# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | *YouTube Music Video Views Prediction* |
| **Requestor** | *Bilal Hasanzadah* |
| **Date of Request** | *August 2024* |
| **Target Quarter for Delivery** | *Q4* |
| **Epic Link(s)** | [*https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube*](https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube)  [*https://musicindustryblog.wordpress.com/2019/01/03/how-youtubes-domination-of-streaming-clips-the-markets-wings/*](https://musicindustryblog.wordpress.com/2019/01/03/how-youtubes-domination-of-streaming-clips-the-markets-wings/)  [*https://marketsplash.com/youtube-statistics/*](https://marketsplash.com/youtube-statistics/)  [*https://www.thinkwithgoogle.com/marketing-strategies/video/music-videos/*](https://www.thinkwithgoogle.com/marketing-strategies/video/music-videos/)  [*https://www.forbes.com/sites/billrosenblatt/2024/03/30/music-industry-annual-reports-show-stable-growth-and-youtube-strength/*](https://www.forbes.com/sites/billrosenblatt/2024/03/30/music-industry-annual-reports-show-stable-growth-and-youtube-strength/)  [*https://pictory.ai/blog/what-are-the-most-popular-genres-on-youtube-in-2023?el=0035&htrafficsource=pictoryblog&hcategory=video*](https://pictory.ai/blog/what-are-the-most-popular-genres-on-youtube-in-2023?el=0035&htrafficsource=pictoryblog&hcategory=video)  [*https://thumbnailtest.com/stats/youtube-music/*](https://thumbnailtest.com/stats/youtube-music/) |
| **Business Impact** | *The predictive model will be able to recommend artists and label companies to select the music tracks for which to make videos for YouTube. This will tend to bring in increase popularity and revenue generation for the Artists and Labels through YouTube.* |

## Business Opportunity Brief

**Business Statement:**

The goal is to bring in views and increase popularity of Artists on YouTube through analyzing. The prediction will be targeted at selecting which tracks to make video for.

**The specific ask:**

Perform analysis on the dataset to investigate and summarize its main characteristics, create data visualization to find insights. Do feature engineering and model building for prediction whether a video should be produced or not.

**Business Problem Statement**:

How can artists and sound engineers optimize music quality and reach, while managers and record labels identify tracks most likely to attract the highest views on YouTube? By analyzing song features, we aim to predict and enhance the success of tracks, guiding strategic decisions for music video production.

## Supporting Insights

* **Market Share and Revenue**: There has been a stable rise in YouTube Music market share from 8% in 2021 to 8.9% in 2023 with predictive rise to cross 9% in January 2024. The revenue has also reached $1.1 billion dollars in February 2024.
* **Trends**: Within the age group of Americans who are 35 to 55 and up say they visit YouTube for music related content at least once a week. And those people who watch music on YouTube, 68% watch it when they feel nostalgic. Baby Shark Dance video acquired more than 12.3 billion views in 2023.
* **Competitor:** The key competitors of YouTube Music are Spotify, Apple, Amazon Music and Pandora. YouTube has heavy competition from them in streaming but when it comes to views they lack behind as the other platforms represent themselves for just streaming and not viewing music videos yet. Even though Google started YouTube Music in 2018 it has become a more popular choice than Apple Music and Amazon Music.

## 1.2 Project Gains

1. Revenue Gains:

* Targeted Marketing and Promotions: By analyzing audience engagement metrics like views, streams, liked, and comments, the company can identify the most promising tracks and artists.
* Optimized Release Strategies: With the various attributes like danceability, energy, speechiness, loudness etc. correlate with higher engagement can guide the development of a potential music video leading to increasing sales and streaming numbers.

1. Quality Improvements:

* Content Curation: Taking leverage of the insights on the KPI like danceability, energy, and valence, the company can curate playlists and recommend content more effectively, enhancing the user experience and keeping listeners engaged.
* Artist Development: Providing artists with data-driven insights into their music’s performance can help them refine their style and production, leading to higher quality music that resonates better with the audience.

1. Cost and Time Savings:

* Efficient Resource Allocation: data analysis can help the company allocate promotional budgets more efficiently by focusing on the most promising tracks and artists. This reduces wasted spending on less effective campaigns.
* Automation Of Reporting: Implementing automated systems for tracking and analyzing performance metrics can save time and reduce the manual effort required for data collection and reporting.

1. Implications of Inaction:

* Missed Opportunities: Without leveraging data analytics, the company risks missing out on key trends and opportunities to capitalize on successful tracks or emerging artists. This could result in lost revenue and market share.
* Competitive Advantage: Competitors who use data-driven strategies will have an advantage in attracting and retaining audiences. Inaction could lead to a decline in the company’s relevance and influence in the market.

## 2.0 Analytics Objective

Key questions:

1. Audience engagement and Growth:

* What factors like the song attributes most strongly influence the number of views and streams a song receives on YouTube?
* How do engagement metrics, for example the likes, comments, and shares correlate with overall song popularity and audience growth?

1. Content Performance:
   * Which specific attribute, for example the danceability, energy and valence are associated with higher engagement and popularity of songs?
   * Are there differences in performance metrics across different types of releases (albums vs singles)?
2. Revenue Optimization

* How can the company maximize revenue from streaming, ads, and other channels based on the analysis of song performance and audience demographics?
* What are the optimal pricing and marketing strategies for maximizing revenue from digital platforms?

1. Market and Audience Segmentation

* What are the key psychographic characteristics of the audience for different music genres and artists?
* How can the company effectively segment its market to tailor marketing strategies and product offering?

1. Predictive Insights

* What early indicators suggest a song or artist is likely to become popular?

Assumptions:

* Data Availability and Accuracy: This dataset is assumed to be accurate and representative of actual user engagement and preferences. It includes all necessary data points such as views, streams, liked, and song attributes.
* Consistency in Data Collection: The data has been consistently collected across different platforms (Spotify and YouTube).
* Generalizability: The findings from this analysis are assumed to be generalizable to broader market segments and not limited to the specific dataset.
* Platform Independence: While there may be platform-specific trends, it’s assumed that key insights and recommendations can be applied across different music streaming and video platforms.

Hypotheses:

1. H1: Song Attributes Influence Popularity

* Hypothesis: Song with higher danceability, energy, and valence scores are more likely to achieve higher views and streams.
* Test: Analyze the correlation between these attributes and engagement metrics.

1. H2: Engagement Metrics Predict

* Hypothesis: Engagement metrics (likes, comments) are positively correlated with attributes such as danceability and energy.
* Reasoning: Tracks that are more danceable or energetic may be engaging, prompting listeners to like or comment.

1. H4: Optimized Content and Marketing Strategies Enhance Engagement

* Hypothesis: Utilizing insights from track attributes and engagement metrics can help content creators and marketers optimize content strategy and marketing efforts, leading to increased engagement.
* Reasoning: Data-driven strategies can better align content and marketing efforts with audience preferences.

## 2.1 Other related questions and Assumptions:

Other Related Questions:

1. How do different track attributes (for example loudness, tempo, valence) correlate with each other?

* Purpose: Understanding these relationships can help in identifying patterns and potential combinations that drive engagement.

1. What is the impact of the artist’s popularity (if measurable from views or streams) on the success of new tracks?

* Purpose: To determine if already popular artists consistently perform better and how this can be accounted for in analyses.

1. Are there significant differences in engagement metrics (views, liked, comments) across different tracks lengths?

* Purpose: To explore if longer or shorter tracks have a measurable impact on user engagement and popularity.

1. What role does the presence of featured artists or collaborations play in track popularity?

* Purpose: To assess whether collaborations increase a track’s reach and engagement.

Assumptions:

1. Data Completeness and Quality:

* Assumption: The data dataset includes all necessary and accurate track attributes and engagement metrics. Missing data or inaccuracies could lead to biased or incomplete analysis.

1. Homogeneity of Platforms:

* Assumption: Engagement metrics from different platforms (Spotify and YouTube) are comparable. Differences in platform user base, behavior, and engagement measurement might affect the comparability.

1. Stable Trends:

* Assumption: Trends in music consumption and preferences have remained stable during the period covered by the data. Rapid changes in trend could affect the validity of the finding.

1. Representativeness:

* Assumption: The dataset is representative of a broader population of music tracks and artists. If the dataset is skewed towards certain artists it may not generalize well.

1. Constant External Factors:

* Assumption: External factors, such as changes in platforms algorithms, marketing efforts, or global events, are constant or have minimal impact during the analysis period. These factors can significantly influence engagement metrics.

1. Uniformity in Engagement Metrics:

* Assumption: The way engagement metrics like views, likes, and comments are measured and reported is uniform and reliable. Any inconsistencies could affect the analysis outcomes.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  |  |
|  | Key Performance Metrics:   1. Track Popularity Metrics:  * Total Views (YouTube): * Definition: The total number of times a track has been viewed on YouTube. * Success Indicator: A significant increase in views or streams after implementing insights from the analysis.  1. Predictive Model Performance:  * Prediction Accuracy: * Definition: The accuracy of predictive models in forecasting tracks popularity based on early engagement metrics and track attributes. * Success Indicator: High accuracy (measured by metrics like RMSE, MAE, or R square) in predicting track popularity. * Model Robustness: * Definition: The model’s ability to generalize across different tracks, artists. * Success Indicator: Consistent performance across various subsets of the data. * Comparing the performance of Linear Regression, Decision Trees, Random Forest, and Neural Networks.  1. Content Strategy Effectiveness:  * Playlist Placement and Performance: * Definition: The performance of tracks placed in curated playlists, measured by subsequent engagement metrics. * Success Indicator: Tracks placed based on data-driven insights active higher engagement compared to those placed without such insights.   Driver and Key Metrics:   1. Data Quality and Completeness:  * Ensuring high-quality, comprehensive data to support accurate analysis and modeling.  1. Model Validity and Interpretability:  * Developing models that are not only accurate but also interpretable, allowing stakeholders to understand and trust the insights.  1. Alignment with Business Objectives:  * Ensuring the analysis and resulting strategies align with overarching business goals, such as increasing market share, enhancing user experience, and maximizing revenue.  1. Stakeholder Engagement:  * Continuous communication with key stakeholders to ensure that the analysis addresses their needs and that the results are actionable.  1. Iterative Refinement:  * Regularly refining the models and strategies based on feedback and new data to ensure continuous improvement. |
|  |  |

## 2.3 Methodology and Approach

Type of Analysis

1. Descriptive Analysis:

* To summarize and describe the key characteristics of the dataset, including track attributes and engagement metrics.
* Techniques: Summary statistics, data visualization

1. Predictive Modeling:

* To develop models that predict the popularity and engagement of music tracks based on their attributes and early metrics.
* Techniques: Linear regression, Forward Regression, Backward Regression, Stepwise Regression and potentially other machine learning methods.

1. Correlation and Causal Analysis:

* To identify and verify relationships between track attributes and engagement metrics.
* Techniques: Pearson correlation, Chi-square test, and regression analysis to explore relationships and causality.

Methodology:

1. Data Preparation and Exploration:

* Data Collection and Cleaning:
* Collect data from Spotify and YouTube here acquired from Kaggle, ensuring that track attributes and engagement metrics are accurately captured.
* Handle missing values, outliers, and ensure data quality by standardizing and normalising where necessary.
* Feature Engineering:
* Create new features if required, such as interaction between track attributes.
* Convert categorical variables, if any, into numerical formats using encoding techniques like one-hot encoding.
* Initial Descriptive Analysis:
* Generate summary statistics and visualizations to understand the distribution and relationship between variables.

1. Exploratory Analysis and Hypothesis Testing:

* Exploratory Data Analysis (EDA):
* Conduct EDA to identify trends, patterns and anomalies in the data.
* Visualize relationships between track attributes and engagement metrics using scatter plots, histograms, and heatmaps.
* Hypothesis Testing:
* Test the hypotheses outlined in the analytics objectives using statistical methods.
* Use linear regression to test the impact of danceability on views count.

1. Predictive Modeling:

* Model Building:
* Start with a Linear Regression model to identify key variables that influence track popularity and engagement.
* Explore additional models such as advanced methods such as Random Forest and Neural networks, to identify key variables and capture complex non-linear relationships.
* Model Evaluation:
* Evaluate models using metrics such as accuracy, precision, mean absolute error, and root mean square error.
* Use cross-validation to ensure the robustness of the models and avoid overfitting.

1. Validation and Refinement:

* Validation:
* Validate the models using a separate test dataset or through cross-validation.
* Ensure that the models generalize well to new data and are not overfitting the training data.
* Refinement:
* Refine models based on validation results, potentially iterating on feature selection, hyperparameter tuning, and algorithm choice.

1. Strategic Recommendations:

* Insights and Recommendations:
* Provide insights based on the analysis, such as which track attributes are most predictive of popularity and engagement.
* Offer strategic recommendations for content curation, marketing, and promotional strategies based on the findings.

1. Reporting and Presentation:

* Final Report:
* Compile a comprehensive report detailing the methodology, analysis, findings, and recommendations.
* Include visualizations, model outputs, and interpretation of results.
* Prepare a presentation for stakeholders summarizing the key insights and recommendations, focusing on actionable outcomes and strategic implications.

## 3.0 Population, Variable Selection, considerations

**Audience/population selection:**

* The population consists of music tracks from various artists available on Spotify and YouTube music.
* The analysis focuses on tracks that have measurable engagement metrics (views, likes, comments) and relevant audio features.

**Observation window:**

* The timeframe for data collection spans from early 2000s to 2010s. This period is chosen to capture sufficient data for robust analysis while accounting for trends over time.
* The observation window also considers the availability of data and the typical lifecycle of track popularity.

**Inclusions:**

* Tracks with complete data for key engagement metrics (views, likes, comments) and audio features ( danceability, energy, valence).
* Tracks available on both Spotify and YouTube music to ensure consistency in analysis.

**Exclusions:**

* Non-musical content (for example podcasts, spoken word) that might be present in the dataset.

**Data Sources:**

* Primary data sources:
* Spotify: Provides detailed audio features, track metadata, and streaming metrics.
* YouTube music: Provides engagement metrics such as views, likes, and comments.

**Audience Level:**

* Track Level: The analysis is conducted at the track level, focusing on individual music tracks rather than aggregated data at the artists or album level.

**Variable Selection:**

* Dependent Variable:
* Audio Features:
* Danceability: How suitable a track is for dancing
* Energy: A measure of intensity and activity.
* Valence: The positivity of the track.
* Tempo: The speed or pace of the music.
* Loudness: The overall volume level.
* Track metadata:
* Track duration: Length of the track in milliseconds.
* Release date (to be extracted) : Date when the track was released

**Derived Variables:**

* Engagement Ratios:
* Like-to-view ratio: Likes divided by views, indicating engagement relative to reach.
* Comment-to-view ratio: Comments divided by views, indicating deeper engagement.
* Popularity Index:
* A composite score that combines views, likes, comments, and streams to provide an overall measure of a track’s popularity.

**Assumptions and data limitations:**

* Data availability and completeness: Assumes that the available data accurately reflects the track’s engagement and audio features. Incomplete data may lead to biased conclusions.
* Consistency across platforms: Assumes that engagement metrics are comparable across Spotify and YouTube music, despite potential differences in user behavior and platform algorithms.
* Data quality: Assumes that the data from primary sources are reliable and has been consistently collected.

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| *Churn rate being inflated by counting multiple contracts from the same rooftop as individual observations.* | *Low* |  | *Once analysis begins, we can quantify the inflation. However, this approach allows us to compare how the same dealer performed across different contracts and find useful patterns.* |
| Data quality issues | Medium | Medium | This project may encounter issues with the incomplete or inaccurate data, leading to biased results and unreliable insights. |
| External factors influencing engagement | Medium | Low | External events or trends, such as viral challenges or major cultural events, may artificially inflate or deflate engagement metrics. |
| Overfitting in predictive models | Medium | Medium | The risk of models being too specifically tailored to the training data, reducing their generalizability to new data. |
| Bias in engagement metrics | Medium | Medium | Differences in how users engage across platforms or demographics can introduce bias, potentially skewing results |
| Changes in music industry trends | Medium | Low | Rapid changes in music consumption trends or industry dynamics could impact the relevance and applicability of the analysis. |
| Dependence on external data sources | Medium | low | Heavy reliance in external data sources such as API can pose a risk if these sources become unavailable or change data formats. |

## 5.0 Deliverable Timelines

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Assessment | *First meeting with advisor to check dataset* | *Define business statement. Analysis plan delivered.* | *5* | *07/15/2024* |
| 2. | Data Exploration & Analysis | *Exploration Data and preprocessing data.* | *Conduct EDA to know the initial data; check distribution of variables, graph that generate if data has outliers, missing values etc.* | *5* | *07/22/2024* |
| 3. | Creating Models | *Develop different ML models and evaluate them,* | *Develop decision trees, logistic regression, random forest, etc. evaluate the models, and optimize them.* | *10* | *07/31/2024* |
| 4. | Governance and Documentation | *Create detailed documentation of the project and its data governance.* | *Create paper using framework. Creating insights.* | *5* | *08/05/2025* |
| 5. | Presentation | *Prepare PowerPoint Presentation* | *Conduct a final presentation that shows the project and all analytical cycle with insights,* | *5* | *08/15/2024* |