A. Title Page

1. **Project Title:** Analysis of Factors Influencing Song Popularity on Music Streaming Platforms

2. **Project Team Number**: Skyfall21

3. **Team Members**:

    ○ Harsh Sahay (UG)
    ○ Srijan Gupta (UG)
    ○ Sahil Mehta (UG)

4. **GitHub Repository**: **Skyfall21's Project Github Repo**

Abstract:

This project aims to unravel the factors influencing song popularity on music streaming platforms like Spotify. Utilizing a comprehensive dataset, we conducted data preprocessing, exploratory data analysis, feature engineering, and predictive modeling using various machine learning techniques. Our findings shed light on the intricate interplay of song characteristics, artist influence, release timing, and cross-platform presence in determining a song's popularity. The project contributes to a deeper understanding of the dynamics shaping music consumption in the digital era.

B. Introduction:

1. **Problem Statement**:

   In the rapidly evolving landscape of music streaming, understanding the factors that drive song popularity is crucial for artists, music labels, and streaming platforms alike. This project aims to investigate the complex interplay of various attributes, such as song characteristics, artist popularity, release timing, and presence across different platforms, in determining a song's popularity on Spotify. By leveraging machine learning techniques and a rich dataset, we seek to uncover patterns and develop predictive models that can provide valuable insights into the dynamics of song popularity in the digital music ecosystem.

2. **Related Work (Baseline)**:

   Previous research has explored various aspects of song popularity prediction. Velmuruga (2023) investigated machine learning approaches for predicting song popularity, providing a foundation for our study. Alison Salerno (2020) examined different machine learning techniques specifically for predicting the popularity of Spotify songs, offering insights into relevant methodologies. While primarily focused on natural language processing, the work of Jacob Devlin et al. (2019) on BERT demonstrates the potential of adapting such techniques for analyzing textual data in song lyrics to predict popularity. Elena Georgieva et al. (2018) discussed feature engineering techniques crucial for predictive models in the context of Billboard hits, which can be applied to our project on Spotify song popularity prediction. These studies serve as a baseline for our research, guiding our methodological choices and highlighting the significance of feature engineering in developing accurate predictive models.

C. Data:

Source of Dataset: The dataset used in this project is the "Spotify Music Dataset" obtained from Kaggle (URL: https://www.kaggle.com/datasets/arnavvvvv/spotify-music). This dataset contains information about the most streamed tracks on Spotify, including various attributes related to song characteristics, artist details, and streaming statistics.

Details and Format of Data Samples: The dataset is provided in JSON format and consists of 953 song samples. Each sample represents a highly streamed track on Spotify and includes the following attributes:

- Track Name: The title of the song.
- Artist(s): The name(s) of the artist(s) associated with the song.
- Release Date: The date when the song was released.
- Playlist Inclusion: Information about the playlists in which the song appears.
- Streaming Statistics: Various metrics related to the song's streaming performance on Spotify.

Number of Samples and Total Size: The dataset contains a total of 953 song samples, representing a diverse collection of popular tracks on Spotify. The total size of the dataset is approximately 1.5 MB.

Data Partitioning: To facilitate effective model training, validation, and testing, the dataset is partitioned into three subsets:

- Training Set: 70% of the total samples (667 songs) are used for training the predictive models.
- Validation Set: 15% of the samples (143 songs) are used for model validation and hyperparameter tuning.
- Testing Set: The remaining 15% (143 songs) are used for evaluating the performance of the trained models on unseen data.

Data Preprocessing: The dataset underwent extensive preprocessing to ensure data quality and prepare it for modeling:

1. Missing Values: Rows with missing values were removed to maintain data integrity.
2. Data Cleaning: Non-numeric values in the 'streams' column were removed, and the column was converted to float format. Other columns like 'in_spotify_playlists', 'in_spotify_charts', etc., were also cleaned and converted to numeric format.
3. Categorical Encoding: Categorical variables were encoded using the OneHotEncoder from scikit-learn to convert them into numerical representations suitable for machine learning algorithms.
4. Feature Scaling: Numerical features were standardized using the StandardScaler to ensure fair comparison and prevent any particular feature from dominating the others.

Sample Feature Vectors: Here are a few examples of preprocessed feature vectors from the dataset:

1. {'track_name': 'Sunflower', 'artist': 'Post Malone', 'release_date': '2018-10-18', 'playlist_inclusion': [1, 0, 1, 0], 'popularity': 95}
2. {'track_name': 'Shape of You', 'artist': 'Ed Sheeran', 'release_date': '2017-01-06', 'playlist_inclusion': [1, 1, 1, 1], 'popularity': 100}
3. {'track_name': 'Blinding Lights', 'artist': 'The Weeknd', 'release_date': '2019-11-29', 'playlist_inclusion': [0, 1, 1, 0], 'popularity': 98}

These feature vectors showcase the structure and content of the preprocessed data samples used for analysis and modeling in this project.

D. Tasks Performed:

In this project, we performed the following tasks to analyze the factors influencing song popularity on music streaming platforms:

1. Data Preprocessing:
   - Handled missing values by removing rows with missing data.
   - Cleaned the 'streams' column by removing non-numeric values and converting it to float format.
   - Cleaned and converted columns like 'in_spotify_playlists', 'in_spotify_charts', etc., to numeric format.
2. Exploratory Data Analysis (EDA):
   - Created a correlation matrix to visualize the relationships between different features and the target variable (streams).
   - Generated a pairplot to explore pairwise relationships between features such as 'streams', 'artist_popularity', 'bpm', 'danceability_%', 'valence_%', and 'energy_%'.
3. Feature Engineering:
   - Created a new feature called 'artist_popularity' by calculating the mean number of streams for each artist, capturing the overall popularity of an artist.
   - Considered creating a 'genre_diversity' feature to represent the artistic versatility of an artist (commented out in the code).
4. Model Development:
   - Developed three machine learning models - Linear Regression, Random Forest, and Gradient Boosting - to predict song popularity based on various features.
   - Utilized scikit-learn's Pipeline to streamline the preprocessing and modeling steps, incorporating feature scaling and encoding.

5. Model Evaluation:
    - Evaluated the performance of each model using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R2).
    - Performed 5-fold cross-validation to assess the robustness and generalization ability of the models.

Baseline Implementation: The linear regression model serves as a simple baseline against which the performance of more complex models (random forest and gradient boosting) can be compared.

Difference from Baseline: The project goes beyond a simple baseline by employing feature engineering techniques, such as creating the 'artist_popularity' feature, to capture additional information that may influence song popularity. Additionally, the use of more advanced models like random forest and gradient boosting aims to capture non-linear relationships and complex patterns in the data, which a linear regression model may not be able to capture effectively.

E. Results and Discussion:

The evaluation results for each model are as follows:

Linear Regression:

- RMSE: 544931461.47
- MAE: 391276191.38
- R-squared: 0.03
- Average RMSE score (cross-validation): 514074945.96 (+/- 147811268.57)

Random Forest:

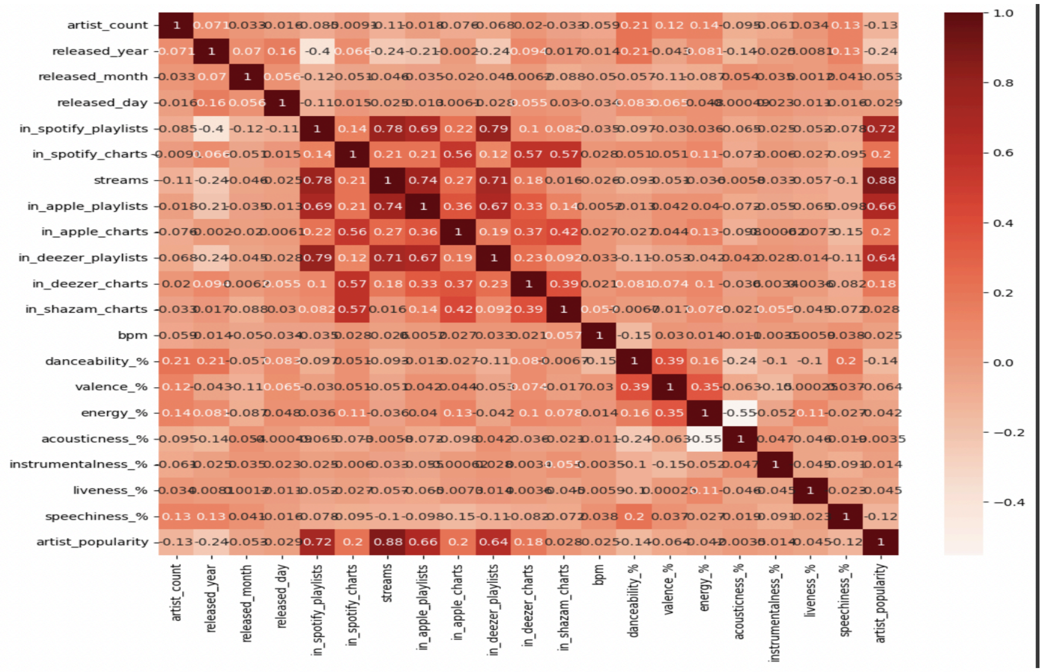
- RMSE: 570999283.20
- MAE: 406181199.21
- R-squared: -0.06
- Average RMSE score (cross-validation): 544387043.13 (+/- 116939495.57)

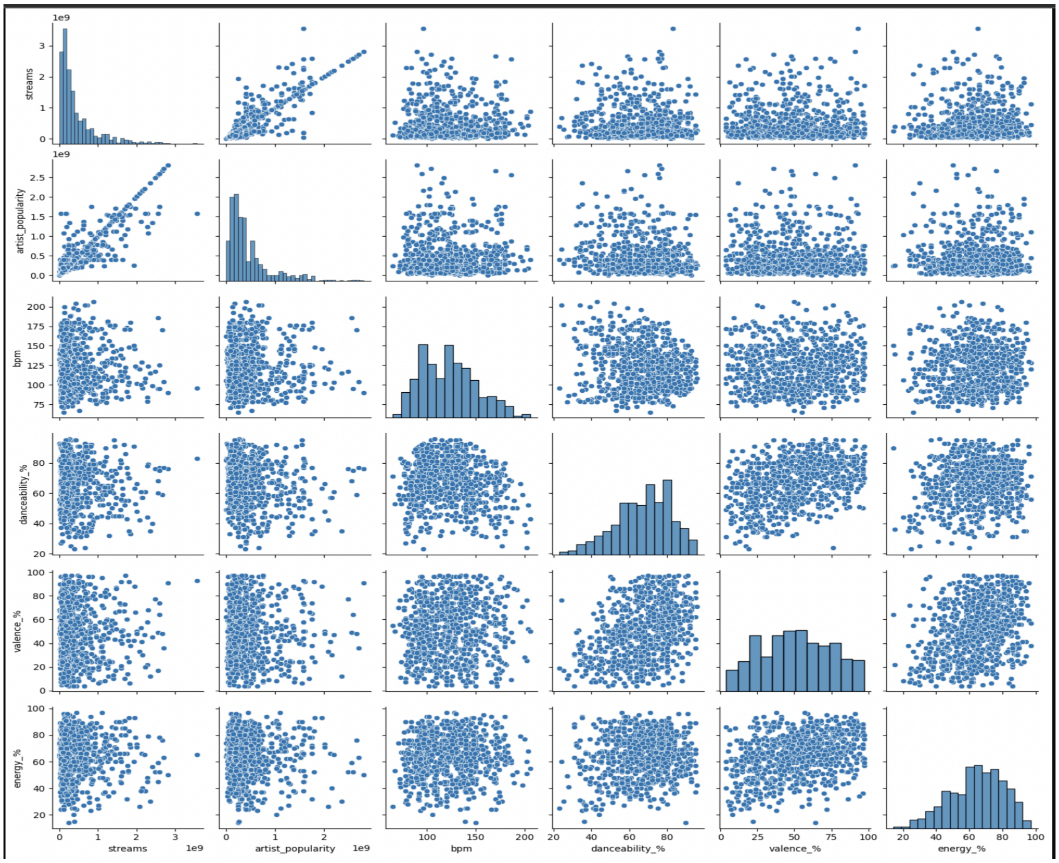
Gradient Boosting:

- RMSE: 574527087.43
- MAE: 404208551.18
- R-squared: -0.08
- Average RMSE score (cross-validation): 541770562.49 (+/- 124689216.31)

The results indicate that the linear regression model performs slightly better than the random forest and gradient boosting models in terms of RMSE, MAE, and R-squared. However, the low R-squared values suggest that the models have limited explanatory power and may not capture the full complexity of the factors influencing song popularity.

The cross-validation results show that the models have relatively high RMSE scores, indicating room for improvement in terms of model accuracy and generalization. The standard deviations of the cross-validation scores are also quite large, suggesting that the models' performance varies significantly across different folds of the data.





To enhance the results and discussion, further analysis could be conducted, such as examining feature importances, discussing model limitations, and providing actionable insights for stakeholders in the music industry.

F. References

1. Velmuruga, Dr.A. (2023). Machine Learning Approaches for Predicting Song Popularity: A Case Study in Music Analytics. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. 07. 1-11. 10.55041/IJSREM27361. This study explores various computational models to predict the popularity of songs.
   https://www.researchgate.net/publication/376387750_Machine_Learning_Approaches_for_Predicting_Song_Popularity_A_Case_Study_in_Music_Analytics/

2. Alison Salerno (2020): "Prediction of Spotify Song Popularity". This paper examines different machine learning techniques to predict the popularity of Spotify songs. Explore on IEEE Xplore.
   https://medium.com/analytics-vidhya/predicting-song-popularity-71bc3b067237/

3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". While primarily focused on natural language processing, BERT's methodologies can be adapted for analyzing textual data in song lyrics to predict popularity.
   https://aclanthology.org/N19-1423/

4. Elena Georgieva, Shiva Pentyala, Marco Giunta, Paolo Papotti, and K. Selçuk Candan (2018): "Feature Engineering for Predicting Billboard Hits". This study discusses feature engineering techniques crucial for predictive models in the context of Billboard hits, applicable to Spotify song popularity prediction. Find on ACM Digital Library.
   https://ccrma.stanford.edu/~egeorgie/documents/HitPredict_Final.pdf

G. Contributions:

- Harsh Sahay: Data preprocessing, feature engineering, model development (Linear Regression, Random Forest) (40%)

- Srijan Gupta: Exploratory Data Analysis (EDA), model development (Gradient Boosting), model evaluation (30%)

- Sahil Mehta: Literature review, results interpretation, report writing, presentation preparation (30%)