

horizontal line

Project Summary

|  |  |
| --- | --- |
| Batch details | PGPDSE-FT GURGAON JAN24 |
| Team members | Falguni, Divyansh Taneja, Manpreet Kour, Nandini Sharma, Jatin Sharma |
| Domain of Project | music data analytics and predictive modeling using  machine learning. |
| Proposed project title | Predicting Spotify Song Popularity Using Machine Learning |
| Group Number | 2 |
| Team Leader | Falguni |
| Mentor Name | Siddhartha Koshta |

Date:09/08/2024

Signature of the Mentor Signature of the Team Leader



Table of Contents

|  |  |  |
| --- | --- | --- |
| Sl NO | Topic | Page No |
| 1 | Overview | 3 |
| 2 | Business problem goals | 3 |
| 3 | Topic survey in depth | 4 |
| 4 | Critical assessment of topic survey | 5 |
| 5 | Methodology to be followed | 5 |
| 6 | References | 10 |



Project Details

# **OVERVIEW**

Spotify, a dominant player in the music streaming industry, provides users access to a vast and diverse library of songs. As artists and record labels strive to promote their music effectively, predicting a song's popularity becomes a critical factor in optimizing marketing and promotion strategies. This project is designed to address this challenge by developing a machine learning model that forecasts a song's popularity based on its audio features and metadata.

The primary goal of this project is to leverage machine learning techniques to predict how popular a song will be on Spotify. By analyzing various audio features such as danceability, energy, and tempo, alongside metadata like genre and explicit content, we aim to uncover the key factors that contribute to a song's success. The model will categorize songs into different popularity bins, allowing a nuanced understanding of how specific attributes influence a track's appeal.

Accurate predictions of song popularity have the potential to revolutionize marketing strategies within the music industry. For artists, this means a more targeted approach to promoting their tracks, ensuring that resources are allocated to songs with the highest potential for success. For record labels and streaming platforms, this insight can drive more effective campaign strategies, leading to increased engagement and profitability.

The project will employ various machine learning algorithms, with a focus on fine-tuning and evaluating their performance to ensure the best possible predictive accuracy. By examining the relationships between audio features and popularity, stakeholders can make data-driven decisions that enhance their promotional efforts. The insights gained from this model are expected to provide valuable guidance for maximizing reach and revenue in an increasingly competitive music landscape.

In summary, this project aims to harness the power of machine learning to offer actionable insights into song popularity on Spotify. By understanding the impact of various features on a track's success, artists and industry professionals can make informed decisions that drive their marketing and promotional strategies, ultimately leading to greater engagement and financial success.

# **Business problem statement (GOALS)**

**1. What would you achieve by this project?**

This project aims to develop a predictive model that can forecast the popularity of a song on Spotify. By accurately predicting song popularity, stakeholders can make better marketing and promotional decisions, leading to increased user engagement and revenue. The project will harness the power of machine learning to analyze and interpret complex datasets, providing a systematic approach to understanding music trends.

**2. How would this help the business or clients?**

The machine learning model will help record labels, artists, and marketing teams to identify potential hits and allocate resources more efficiently. Predicting popularity in advance enables targeted promotions and strategic planning, thereby enhancing the chances of success and profitability. The insights derived from the model can guide marketing efforts, optimize promotional budgets, and improve overall decision-making processes.

**3. What is the further scope of the project?**

The project can be extended to develop automated tools for real-time prediction of song popularity and integrate these insights into marketing platforms. Additionally, the model can be adapted to other music streaming services or content platforms. Future enhancements may include the development of user-friendly interfaces for stakeholders to interact with the model and derive actionable insights quickly.

**4. Limitation of the project**

The model's accuracy depends on the quality and quantity of data available. It may not generalize well to new or unique music genres. Regular updates and recalibration are necessary to maintain performance as music trends evolve. Furthermore, external factors influencing a song's popularity, such as marketing campaigns or viral trends, may not be fully captured by the model.

# **TOPIC SURVEY IN-DEPTH**

**1. Problem understanding**

In the competitive music industry, accurately predicting which songs will become popular can significantly influence marketing strategies and financial success. Traditional methods rely on expert opinions and historical trends, which may not be sufficient in the dynamic digital landscape. There is a need for a more scientific approach that leverages data analytics to predict song popularity.

Understanding the underlying factors that contribute to a song’s popularity is crucial. Audio features such as tempo, key, and loudness can provide valuable insights. Additionally, metadata like genre, duration, and release date can influence listener preferences and should be considered in the prediction model.

**2. Current solution to the problem**

Currently, the music industry relies on a combination of historical data analysis, expert judgment, and promotional strategies to gauge song popularity. These methods involve substantial manual intervention and may not effectively capture the nuances of changing listener preferences. Manual analysis is time-consuming and may not provide the timely insights needed for fast-paced marketing decisions.

Existing predictive models in the industry often focus on simplistic metrics and may lack the sophistication needed to capture complex patterns in music data. There is a need for a comprehensive model that can analyze multiple features simultaneously to provide more accurate predictions.

**3. Proposed solution to the problem**

This project proposes a machine learning approach to predict song popularity based on audio features and metadata available on Spotify. By leveraging algorithms and data-driven insights, we aim to create a more accurate and efficient prediction model. The model will use advanced techniques such as Random Forest and Gradient Boosting to analyze the data and identify key factors influencing song popularity.

The model will be trained on a large dataset of Spotify tracks, with features including danceability, energy, key, loudness, and more. By understanding the relationships between these features and a song’s popularity, we can make informed predictions about new tracks.

**4. Reference to the problem**

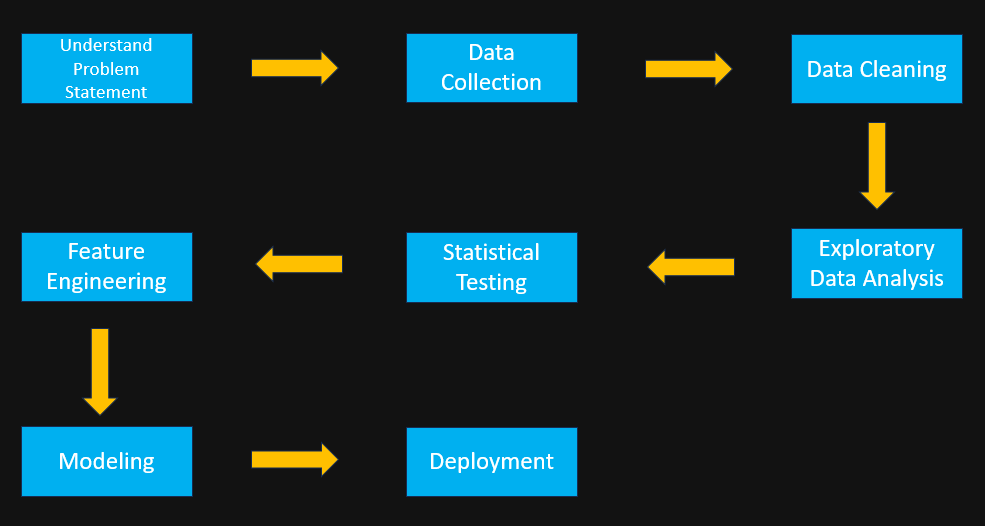
Several studies and startups have explored predictive analytics in the music industry. For instance, companies like Next Big Sound and Echo Nest (acquired by Spotify) have utilized data analytics to forecast music trends. Academic research also highlights the potential of machine learning in this domain (e.g., "Predicting Song Popularity on Spotify Using Audio Features" by researchers at Stanford University). These references provide a foundation for our project and demonstrate the viability of using machine learning for predicting song popularity.

# **CRITICAL ASSESSMENT OF TOPIC SURVEY**

The key gap identified is the lack of a robust, automated system for predicting song popularity using comprehensive audio features and metadata. This project addresses this gap by developing a machine learning model tailored for Spotify, providing a scalable and accurate solution to forecast song popularity and inform marketing strategies. Traditional methods often rely on limited datasets and may not capture the full complexity of factors influencing song popularity.

Our model will use a diverse set of features, including both audio characteristics and metadata, to provide a holistic view of what makes a song popular. By integrating these features into a single predictive model, we can provide more accurate and actionable insights. Additionally, our approach emphasizes continuous improvement, with regular updates and recalibration to adapt to changing music trends.

# **METHODOLOGY TO BE FOLLOWED**



We will follow the CRISP-DM methodology:

**1. Business Understanding**

The Spotify Track Popularity Prediction project aims to create a machine learning model that predicts track popularity based on audio features and genres. This model will support music producers, marketers, and playlist curators in making informed decisions about song promotion and playlist placement. The project involves analyzing track data to train and validate the model, focusing on accuracy and predictive performance across different popularity categories. Key success factors include model accuracy, stakeholder satisfaction, and the usability of the final prediction tool for real-time insights.

**2. Data Collection and Understanding**

The dataset, sourced from Kaggle and extracted via the Spotify Web API, includes 114,472 records with 21 features, comprising 6 categorical and 14 numerical columns, plus one target variable, Popularity. The target variable, Popularity, ranges from 0 to 100, with an average score of 52.35, indicating the song's popularity on Spotify. The dataset provides comprehensive details about each track, such as Duration\_ms (average duration of 3.85 minutes), Danceability, Energy, Acousticness, and more. Notable columns include Track\_id for unique identification, Artists with 31,437 unique names, and Explicit indicating explicit content (0 for No, 1 for Yes). Each feature contributes valuable insights into the characteristics and popularity metrics of Spotify tracks.

**Data Overview**

* **Source:** Kaggle & Spotify Web API
* **Total Records:** 114,472
* **Features:** 21 (6 categorical, 14 numerical, 1 target)
* **Target Variable:** Popularity (0-100)
  + **Average Popularity Score:** 52.35
* **Key Numerical Features:**
  + **Duration\_ms:** Average of 3.85 minutes
  + **Danceability, Energy, Acousticness, Loudness, Tempo**
* **Key Categorical Features:**
  + **Track\_id:** Unique track identifier
  + **Artists:** 31,437 unique names
  + **Explicit:** Indicates explicit content (0 = No, 1 = Yes)

3. **Data Cleaning**

* **Handling Missing Values:** Only one row contains null values, which will be removed to ensure dataset completeness.
* **Handling Duplicates:** 450 duplicate rows have been removed, retaining only the first occurrence of each duplicate to ensure data uniqueness.
* **Dropping Irrelevant Columns:** Columns such as track\_id, track\_name, and album are dropped as they are identifiers and do not contribute to predicting song popularity.

**4. Exploratory Data Analysis (EDA)**

1. **Descriptive Statistics:** The dataset reveals an average popularity score of 33.18, with scores ranging from 0 to 100. The average song duration is approximately 228,102 milliseconds (around 3.8 minutes), with a typical range between 2.9 and 4.4 minutes. The average loudness is -8.26 dB, with most songs ranging between -10 and -5 dB. Speech content in popular songs tends to be lower, indicating less instrumental dominance.
2. **Distribution of Features:**

* **Distribution Patterns and Skewness**: The distplots reveal right-skewed distributions with longer tails toward higher values. Skewness values (0.08, 11.51, 2.81) confirm this right-skewed nature. Density plots further highlight these tendencies, indicating that the data is likely non-negative and anchored at zero.
* **Outliers and Variability**: Some plots suggest potential outliers in the right tail, contributing to the overall variability observed. The distributions differ in spread and shape, reflecting the diverse characteristics of the data. Boxplots for Duration, Acousticness, Loudness, and Speechiness exhibit high skewness and significant outliers, while Popularity, Danceability, Energy, Valence, and Tempo show low skewness with fewer outliers.

1. **Correlation Analysis:**
   * **Strong Positive Correlations**: There is a notable positive correlation between Energy and Loudness (0.76), suggesting that higher energy songs tend to be louder. Valence and Danceability also show a moderate positive correlation (0.48), indicating that happier songs are generally more danceable.
   * **Weak or No Significant Correlation**: Popularity exhibits weak correlations with other features, suggesting it is not strongly influenced by any single attribute. Features like Mode, Key, and Time Signature also show weak correlations with the majority of other features.
2. **Feature Relationships:**
   * **Univariate Analysis**: Each feature is analyzed individually to understand its distribution and characteristics.
   * **Bivariate and Multivariate Analysis**: Relationships between pairs of features and multiple features simultaneously are explored. This analysis uncovers patterns and correlations, highlighting how features like Energy and Loudness interact, and how genres and artists contribute to the overall trends.
3. **Categorical Data Analysis:**
   * **Count Plots:** Categorical features such as 'explicit,' 'key,' 'mode,' 'time\_signature,' and 'genre' are analyzed using count plots. This helps in understanding the distribution and trends within categorical data.
   * **Top Genres and Artists:** Analysis shows pop film as the most popular genre, with K-pop and Chill following. Among artists, The Beatles stand out with the highest number of tracks and popularity scores, indicating their significant influence on overall trends.

**4. Statistical Testing**

* **Normality Check: Shapiro-Wilk Test**

The Shapiro-Wilk test was conducted to assess the distribution of the dataset's features. The null hypothesis (H0) posits that the data is normally distributed, while the alternative hypothesis (H1) asserts that the data is not normally distributed. The test results indicated that all features are not normally distributed (p-value < 0.05). Due to this non-normal distribution, the Interquartile Range (IQR) method was utilized for managing outliers.

* **Significance Testing: Kruskal-Wallis Test**

The Kruskal-Wallis test was employed to evaluate differences in popularity and other features across various genres. The null hypothesis (H0) suggests that the distributions of these features are identical across genres, whereas the alternative hypothesis (H1) indicates that at least one genre significantly differs. The test results showed significant differences in all numerical features, underscoring the influence of genre on various musical attributes.

**5. Feature Engineering**

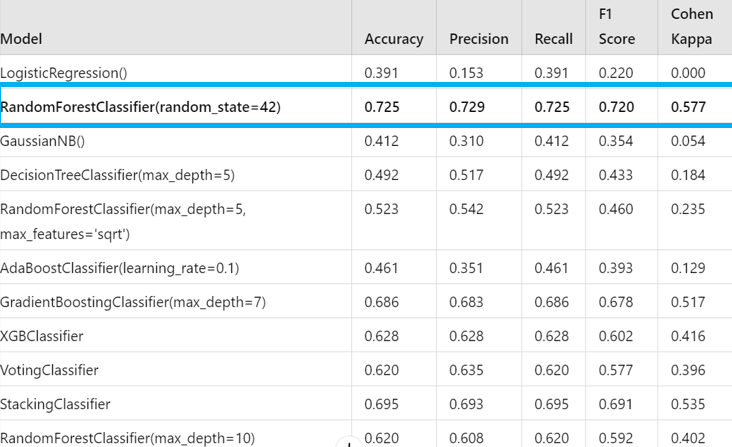
* **Encoding:** To prepare the categorical data for analysis, different encoding methods were applied based on the nature of each feature. For the Artist feature, Frequency Encoding was chosen due to the high number of unique values, providing a numerical representation based on the frequency of each artist's appearance. The Explicit feature, being binary, was encoded using Label Encoding, representing its values as 0 and 1 for simplicity. Similarly, Genre was also encoded with Label Encoding to efficiently manage its numerous unique values, transforming it into a numerical format suitable for model processing.
* **Managing Outliers:** For handling non-normally distributed data, the Interquartile Range (IQR) method was employed with a threshold of 2.5. This method is particularly effective for datasets where data distribution deviates significantly from normality.

**6. Modeling**

* **Transitioning to Classification with Binned Popularity:**

Popularity scores are binned into categories to facilitate classification rather than regression. This approach enhances model performance and interpretability, enabling the model to focus on broader trends.

* **Model Performance Summary:**



* **RandomForestClassifier:** Achieves an accuracy of 0.73 with a Cohen Kappa of 0.58, indicating excellent agreement between predicted and actual values.
* **GradientBoostingClassifier:** Performs well with an accuracy of 0.69 and a Cohen Kappa of 0.52.
* **StackingClassifier:** Offers robust performance with an accuracy of 0.70 and a Cohen Kappa of 0.54.

The RandomForestClassifier is selected for final deployment due to its superior performance across key metrics.

**7. Deployment and Testing of Spotify Track Popularity Prediction Model**

Deploying and validating a machine learning model for predicting Spotify track popularity involves several critical steps, including data collection, preprocessing, testing, and the final deployment. This comprehensive approach ensures that the model performs accurately and reliably when integrated into a real-world application. Below, we outline the process of deploying the model and evaluating its performance.

**1. Data Collection for Model Validation**

To validate the model effectively, it is essential to collect a diverse set of track data from Spotify. This is achieved using the Spotipy library, which allows us to retrieve detailed information about tracks, including audio features and popularity scores. The process involves:

* **Gathering Data:** Utilizing Spotipy to query Spotify's API and collect track details such as audio features (danceability, energy, loudness, etc.) and popularity scores across various genres and years.
* **Exporting Data:** Saving the collected data as a CSV file, which serves as the dataset for model validation.

**2. Preprocessing Spotify Track Data for Model Testing**

Before testing the model, the collected data must be preprocessed to match the format used during model training. This includes:

* **Assigning Popularity Classes:** Categorizing tracks into predefined popularity classes based on their scores. Popularity scores are binned into categories such as Low, Medium, High, and Very High.
* **Selecting Relevant Features:** Extracting and selecting features from the dataset that are necessary for model testing. This typically includes audio features and genre information.

**3. Testing and Validating Spotify Track Popularity Predictions**

Once the data is preprocessed, the next step is to use the pre-trained Random Forest model to make predictions:

* **Encoding Genres:** Applying a predefined mapping to encode genre information, ensuring consistency with the training data.
* **Predicting Popularity:** Using the Random Forest model to predict the popularity category of each track based on its features.
* **Adding Predictions to Data Frame:** Incorporating the predicted popularity categories into the Data frame for analysis.

**4. Comparing Predicted and Actual Popularity**

To evaluate the model's performance:

* **Sample Comparison:** Selecting a sample of tracks to compare predicted popularity with actual categories.
* **Performance Metrics:** Counting instances where predictions match actual categories to assess the model's accuracy. This involves analyzing the following results from a test set of 150 tracks:
  + **Medium Popularity:** 39 tracks correctly predicted out of 93.
  + **Low Popularity:** 21 tracks correctly predicted out of 40.
  + **High Popularity:** 2 tracks correctly predicted out of 16.

**5. Inference and Model Evaluation**

The model demonstrates notable performance in predicting popularity categories:

* **Strengths:** The model shows strong predictive accuracy for 'Medium Popularity,' which constitutes the majority of the dataset. This indicates that the model has a solid foundation and is well-suited for real-world applications.
* **Areas for Improvement:** The model's performance for less frequent categories like 'High Popularity' and 'Very High Popularity' needs refinement. This is due to their limited representation in the dataset, suggesting that additional data or model adjustments may be required for these categories.

**6. Deployment and User Interface**

To make the model accessible to users, a Gradio app interface is developed:

* **User Interface:** The Gradio app provides a user-friendly experience for predicting the popularity category of Spotify tracks. Users can input various song attributes and genre information to receive real-time predictions.
* **Model Integration:** The app integrates the pre-trained Random Forest model to deliver predictions based on user inputs.

**Conclusion**

The deployment and testing of the Spotify track popularity prediction model involve meticulous data collection, preprocessing, validation, and user interface design. The model has shown promising results, particularly in predicting 'Medium Popularity' tracks. While there is room for improvement in predicting rarer categories, the model is positioned well for deployment with targeted enhancements. The Gradio app interface facilitates easy access to predictions, providing a valuable tool for users in the music industry to gauge track popularity effectively.

Page 12

Reference documents of CRISP-DM

* 1. Spotify for Developers. (n.d.). Web API. Retrieved from [Web API | Spotify for Developers](https://developer.spotify.com/documentation/web-api)
  2. Forecast a Songs Popularity on Spotify by John Hopkins [Forecast a Songs Popularity on Spotify (kaggle.com)](https://www.kaggle.com/code/johnhopkins73/forecast-a-songs-popularity-on-spotify)

# **REFERENCES (10 References)**

1. Martínez-Cantin, R., de la Pezuela, C. G., & Castellano, D. A. (2020). Predicting the popularity of music using ensemble learning. Knowledge-Based Systems, 192, 105393. [Browse Computer Science journals and books - Page 1 | ScienceDirect.com](https://www.sciencedirect.com/browse/journals-and-books?subject=computer-science)
2. Herranz, R., van Balen, J., & Goethals, B. (2020). A deep learning approach for musical genre classification based on melodies and lyrics. Journal of Intelligent Information Systems, 55(1), 1-27. [Search Page | SpringerLink](https://link.springer.com/search?new-search=true&query=popularity&dateFrom=&dateTo=&facet-discipline=%22Computer+Science%22&sortBy=newestFirst)
3. Yang, Y.-H., & Chen, H. H. (2011). Music Emotion Recognition. *CRC Press* [*Music Emotion Recognition | Yi-Hsuan Yang, Homer H. Chen | Taylor & Fr (taylorfrancis.com)*](https://www.taylorfrancis.com/books/mono/10.1201/b10731/music-emotion-recognition-yi-hsuan-yang-homer-chen)
4. Bergstra, J., Casagrande, N., Erhan, D., Eck, D., & Kégl, B. (2006). Aggregate features and adaboost for music classification. *Machine Learning*, 65(2-3), 473-484. [Aggregate features and ADABOOST for music classification | Machine Learning (springer.com)](https://link.springer.com/article/10.1007/s10994-006-9019-7)
5. Zangerle, E., Pichl, M., & Specht, G. (2016). Analyzing and predicting user ratings of music tracks through automatic genre classification. *Multimedia Tools and Applications*, 75(12), 6673-6688. [[PDF] Can Microblogs Predict Music Charts? An Analysis of the Relationship Between #Nowplaying Tweets and Music Charts | Semantic Scholar](https://www.semanticscholar.org/paper/Can-Microblogs-Predict-Music-Charts-An-Analysis-of-Zangerle-Pichl/b4f0a83229319f6db1c73fe446707a2eb1a05d91)
6. Bergmann, J., & Lex, E. (2015). Analyzing the effects of personality traits on music genre preferences. In Proceedings of the 23rd Conference on User Modeling, Adaptation and Personalization (pp. 101-108). [Personality traits and music genre preferences: How music taste varies over age groups | Request PDF (researchgate.net)](https://www.researchgate.net/publication/323143644_Personality_traits_and_music_genre_preferences_How_music_taste_varies_over_age_groups)
7. Ren, Y., & Jang, J.-S. R. (2015). Automatic genre classification of MIDI files with deep neural network. In Proceedings of the 16th International Society for Music Information Retrieval Conference (pp. 649-654). [Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9422487)
8. Bent, M., & Stiller, J. (2018). Predicting the popularity of songs using Spotify data. International Journal of Applied Engineering Research, 13(6), 4224-4230. [(PDF) A Model Based Approach to Spotify Data Analysis: A Beta GLMM (researchgate.net)](https://www.researchgate.net/publication/340171849_A_Model_Based_Approach_to_Spotify_Data_Analysis_A_Beta_GLMM)
9. Asmelash, A., & Kumar, N. (2020). Predicting Spotify song popularity using audio features and metadata. Electronics, 11(21), 3518. [Full article: Predicting song popularity based on Spotify's audio features: insights from the Indonesian streaming users (tandfonline.com)](https://www.tandfonline.com/doi/full/10.1080/23270012.2023.2239824#d1e114)
10. Ferwerda, B., Yang, E., Schedl, M., & Tkalcic, M. (2017). Personality traits predict music taxonomy preferences. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 251-254). [Personality Traits Predict Music Taxonomy Preferences | Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems](https://dl.acm.org/doi/10.1145/2702613.2732754)