Course: DSO 528, Blended Data
Author: Yixuan L., Tsubasa L., Katherine W., Camilla Z., Hanwei C., Jewel L.
Date: December 7, 2024
Subject: Final Team Project Executive Summary

To help Universal Music Group (UMG) optimize its song promotion strategy and identify trends in popular music, we developed machine learning models to predict song popularity using historical Spotify data through March 2020. Our analysis began with exploratory data analysis to uncover characteristics of successful tracks and market trends. We then experimented with various machine learning models, selecting a model that maximized net profit as our key performance indicator. Finally, we interpreted the model's results and provided actionable business recommendations to enhance Universal Music Group's promotion strategies.

We were provided with two different datasets to manipulate for our analysis, however, we only used the first dataset supplied since the first dataset had around 26,000 rows, meaning the dataset had more data to train on relative to the second dataset. However, we cleaned and enhanced the dataset using the Spotify Developer API, which will be touched upon later.

The dataset in its original state had 26,266 rows and 19 columns, with each row representing a 'track' with a track ID, name, artist, and essential song-level attributes such as danceability, energy, tempo, genre, and the binary target variable "popular". The dataset focused on high-level characteristics but lacked detailed genre information and feature combinations.

For data cleaning, data exploration revealed several anomalies: there were duplicate track IDs with multiple genre categories represented in different rows by track, a large chunk of data had release date years set as 1905 or 0, and similarly, a portion of the data had release dates in January, and there was a minimal amount of missing data in some of the columns. For the duplicate track IDs based on genre, we first grouped by track ID and expressed the different genre categories for each track ID in boolean form through one-hot encoding. This meant our dataset would have unique track IDs with multiple genres expressed based on the boolean value of the five genre columns per track ID. For the release date anomalies, we speculated that the anomalies could have been because of placeholder values. To address this issue, we connected to a Spotify API and matched data by track name and artist to obtain proper release dates for the songs in our dataset. We also performed feature engineering such as combining some categorical variables that are closely associated, or the concatenation of release months and years to make date columns for time series analysis. Our cleaned dataset had 22,944 rows and 49 columns.

For EDA, we explored the distributions and interactions of the different variables within our dataset. We summarized the data by popularity based on artist, most common release dates, genre, specific musical characteristics, etc. We also looked at time series graphs of how variables such as danceability, speechiness, and valence changed across popular, unpopular, and all songs within the dataset based on release date. Notable findings include: David Guetta being the artist with the highest number of popular songs, the release date anomalies touched upon before, Pop being the most popular genre followed by Latin, R&B, and Rap, popular songs having a distribution centered around a higher danceability value relative to unpopular songs, and the trends of popular songs having increasing danceability and speechiness across release years relative to unpopular songs. We also created correlation heat maps to observe the relationships between the variables.

After EDA, we tried feature engineering through Vader sentiment analysis and created a sentiment score for tracks based on their title. This was the only 'new' variable created for our modeling besides the variables created during the EDA process.

The profit equation based on Universal Music Group's business problem is $150,000 in revenue from popular songs, $20,000 in revenue from unpopular songs, and investment costs of $30,000 for each song. This leads to a simplified profit equation of $120,000 profit per correctly classified popular song, and a $10,000 loss per incorrectly classified song. Since the financial gains greatly outweigh the losses, many of our models use a generous probability cutoff.

For modeling, we created custom functions to evaluate model performance and KPIs. We then split our data into training and testing sets with a 70/30 split and began modeling. We started with a baseline logistic regression model including all the engineered variables and achieved a $87.99 million profit at a probability cutoff of 0.07. We then tried a decision tree model with the same variables and achieved a $83.13 million profit at a probability cutoff of 0. Then we moved on to ensemble methods such as random forest, bagging, and boosting. Out of all the ensemble methods, our XGBoost model outperformed all the models we tried with a profit of $91.86 million at a probability cutoff of 0.08. The final model we tried was a neural network model, which achieved a profit of $88.04 million at a probability cutoff of 0.06. As we modeled, we kept a close eye on the performance metrics of accuracy, precision, recall, and ROC AUC and saw how the changes in the metrics affected net profit.

After comparison, the XGBoost model emerged as the best-performing model, providing an optimal balance of precision, recall, and financial outcomes. It excelled in capturing complex patterns within the Spotify dataset, including track metadata and listener behavior, and demonstrated superior performance in terms of balancing accuracy, precision, recall, and maximizing profit. With an optimal probability threshold of 0.08, the model delivered a record $91.86 million profit, ensuring its feasibility and practicality for UMG. By achieving high recall, the model ensures that most potential hit tracks are identified, while sufficient precision minimizes wasted resources, directly supporting UMG's business objective of optimizing promotional strategies.

Thus, by integrating the XGBoost model, UMG can unlock substantial benefits across its operations. In music production, the model helps strategically balance critical song features such as danceability, energy, tempo, and valence—key predictors of popularity, which enable UMG to focus production efforts on tracks with the highest potential for success while tailoring production to align with regional preferences, such as prioritizing Latin genres in target markets. For playlist curation, the model can predict the popularity and profitability of bundled songs, allowing UMG to optimize curated playlists and ensure high-performing tracks are highlighted in ways that resonate with listeners. Playlists such as "Workout" or "Feel-Good Vibes" can feature tracks with high energy and danceability, driving listener discovery and boosting engagement. In song promotion, the model facilitates targeted marketing by prioritizing tracks with high predicted popularity probabilities. Using genre and feature-level insights, UMG can design tailored campaigns, such as promoting high-energy pop tracks during summer festivals or focusing on high-danceability Latin tracks in regions where they thrive. Furthermore, XGBoost supports collaboration strategies by analyzing trends and metadata to recommend artist pairings that maximize visibility and audience reach. These strategies collectively empower UMG to make data-driven decisions that maximize ROI and audience engagement.

Yet, the given strategies based on dataset 1 and the XGBoost model could have two key limitations. First, it relies on historical Spotify data, which may not capture future trends. Regular retraining with updated data is essential to stay aligned with evolving listener behavior. Second, while data-driven insights are valuable, artistic creativity remains critical for groundbreaking hits. UMG should balance the model's recommendations with artistic intuition to achieve commercial and artistic success. To enhance its effectiveness, we recommend integrating real-time metrics, developing region-specific models to tailor strategies, and implementing a feedback loop to refine predictions. These steps will boost song success rates, maximize ROI, and solidify UMG's position as an innovative, data-driven industry leader. By combining data insights with creative freedom, UMG can thrive in the rapidly evolving music market.