**Syed Ishmum**

**July 31st, 2023**

**BrainStation**

**Data Science, cohort of May 15th - Aug 4th, 2023**

<sydishmum@gmail.com>

[linkedin.com](https://www.linkedin.com/in/syed-ishmum/)

[github.com/sai-gh](https://github.com/sai-gh)

Capstone Business Report

# Executive Summary

This business report presents the findings and results of our project, which aims to develop a popularity prediction model using the Spotify dataset. The primary question we sought to answer was whether we could accurately predict the popularity of songs on Spotify based on various features. Our project holds significant business value as it provides the music industry, including record label companies and upcoming artists, with a powerful tool to understand which songs are likely to become popular and make informed decisions accordingly.

# Background and Business Value

The music industry is highly competitive, with countless songs being released every day. For record label companies and artists, identifying potential hits among this vast pool of songs is a challenging task. Our popularity prediction model addresses this problem by using data science techniques to analyze the features of songs and predict their popularity. By accurately identifying popular songs, record label companies can optimize their resource allocation, focusing on promoting and supporting songs with higher chances of success. For new upcoming artists, our model offers insights into which aspects of a song contribute to its popularity, aiding them in making better artistic choices and increasing their chances of success in the industry.

# Dataset Details

The dataset used for this project was sourced from Spotify’s open API under Open Database License (ODbL). This data set has 114k rows and 21 columns out of the box with 125 different genres. It also contains information about various songs, including features like danceability, energy, and loudness, among others.

The data was collected by Spotify to enhance user experience and improve their song recommendation system. It consists of structured data with multiple features and a target variable indicating song popularity. The dataset was preprocessed and cleaned to handle missing values and convert categorical features, making it suitable for modeling.

# Data Cleaning and Preprocessing

During the data cleaning and preprocessing phase, we handled missing values and transformed categorical features into numeric representations. We also performed feature engineering such as binarization and one-hot encoding to extract relevant information and prepare the data for modeling. The next step was to drop all the non-numerical columns that we were not able to convert, in the hopes of doing future modeling that would be able to read also non-numerical features such as Neural Network model.

Exploratory Data Analysis (EDA) was conducted to gain insights into the distribution of features and their relationships with song popularity. High-level visuals, such as histograms and scatter plots, were employed to present the findings to a non-technical audience.

A comparison of a bar graph

Description automatically generated

Figure : Bar graph comparison of correlation changes before and after Binarization of target variable

# Modeling and Results

Our project involved building and training several machine learning models, including Decision Trees, Random Forest, and XGBoost, to predict song popularity. First, we ran just baseline for each model without tuning to see the overall accuracy and precision. After We evaluated these models using appropriate metrics such as precision, recall, and accuracy. Once we had a breakdown of every model’s performance, we looked at the top performing models and started doing hyperparameter tuning for precision on them and test them with train data, validation data and then test data multiple times to assure the best results possible.

The Precision-tuned XGBoost model stood out as the best performer, achieving high precision while maintaining a good recall value. This model accurately identifies popular songs and minimizes false positives, making it a valuable tool for the music industry.

A screenshot of a graph

Description automatically generated

Figure : Baseline performance metrics of all the models.

A screenshot of a graph

Description automatically generated

Figure : Hyperparameter tuned XGBoost performance metric as a final model.

# Findings and Conclusion

The project's results align with our initial goals, showcasing the effectiveness of the popularity prediction model. The practical value of our model for the music industry exceeds our expectations, as it provides essential insights and aids in decision-making for both record label companies and upcoming artists. The model's applications include resource optimization, targeted marketing, and enhanced success rates for artists.

XGBoost, tuned for precision, accurately identifies popular songs with high confidence. This precision-driven approach benefits new artists and record labels, aiding in the identification of potential hits and optimizing marketing efforts. Additionally, the model's precision makes it well-suited for handling user data, enabling personalized music recommendations and enhancing user satisfaction. The high accuracy and precision of the model enhances its practical applications in the music industry and user experience.

Looking ahead, integrating user data and live data into the model will further enhance its performance and address cold start issues. Additionally, continuous refinement of the model using real-time feedback from users will ensure its relevance and accuracy in a rapidly changing music landscape. This project serves as a foundation for future research and development in the domain of song popularity prediction, offering immense potential for improving decision-making in the music industry.