Project Executive Summary: Music in the 2010s and Early 2020s

Team 3: Jesslyn Noorjono, Sherleen Lee, Yeunbin Cho, Yangruiqi Li, Stanley Toh

This project examines the key characteristics of successful songs to help Universal Music enhance its promotion strategies using Spotify data. The datasets include two subsets: Spotify 1, which captures long-term trends with 26,266 entries of historical songs released before March 2020, and Spotify 2, a smaller, more recent dataset with 952 entries covering popular songs updated through 2023. Together, these datasets offer a clear view of how music trends have evolved over time and what factors contribute to a song's success in today's market.

Spotify 1 was selected for both exploratory data analysis and modeling. Its key metric, popular, reflects both the total number of plays and how recent those plays are, offering a dynamic measure of a song’s current success. In contrast, Spotify 2’s metric, streams, represents cumulative success up to the data collection point without accounting for whether a song is currently trending. This distinction makes Spotify 1 more suitable for identifying actionable insights into song popularity.

The exploratory data analysis identified significant trends in the music landscape. Genres such as pop, Latin, and R&B dominate, with Latin music experiencing remarkable growth in recent years, even surpassing rap and pop in some instances. Popular songs share distinctive characteristics, including high danceability, energy, and frequent use of major keys like Db/C#, C, and G, which evoke positive and uplifting emotions. Most successful tracks are studio-produced, feature prominent vocals, and align with modern listening habits, with an average duration of approximately 3:20 minutes. Key release months—January, June, and November—were identified as critical periods for maximizing exposure, coinciding with promotional cycles and seasonal engagement. External factors, such as artist fame, viral trends on platforms like TikTok and Instagram, and collaborations with well-known artists, also play a significant role in boosting song visibility and success.

Based on the feature importance analysis, key variables contributing to song success include instrumentalness (30.1%), loudness (13.9%), playlist genre (12.9%), and energy (9.8%), highlighting the significant role of musical intensity and genre. Other features such as acousticness (5.9%), danceability (6.5%), and duration (5.4%) also play a moderate role. In contrast, variables like release month (1.97%), key (0.5%), and mode (0.13%) showed minimal impact and were excluded from the modeling process.

The Random Forest model was selected for predictive modeling. This tree-based ensemble method is well-suited for capturing complex, non-linear relationships and handling mixed data types without extensive preprocessing. Songs' popularity can depend on a combination of factors (e.g., energy and loudness interacting with danceability), which Random Forest handles well through its decision tree-based approach. Its ability to aggregate predictions across multiple decision trees reduces overfitting, ensuring robust generalization to unseen data.

To maximize the model's performance, three approaches were employed. After testing various parameters, it was found that simpler parameter settings led to higher metrics and expected payoff. As a result, in this model, the number of estimators was set to 200 and the max depth to 15, with default values used for the minimum samples split and minimum samples leaf. Class weighting was also applied to address the class imbalance, assigning a higher weight of 15 to popular songs to improve recall. Then the decision threshold was lowered to 0.2 to prioritize recall, ensuring that 98% of actual popular songs were correctly identified, though this resulted in a higher False Positive Rate (FPR) of 78%. The model achieved an AUC of 0.78, indicating reasonable performance in distinguishing between popular and non-popular songs. With a cutoff of 0.22, the model generated an annual expected payoff of $155,480,000 on testing data.

In the real business world, particularly in the music industry, investment in promotions is a significant financial commitment. For a company like Universal Music, promoting a track involves not only marketing efforts but also contracting the song, which requires substantial investment. This is where the metrics of recall and expected payoff become critically important. Maximizing recall ensures that the company captures as many high-revenue tracks as possible, minimizing the risk of missing out on potential hits. By also focusing on expected payoff, the company ensures that its $30K per track investment is directed towards the most promising tracks, optimizing the return on investment. Therefore, balancing recall to identify potential hits and using expected payoff to allocate the promotion budget effectively is critical for making profitable decisions and driving revenue in the competitive music industry.

An adjustment was made to account for a 20% probability that initially unpopular songs could later gain popularity. This adjustment increased the expected payoff to $290,304,000 at a cutoff of 0.10, representing an improvement of $135,250,000. This refinement reflects a more optimistic forecast by acknowledging the potential for delayed success and reducing the risk of misclassifying songs as unpopular.

Overall, this project highlights the importance of aligning data-driven strategies with evolving trends in the music industry. In such a fast-paced industry, where consumer tastes and trends change rapidly, understanding what is recent or even anticipating future trends is crucial. This is evident in the feature importance analysis, which shows that while most variables contribute to the story, their significance alone may not be strong enough to drive critical financial investment decisions for large corporations like Universal Music. Therefore, in industries like entertainment, it is essential to complement quantitative analysis with strong domain knowledge. By integrating deep industry insights with data-driven strategies, companies can make more informed decisions that align with both market trends and consumer preferences.