding and accessing new music even easier. Total music streaming subscribers worldwide in the third quarter of 2023 totalled 713 million (Statista Research Department 2024) highlighting the prevalence of online music consumption and how catering to music platform users’ tastes is undoubtedly important. As of the end of 2023, Spotify had 602 million monthly users, which includes 236 million paid premium subscribers globally (Spotify 2024). While Spotify is the category leader in music streaming, consistent user engagement is crucial. The ability to predict track popularity can enhance recommendation algorithms, improve user satisfaction, and shape the strategic content that Spotify may include on its platform. This project will aim to develop a predictive model to estimate a song’s popularity on Spotify by analysing audio features and providing practical implications and suggestions for future studies.

Figure 1 - Pie chart of the global music streaming subscription market in 2022 (McCain 2022)

## Problem Investigation

This project will investigate the audio features contributing to a track’s popularity on Spotify. By analysing the metadata and audio features of songs, the goal is to create a predictive model that can project the popularity of a track. Accurately predicting a song’s popularity will enable Spotify to refine its recommendation algorithms and improve user experience.

This problem is valuable for Spotify to address for various reasons. Understanding the elements of what contributes to track popularity can lead to better recommendations and user engagement. Uncovering these insights can give Spotify a competitive advantage in the strategic campaigns it leads for its content. Additionally, improving user satisfaction can drive subscription growth and reduce user churn, subsequently leading to increased revenue. While Spotify’s premium subscribers and active users continue to grow, improved recommendation accuracy will continue to have a significant impact on Spotify’s market position.

Spotify currently uses complex algorithms for their recommendations which are continuously being improved. The popularity of a track is determined by an index based on user engagement such as total play counts and the recentness of the track. Current challenges leading to a necessity to continuously improve the accuracy of these algorithms include loss of user engagement due to subpar recommendations and content saturation. Spotify has reported that users consume 10% less music on their platform when they are continuously being shown songs that are not to their preference (Spotify Engineering 2023). There is also a gap in effective content delivery to the right users at the right time as on average 1.8 million songs are uploaded to Spotify every month (Shewale 2023). These factors can lead to lost engagement and an increase in churn rates if users are not finding content that appeals to them.

The desired state for Spotify is to develop a more accurate predictive model for song popularity, curate better editorial playlists, identify potential hits as early as possible and provide more personalised recommendations. Spotify needs to maintain a continuously appealing catalogue of music to increase user engagement and ultimately obtain greater subscription revenue.

Research into this area has been ongoing with multiple projects from external parties as well as Spotify’s R&D department also providing insights into these features. Previous studies have discovered that audio features from Spotify’s metadata such as danceability and valence can be meaningful predictors of a song’s popularity however there is still room to refine these predictive models.

# **Industry/domain**

The industry this project will be focusing on is the music streaming industry, with the major companies being platforms such as Spotify, Apple Music and Amazon Music. The music streaming industry allows users on their platform to listen to music digitally over the internet.

The music streaming industry is constantly growing and remains the dominant means of music consumption however it is facing many challenges. According to the Global Music Report by IFPI, subscription revenues grew by 11.2% with now streaming services taking up 67.3% of the global revenue (IFPI 2024). This emphasises the significance for musicians to upload their music to these platforms to be accessed. Challenges include streaming fraud and additionally, AI has also disrupted the industry as new challenges and opportunities for the music industry to adapt to issues, such as copyright infringements on generative AI.

## Value Chain

### Key Concepts

On-demand streaming is the key factor in the music streaming industry allowing users to play music of their selection digitally over the internet. This is usually supported with two different services; subscription services in which users pay a fee monthly to access the music without advertisements, and free access to music with advertisements. Users can also create playlists of their own as well as play editorial playlists generated by the platform. Revenue is distributed to artists and record labels in the form of royalties based on streaming data. Music streaming platforms also use user data analytics to provide recommendations and improve user experience.

This project is relevant to other industries that also rely on user data analytics, such as video streaming services, which use comparable algorithms to enhance content delivery. The diverging factor is that the metadata that will be analysed for audio tracks in this project will be different to the metadata in other industries.

# **Stakeholders**

Key stakeholders for this project will include Spotify’s R&D teams as they will use the model to refine the recommendation algorithms and improve user experience. Concurrently, Spotify’s management and product teams will leverage the results to enhance features on the platform as well as to make decisions on acquiring and promoting new content from artists. Marketing teams will also require the insights to understand popular trends to target their promotions effectively. Spotify users are the end beneficiaries who will experience an enhanced streaming platform.

It is important to address this problem as accurate predictions on song popularity can lead to improved recommendation algorithms leading to increased user engagement. This improves user loyalty and increases subscription rates which will contribute to Spotify’s strategic goals and directly impact revenue.

Expectations from stakeholders include accurate models that are accurate and scalable to be integrated into existing systems. Results should show an improvement in user engagement and reduced churn. The predictions from the model should be precise so that it can reliably drive campaigns and advertising strategies from marketing teams.

# **Business Question**

The main business question that will be answered is: **"To what extent do audio features of songs contribute to a track's popularity on Spotify, and can we predict the popularity of a track on Spotify using its audio features?"** This question will aim to utilise the audio metadata of songs to predict the potential popularity of tracks, which will contribute to improving Spotify’s recommendation systems and enhance user experience content strategies.

## Business Value

Answering this business question will hold significant business value for Spotify in multiple ways. Accurate predictions of song popularity can substantially improve recommendations which is critical for higher engagement and user retention. A 5% improvement in recommendations which leads to even a 1% increase in user retention would translate to 6.02 million active users. If each use retained generates $3 per month in revenue, this would entail an increase of $217 million in annual revenue. If marketing is also considered, streamlining the marketing budget by 10% due to less spending on promoting tracks less likely to succeed could equate to savings of $43 million annually. Additionally, providing insights to artist labels can help strengthen collaborations with a more diverse range of artists, anticipating that these partnerships can contribute an additional $10 million annually. Thus, the estimated total business value of this project is $270 million.

## Required Accuracy

The required accuracy for predicting track popularity should be 85% or higher, as this ensures that the majority of predictions are beneficial. For regression, this would entail that the predictive model aims for an R-squared (R²) value of 0.85 or higher and a Mean Absolute Error (MAE) represents less than 10% of the range. A high R² indicates that our model explains a substantial portion of the variance in popularity scores, while a low MAE ensures that our predictions are close to the actual values. This combination of metrics ensures that our predictions are both accurate and reliable, optimizing marketing efforts and maximizing promotional opportunities.

# **Data Question**

The data question to answer our business problem will be: “**To what extent do audio features of songs contribute to a track’s popularity on Spotify**, and can we predict the popularity of a track on Spotify using its audio features**?**” This will involve identifying which specific audio features are most strongly correlated with high popularity scores on Spotify. The features identified will then be used to develop the predictive model.

Data required to answer this question will be driven by the metadata found in tracks on Spotify. This will include the track name, artist name and popularity, which is on an index from 0 to 100 and is based on total streams and newness of a song. Audio variables will include danceability, valence (how positive a track is) and speechiness (presence of spoken words). These are a value between 0 and 1. Other features include the musical key of a song and tempo in beats per minute. Other contextual data that may be useful is the release date of the tracks and streaming data.

# **Data**

The data is sourced from Kaggle and the dataset is titled “Top Spotify Songs in 73 Countries (Daily Updated)” uploaded by user Asaniczka and can be downloaded at this link: <https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated/data>. The dataset captured daily the top 50 songs on Spotify for 70 countries and the data capture range is from 18 October 2023 to 29 July 2024.

The dataset has 1,029,803 records and 25 columns. The attributes of the data are:

* spotify\_id: The unique identifier for the song in the Spotify database.
* name: title of the song
* artists: Name of the artist(s) of the song
* daily\_rank: The daily rank of the song in the top 50 list.
* daily\_movement: The change in rankings compared to the previous day.
* weekly\_movement: The change in rankings compared to the previous week.
* Country: The ISO code of the country of the Top 50 Playlist. If Null, then the playlist if 'Global Top 50'.
* snapshot\_date: The date on which the data was collected from the Spotify API.
* popularity: A measure of the song's current popularity on Spotify.
* is\_explicit: Indicates whether the song contains explicit lyrics.
* duration\_ms: The duration of the song in milliseconds.
* album\_name: The title of the album the song belongs to.
* album\_release\_date: The release date of the album the song belongs to.
* tempo: Beats per minute, a measure of song tempo
* key: Key of the song
* mode: Mode of the song (major or minor)
* danceability: Percentage indicating how suitable the song is for dancing
* loudness: The overall loudness of the song in decibels.
* valence: Positivity of the song's musical content
* energy: Perceived energy level of the song
* acousticness: Amount of acoustic sound in the song
* instrumentalness: Amount of instrumental content in the song
* liveness: Presence of live performance elements
* speechiness: Amount of spoken words in the song
* time\_signature: The estimated overall time signature of the song.

The data is reliable as it is sourced from actual data from Spotify which use their metrics to track performance. The dataset itself has been aggregated by a third-party uploader on Kaggle through web scraping methodologies via Spotify’s API, and thus may have minor discrepancies which should be validated.

The quality of the raw data is high as it is derived directly from Spotify, which maintains robust data collection. As it was compiled by a third party, potential issues to monitor may include missing values, duplicates and inconsistent formatting.

The data was generated through Spotify’s tracking of top songs daily across various countries, capturing its rank and popularity. Spotify maintains metrics such as play counts and engagement to measure popularity scores and other attributes can be found in track metadata. These metrics can be scraped directly from Spotify’s API and compiled into datasets such as those found on Kaggle.

This specific dataset is updated daily, allowing for continuous access and retraining with new data and can be retrieved automatically through Kaggle’s API. Likewise, as Spotify continually updates its own data, access to similar chart data can be retrieved through Spotify’s API.

# **Data Science Process**

## Data Analysis

### Data Wrangling Pipeline

Data ingestion involves ensuring the data is structured in a way that can be readily used. For the project, the dataset was structured into a pandas DataFrame. Subsequently, the data cleaning process involved handling missing values, converting columns to appropriate data types and handling duplicates. Data enriching involves techniques such as feature engineering to enhance the existing data.

### Exploratory Data Analysis (EDA) Highlights

#### Cleaning

The data was uploaded as a Dataframe into Jupyter Notebook for ease of processing. Majority of the column data types were in line with expectations except for snapshot\_date and album\_release\_data which was converted to date-time object. Mode was also converted from integer to Boolean type. Minimum and maximum figures were inspected through the describe function and were inside expectations. It is noted that popularity is on a scale of 1 to 100, audio features are on a scale of 0 to 1 and that key are integers from 1 to 11, which is not readable as musical keys. Duration was converted from milliseconds to seconds.

Records before 1 January 2024 were dropped as the project focuses on only 2024 data. There were missing values present in name, artists, country, album\_name and album\_release\_date columns. Records in name, artist and album\_release\_date were dropped as they would not be able to be identified. The entire column for album\_name was dropped as the information was not needed. As per the data dictionary, null values in country indicate the Global Top 50 chart. As such null values were imputed with ‘Global’.

Duplicates were a major aspect to deal with as 98.61% of the records were duplicated songs. Because the data had been aggregated over multiple periods, many songs can chart on many days. For the purposes of EDA the duplicates were retained to understand temporal trends. However, for the purposes of building a predictive model, song duplicates can skew the data and lead to biased predictions if songs chart more frequently than others. Aggregating the data based on calculating the rolling average popularity over a 7-day period and then taking this mean across all days and countries mitigated this so the popularity index can be made the focus. Rolling average popularity is chosen so that we can assess a song's actual popularity over time i.e. no short term one hit wonders, with a focus on capturing recent trends.

#### Analysis

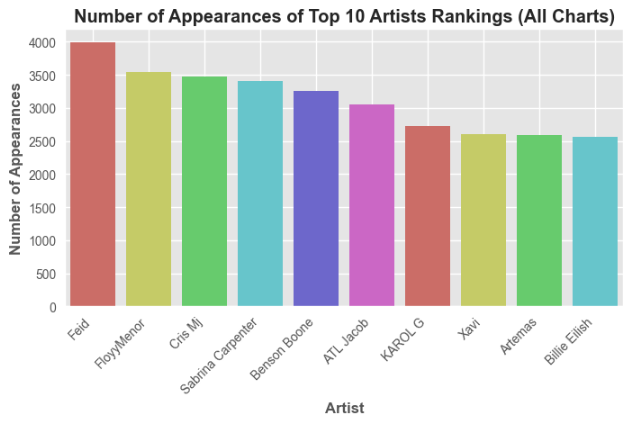
Some of the initial analysis was that the highest global ranking song currently is Who by Jimin. The song has a popularity score of 92 and leans on the higher end of the scale for danceability, energy and valence.

Figure 2 - Which artists have the most appearances in the top rankings?

When asking the question above there are some artists not as prevalent in the west. As the chart data has included data from 70 countries, it is a given that there would be artists more prevalent in other parts of the world.

A screenshot of a computer screen

Description automatically generatedA colorful pie chart with numbers and letters

Description automatically generatedFeid has the most appearances in the Top 50 chats globally however most of the placements come from South American Spotify charts in 2024 so far. Out of the 4000 times they placed on Spotify charts in 2024 so far, only 0.29% of these were on the Global Spotify Top 50 chart, revealing that their presence may not be that global currently.

Figure 3 - What is the overall popularity of Feid's songs?

Figure 4 - Which countries are listening to the most charted artist so far on Spotify in 2024, Feid?

The mean popularity for each of Feid’s songs placed on Top 50 charts do not place higher than 85 on the popularity index.

A chart with colorful lines

Description automatically generated with medium confidence

Figure 5 - What is the distribution of song popularity across different countries and the global chart?

As expected for the global top 50 chart, the distribution of the popularity index sits around the 85 to 95 mark. Comparing this to other countries with the most spread in popularity such as Lithuania (LT), Iceland (IS), and Bulgaria (BG), having a spread amongst lower popularity scores could indicate that the most popular songs in those countries are not that popular globally on Spotify. This coincides with the analysis above that while Feid is the artist with the most placements on the Spotify charts globally, because they are not listened to on a global scale, even their most popular songs would not be within the ranges of the popularity index on the Global chart.

Coinciding with the box plot, this time series plot shows that the average popularity of songs on the Global Top 50 sits around the 85 to 95 popularity index. However, there are some points where the graph dips, even reaching as low as an average of 35 popularity. From Spotify's perspective this can be concerning if this is due to Spotify's recommendation algorithms and which songs are being promoted. Other reasons could be that fewer major artists released music during this time or a temporary shift in listener preferences.A graph with green lines

Description automatically generated The reason for the dip was due to Taylor Swift whose newest release charted within the Spotify Global Top 50 as soon as it was released. On its immediate release its popularity index had not yet been calculated which is why popularity is 0, however the entire album already ranked within the Global Top 50 due to the sheer number of listeners. On paper would seem contradictory since popularity and the Global Top 50 are linked, but the achievement is something only Taylor Swift can pull off!

Figure 6 - Do the popularity of songs on the global chart fluctuate over time?

#### Distribution

A green bar graph with numbers

Description automatically generated with medium confidence

A graph of different types of data

Description automatically generated with medium confidenceSome key takeaways from the distribution are that the songs in the dataset tend to lean towards higher danceability, loudness and energy. There is an even split between songs with major and minor modes. They also tend to have less acousticness and instrumentalness. Valence has a slightly negatively skewed distribution indicating that songs on the Top 50 lean more slightly happy. Our target variable popularity also leans negatively skewed.

Figure 7 – Distributions of features

#### Outliers

­A screenshot of a graph

Description automatically generated

Figure 8 – Outliers in features

Some variables are highly skewed even though they are not errors and represent the true variability of the data. To ensure more accuracy and reliability of the modelling and make the features more representative of the population the outliers were dealt with. Features with moderate to high skewness were transformed using box-cox to make the data as close to normally distributed as possible. The target variable was also transformed to help with non-linear relationships with predictors. Remaining outliers three standard deviations away were removed using z-score method. For the heavily skewed instrumentalness feature winstorisation was used to limit values by transforming extreme values to their closest value.­

### Reusability of Pipeline

While the pipeline is reusable in the case of moderate size datasets such as the Top 50 charts used, if the pipeline were to be applied to a greater number of songs on Spotify’s database it should be scaled. The pipeline in the notebook has been broken down into functions so that it can be adapted with additional steps if needed. Use of automation tools such as Apache Airflow may help to automate the pipeline for regular updates for new song data.

### Intermediary Data Structures

Data structures used include intermediary Dataframes to store the data after it has been cleaned and processed. During the EDA process visualisation objects such as box charts and plots were utilised to understand patterns in the data. Feature matrices noted as X and Z to represent the two scenario sets were used to feed into the machine leaning algorithms.

## Modelling

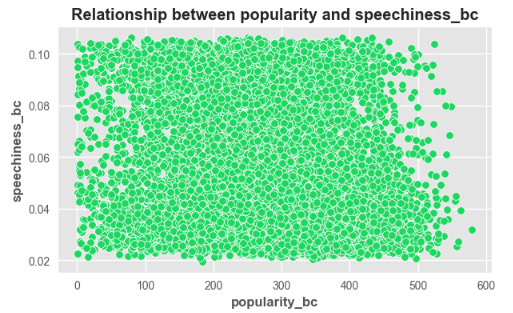
#### Correlation Analysis

A graph with numbers and letters

Description automatically generated with medium confidence

Figure 9 - Correlation analysis of numerical features

Through the correlation analysis above, there is no clear linear relationship between any of the dependent features and the target feature popularity. This indicates that a popular track can have many various attributes and differ from one another. Because of this, there will be a focus on non-linear models and feature engineering. Some of the dependent features do demonstrate collinearity such as energy and loudness, and acousticness and energy.

A green dot diagram with white text

Description automatically generatedThe scatter plots support the details in the correlation analysis as even the highest correlated variables speechiness and loudness are highly varied and do not reveal a clear linear relationship with the target variable popularity.

Figure 10 - Relationship between target variable popularity and feature loudness

Figure 11 - Relationship between target variable popularity and feature speechiness

#### Feature Engineering

Feature engineering is a focus as the current features do not synergise enough with the target variable. New interaction features were created by multiplying features with high collinearity or reveal more deeper patterns in relationships. Interaction features created included energy and danceability, loudness and tempo, and speechiness and danceability.

Some features were also categorised so that the models can handle them non-linearally. As identified earlier, musical keys were being denoted by integers. These were mapped to actual musical keys and combined with mode to show the specific musical tone of the song. Other features categorised include song duration, valence and tempo.

Other non-music features were also extracted for more comprehensive interactions. Temporal features such as day of week and release year were extracted. Aggregated metrics for cumulative and mean artist popularity were also utilised to see the impact of an artist’s popularity on a song’s popularity. All categorical features were encoded with one-hot encoding. Following the feature engineering the variables in the dataset total 58. The data was also standardised to have a mean of 0 and a standard deviation of 1, allowing the features to be modelled on the same scale.

#### Inputs and Output

For the main question, the output label is popularity. It is hypothesised that an artist's popularity may have a large impact on a song's popularity. Because of this, two scenario sets will be modelled to see its impact. X feature matrix (Model 1) includes a focus on audio features and Z feature matrix (Model 2) will include all features including mean\_artist\_popularity and cumulative\_popularity.

#### Using PyCaret to Establish a Baseline

Using Pycaret allows for the comparison of different models as a starting point for analysis. Different feature selection combinations were experimented with to see the effect on regression metrics. The top three models were also chosen for modelling.

A screenshot of a computer

Description automatically generated

Figure 12 - Pycaret output for Model 1

After running through different ratios of features to select, the best ratio to maximise r2 for the audio feature scenario set (Model 1) were all the features i.e. no feature selection used. The top three models outputted for this scenario are outlined in the figure. As hypothesised, non-linear and ensemble models initially perform the best for this set.

A screenshot of a graph

Description automatically generated

Figure 13 - Pycaret output for Model 2

For the second scenario set 2, the best ratio of features that maximise R2 is 0.5, or half of the features being used. As expected, including artist popularity as a feature significantly impacts a song’s popularity, which aligns with real-world scenarios. The three models selected to be focused on in modelling are Gradient Boosting Regressor (GBR), Lasso Regression and Linear Regression.

#### Feature Selection

A graph of different colored bars

Description automatically generatedTwo feature selection techniques were utilised for comparison. Lasso regression was first used, which adds a penalty shrinking to the least important features forcing them to have coefficients of zero, effectively removing them from the model.

Figure 14 - Feature importance for Model 1 using Lasso Regression

For Model 1 19 features were determined as significant with Lasso Regression. The most significant audio feature identified is the loudness of the songs.

A graph with different colored bars

Description automatically generatedFor Model 2 it is clear with Lasso Regression that mean\_artist\_popularity is the most significant determiner in predicting popularity however, audio features are still identified as contributors. 24 features are identified as significant.

Figure 15 - Feature importance for Model 2 using Lasso Regression

Recursive feature elimination (RFE) was then used as a comparison. Due to the computation power needed to test multiple features. Using knowledge from Lasso regression and Pycaret analysis, the top 20 features were selected for RFE.

Following this, common features were identified between the two feature selection techniques for a cross-section of the most optimal features. This produced 4 feature selection sets for each scenario model (one with all features, one from Lasso, one from RFE, and common features shared with Lasso and RFE – 8 total) to be tested.

#### Model Used

##### Training with Default Model Settings

A screenshot of a computer

Description automatically generatedFeature matrices X and Z were split into a train-test split to ensure the robustness of the modelling. The initial modelling round was tested on each model’s default settings and with each feature selection set to analyse which features contribute most to the R2 score.

Figure 16 - Initial modelling with feature selection sets for Model 1

Similar to the findings from the Pycaret analysis, the best R2 score for Model 1 is when utilising all 52 features. Since this was confirmed in both analyses, modelling will continue using all the features identified.

A screenshot of a computer

Description automatically generatedFor Model 2 feature set 4 (common features from Lasso and RFE) performed the best for R2. It is noted that for Lasso Regression, the feature sets had no impact on its score.

Figure 17 - Initial modelling with feature selection sets for Model 2

##### PCA Analysis

A screenshot of a computer

Description automatically generatedPrincipal component analysis (PCA) was then tested to reduce the dimensionality of Model 1 due to the sheer number of features included. However, this did not improve the R2 score so the original X\_train was retained for modelling.

Figure 18 - PCA analysis for Model 1

##### Hyperparameter Tuning

Hyperparameters were then tuned for each of the top models to see if metrics could be optimised. Due to processing time, only three hyperparameters were chosen to be tuned for each model.

A screenshot of a computer

Description automatically generatedEnsemble methods were then added to the training process using the tuned hyperparameters for each model. Bootstrap aggregating (bagging) was applied to each of the base regressors to see its impact. Bagging involves splitting the data to create multiple of the same base regressors on different training data computed parallelly. The final model trained was stacking, which involved combining the top 3 models identified including a final estimator.

Figure 19 – Hypeparameter-tuned models for Model 1

A screenshot of a computer

Description automatically generatedFor Model 1 the model which achieved the best metrics was stacking, with an R2 score of 0.1739 and a MAE of 78.8306. This ensemble model is what is chosen to be tested on unseen data.

Figure 20 - Hyperparameter tuned models for Model 2

For Model 2, the best model performance on training data was Bagging Regressor (Gradient Boosting). This ensemble model is what is chosen to be tested on unseen data.

#### Training Time

Training time varied for each model with the fastest being Linear Regression and Lasso. The ensemble methods and Random Forest took longer to train due to the complexity of the models. Bagging and stacking times also varied depending on the estimators used. Estimators with ensembles took the longest time to train, ranging up to 10 minutes.

#### Model Performance Metrics

Key metrics included in the training process included R2 and Mean Absolute Error (MAE). R2 explains the amount of variance in target variable popularity explained by the model. MAE is a value showing the average of absolute errors between predicted and actual popularity values. As the target variable was transformed during the outlier process, this shifted the popularity index from what used to be 0 to 100, to a minimum of 0.6894 to a maximum of 579.22. This is the range that will be considered when assessing the MAE.

#### Final Model Testing

The final model selected for Model 1, which is the audio feature focus set was stacking, with an R2 score of 0.1991 and a MAE of 78.34. This is an improvement slightly from the training set now explaining about 20% of the variability in track popularity. For the MAE the predicted values now lie 13.53% away from true popularity values. While this is not close to the target metrics, it still identified that audio features play some role in predicting track popularity.

For Model 2, which is the set with artist popularity included, the selected model was Bagging Regressor (Gradient Boosting) with an R2 score of 0.8651 and an MAE of 22.84. This is 3.94% away from true popularity values, which is a great result.

## Outcomes

The results from the training and test results from Model 1 with the audio features focus display that audio features alone only account for a small portion of the variance in song popularity and that in isolation is insufficient to reliably predict song popularity.

A graph showing a green and blue line

Description automatically generatedThis is supported by the plot above which shows a noticeable difference between the prediction line, which is the final model used, and the line of best fit, which would be the optimal model for the data. This entails that even with extensive modelling, either different features or additional data is needed to optimise the model.

Figure 21 - Actual vs predicted popularity values for Model 1

A graph showing a number of different colored bars

Description automatically generated with medium confidence

Figure 22 - Feature importances for Model 1

The most significant features identified in this model are loudness, duration, acousticness, release day, speechiness and energy. While these features are important, they are not sufficient in capturing overall variability in song popularity. Nevertheless, it would be good to monitor these categories in identifying popular songs.

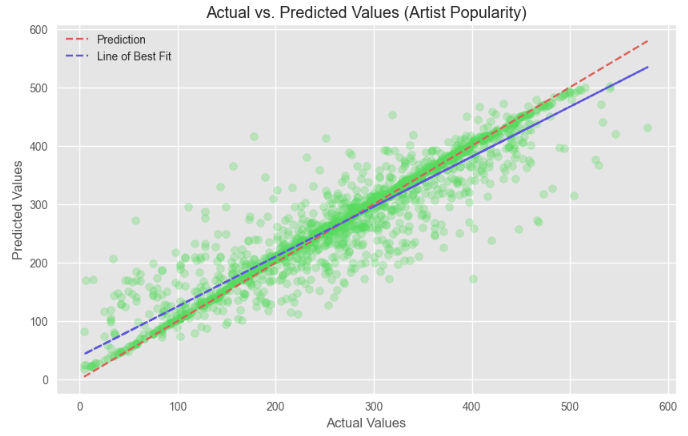
With Model 2, including artist popularity features dramatically improves the model’s predictive accuracy, reflecting the importance of an artist’s existing fanbase on a song’s success.

Figure 23 - Actual vs predicted popularity values for Model 2

The plot shows that the model’s line is much closer to the line of best fit, with most of the data points lying on the predicted line.

A graph with a red line

Description automatically generated

Figure 24 - Feature importances for Model 2

The feature importances show the enormous impact of an artist’s popularity and the influence of the audio features in comparison. This indicates that a well-known artist will be more likely to produce popular tracks, regardless of the specific audio attributes in the song.

The use of ensemble methods as the final models chosen indicates how the techniques can provide a valuable boost in accuracy while also improving robustness.

## Implementation

The top considerations for implementing the models in deployment are scalability, monitoring and data management. The models should be optimised to be able to handle large volumes of data Spotify maintains in real time and integrate seamlessly with Spotify’s existing infrastructure such as its recommendation systems. This may involve using distributed computing frameworks like Apache Spark. The model’s performance should be continuously tracked over time as new data becomes available against model metrics and business KPIs. Automatic retraining on pipelines will allow models to stay updated.

# **Data Answer**

The data question was answered satisfactorily as the goal was to evaluate the extent to which audio features can predict song popularity. While it was identified that audio features in isolation cannot do this alone, it was discovered the artist's popularity is the main determiner, as demonstrated through the model’s metrics exceeding the project goal metrics. The models were successfully developed and tested to address the data question.

The confidence level in the answer is high as supported by strong R2 scores of 0.8652 validated on an unseen test set. The small variance between the training and test sets shows that the models are not overfit, but are consistent, reinforcing the reliability of the results.

Some restrictions that affect the confidence level are the data limitations due to the chosen dataset which may not represent the full diversity of songs on Spotify. A segregated analysis for established versus lesser-known artists can be modelled to see if audio features are more impactful in that space. Additionally, as music trends change over time models should be constantly retrained to maintain accuracy.

# **Business Answer**

The business question was answered satisfactorily as specified earlier; the question addressed was the extent to which audio features can predict tracks on Spotify. While they provide some predictive power, including artist popularity metrics improved the model’s accuracy. This insight is valuable as it suggests that a combination of majority artist-based and supplementary content-based content may be more successful on platforms.

The confidence level in the business answer is high as shown by robust model performance and validation on test sets, showing the model’s ability to generalise to new data. The findings have clear business insights to stakeholders such as the importance of utilising artist popularity in recommendation algorithms. As identified previously, external factors such as social media and changes in music trends would make consistent model updates necessary to maintain the confidence level.

# **Response to Stakeholders**

The overall message to stakeholders would be that artist popularity is a key driver of song success with established artists having a higher chance of releasing viral songs. Stakeholders should consider utilising resources on artists with greater popularity to maximise chart success. Artist popularity should be a key feature in refining recommendation systems as this will likely enhance user engagement.

While audio features such as danceability and valence alone are less effective in isolation, combining a comprehensive approach with both features and artist metrics is better for strategic decision-making. Exploring how audio trends interact with artist popularity could be a recommended approach. Stakeholders should also keep an eye on emerging trends in audio features to maintain a competitive edge. Marketing teams can also tailor promotions emphasising both popular artists and a song’s unique audio features. It can also be insightful to uncover new and upcoming artists.

As music trends continuously develop, models should be consistently updated with new data to monitor performance and refine models.

# **End-to-end Solution**

# **References**

<https://matplotlib.org/stable/gallery/style_sheets/style_sheets_reference.html>

<https://seaborn.pydata.org/tutorial/color_palettes.html>

<https://pycaret.readthedocs.io/en/stable/api/regression.html#pycaret.regression.setup>

<https://developer.spotify.com/documentation/web-api/reference/get-audio-features>

<https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated/data>

IFPI. (2024). *IFPI global music report (2024 edition)*. <https://ifpi-website-cms.s3.eu-west-2.amazonaws.com/IFPI_GMR_2024_State_of_the_Industry_db92a1c9c1.pdf>

McCain, A. (2022, May 23). *29 Music Industry Statistics [2022]: Facts, Trends, And Sales – Zippia*. Zippia. <https://www.zippia.com/advice/music-industry-statistics/>

Shewale, R. (2023, December 15). *Spotify Stats For 2024 (Subscribers, Revenue & Trends)*. Demands. <https://www.demandsage.com/spotify-stats/#:~:text=As%20we%20enter%20into%202024%2C%20Spotify%20has%20574%20million%20monthly>

Spotify. (2024). *Q4 2023 Update*. <https://s29.q4cdn.com/175625835/files/doc_financials/2023/q4/Shareholder-Deck-Q4-2023-FINAL.pdf>

Spotify Engineering. (2023, October 20). *Exclude from Your Taste Profile*. Spotify Engineering. <https://engineering.atspotify.com/2023/10/exclude-from-your-taste-profile/>

Statista Research Department. (2024, May 29). *Global streaming music subscribers 2023*. Statista. <https://www.statista.com/statistics/669113/number-music-streaming-subscribers/#:~:text=In%20the%20third%20quarter%20of%202023%2C%20the%20number>