**Week 10 Writeup -**

**Ryan O’Hara, Jack Metzger**

**Crucial Feature Selection**

Based on Week 10’s results, our KNN-based model achieves optimal prediction quality by leveraging cosine-distance metrics in 45% of the lyrics embeddings, 45% of the metadata, and 10% of the genre.

Among these predictors, the lyrics embeddings were the most significant variables, while the interaction of energy and loudness, referred to as Power, ranked second, indicating the intensity of the song.

Danceability was a powerful indicator for rhythm-oriented styles, and square-root normalized tempo closely matched genre clusters. By strengthening fan-based affinity, artist metadata enhanced local recommendations, while acousticness and valence aided in differentiation between genres, moods, and instrumentation.

Camelot coordinates made harmonic mixing compatible, subgenre designations helped target specific niches, and duration provided subtle clues about musical conventions, such as differentiating between progressive rock and punk.

**Five Predictions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Song Title** | **Test Artist** | **Recommended Song Title** | **Recommended Artist** |
| Waterloo | ABBA | Love Grows (Where My Rosemary Goes) | Edison Lighthouse |
| The King | Sarah Kinsley | Drunk on Halloween | Wallows |
| What You Waiting For? | Gwen Stefani | Bad Romance | Lady Gaga |
| The Night We Met | Lord Huron | High Road | Zach Bryan |
| **Sex on Fire** | **Kings of Leon** | **love is embarrassing** | **Olivia Rodrigo** |

Four of the five song recommendations above were marked as “good” recommendations, meaning they felt like a strong match when listening. What stood out across most of them was how much the lyrics seemed to drive the match - which lines up with how we’ve weighted the model. Since lyrics embeddings make up 45% of the cosine distance calculation, it makes sense that songs with similar lyrical themes, structure, or emotion end up being recommended.

Incorporating song metadata as 45% of the model's feature weighting helped ensure that recommendations stayed within a similar realm of rhythm and energy. Earlier models that relied heavily on lyrics often recommended songs that shared similar themes—such as love, heartbreak, or loneliness—but expressed them through vastly different genres (e.g., rap vs. metal vs. indie rock). By integrating metadata, the updated model was better at preserving the *vibe* of a song—matching not just lyrical content but also sonic elements like tempo, energy, and mood.

Although lyrics and metadata each contribute 45% of the model's distance calculation, it's important to note that the metadata weight is distributed across nine features (e.g., artist, popularity, sqrt\_BPM, dance, sqrt\_acoustic, happy, power, camelot\_sin, camelot\_cos), meaning each contributes roughly 5%. In contrast, the full 45% from the lyrics is carried by a single high-dimensional embedding, making it the strongest individual driver of recommendations—but now balanced by meaningful musical context.

For example, “The Night We Met” and “High Road” both have slower pacing and emotionally reflective lyrics, which likely led to the recommendation. Same goes for “What You Waiting For?” and “Bad Romance” - they both have intense, high-energy vibes and lyrical confidence, which fits how the model picks up on energy and word usage. These recommendations reflect how the model is using lyrics first, then fine-tuning based on metadata features like energy, acousticness, and danceability, and finally genre as a lighter influence.

You can see how the recommendations would shift if certain features changed - like if a song’s energy dropped or the lyrics were more upbeat, it could push it into a different cluster and change what gets recommended (shown by recommendations for lyrically different songs). These examples help show how the weighting of features (especially lyrics) plays out in the actual model behavior.

**Protected Categories**

Our dataset does not include any direct protected categories like race, gender, or age. However, there are a few features that could correlate with protected categories, even if they aren't explicitly labeled. For example, artist metadata might indirectly reflect gender or cultural background based on the artist's name. Also, genre could carry cultural associations - certain genres might skew toward specific age groups or communities, which introduces the possibility of proxy bias.

That being said, our model doesn't use protected features directly, and we don’t have demographic info on users or artists in the dataset. The features we do use - like lyrics embeddings, energy, danceability, tempo, and genre - are musical and technical in nature. It’s still important to keep in mind that bias can sneak in through correlated data, especially if certain genres or moods are underrepresented in the training data. It is important to remember that music of all genres often has racial, ethnic, and cultural origins.

**Bias in Model / Bias Removal Strategies**

As with many machine learning models, bias is an important concern and must be carefully considered during model evaluation. While not all bias can be completely eliminated, it is essential to acknowledge areas where it may arise.

Throughout our project, we observed how the bias-variance tradeoff played out as our recommendation models increased in complexity. Early models like Week 1 and Week 4, which used only lyric embeddings, had high bias and underfit the data by ignoring important musical features, though they maintained low variance and stable performance. As we added more features in Week 8—such as tempo and energy—the models captured a broader range of song characteristics, reducing bias while introducing moderate variance. By Week 9, the models became the most complex, incorporating all available data including genre. This minimized bias by fully leveraging the dataset but also resulted in the highest variance, increasing the risk of overfitting to specific features. Overall, the progression from simple to complex models reflects a clear shift from high-bias, low-variance to low-bias, high-variance, highlighting the importance of balancing these forces to achieve accurate and generalizable recommendations.

One such area of potential bias is the use of vernacular, dialect, and slang in song lyrics. Because songs are often poetic and rhythmic, they frequently incorporate non-standard language—slang, intentional misspellings, or modified words to match tempo or style. This is especially prominent in genres like rap, where slang and explicit language are more prevalent than in genres such as classical music or 60s rock. As a result, the model may learn to overemphasize certain words or phrases that are genre-specific but not necessarily semantically significant. This could lead the model to unintentionally associate certain dialects or styles with stronger predictive weight, reinforcing genre or linguistic bias in its recommendations.

Going further, music genres are historically cultural and come from different cultural roots. Rap/hip-hop traces its origins back to the Bronx, NYC with heavy African American roots, modern classical music can be traced back to 1700s Europe with roots with the Catholic church and folk music, the blues originated in the deep south, United States with heavy African American roots, country in South Appalachia, Jazz in New Orleans, Rock in the USA and UK, etc. (OpenAI). Naturally, each genre will carry its own vernacular, tempo, and patterns that allude to the specific groups and areas these genres were created and greatly influenced by.

However, we recognize that not all bias is inherently harmful in the context of our recommendation model. Since the goal is to suggest songs that align with the original track’s lyrics, message, tempo, vibe, and genre, some degree of bias can actually enhance performance. Initially, we excluded genre labels to avoid introducing genre selection bias. But in our final model, we deliberately reintroduced genre as 10% of the distance metric. This controlled inclusion of bias improved the quality of recommendations by reducing mismatches—such as suggesting songs from entirely different genres or moods—and helped ensure the model better captured the intended feel of the original song. Furthermore, given music’s deep historical and cultural roots—shaped by race, geography, and community—preserving some genre-related bias allows the model to retain the cultural uniqueness and authenticity embedded in different musical styles.

To combat some of the bias in our model, one potential strategy would be to rebalance the training set. This would involve creating a dataset with equal representation across genres, tempos, and lyrical characteristics, ensuring that no single style dominates the training process. However, a key drawback of this approach is that it would artificially alter the natural distribution of songs, disrupting the “random sampling” method originally used. As a result, the training set would no longer reflect real-world listening patterns or genre popularity, such as those found on platforms like Spotify.

Another approach could involve implementing a post-recommendation check to ensure that KNN outputs follow a more balanced genre distribution. While this might help reduce overrepresentation of popular genres like rap or rock, it risks interfering with the model’s core purpose: recommending the most similar song based on cosine distance. The goal is not to achieve genre equity in every recommendation, but rather to provide the best match to the original song based on lyrical and musical similarity. This is why we ultimately chose to reintroduce a small amount of genre bias into our final model by weighting genre at 10% of the distance metric. This intentional bias helped the model produce more favorable results by nudging recommendations to stay within a similar musical realm. In doing so, we embraced the idea that not all bias is harmful, as stated above,—particularly in music.

This tradeoff highlights the challenge of reducing bias without sacrificing the authenticity and representativeness of the music landscape.

**Risks of Model Use**

Two of the biggest risks we discussed that could come from this model include:

1. Representation : If the model is trained on a dataset that underrepresents certain artists, especially indie or emerging artists from marginalized backgrounds - those voices could get pushed out of recommendations entirely. That has downstream effects on exposure, income, and cultural visibility, especially for smaller or newer artists trying to break through.
2. Platform-level business risk : If the recommendations don’t feel diverse or fresh, users might disengage from the app entirely. If stakeholders (like record labels or the public) perceive that the system is biased toward mainstream content or major label artists, it could damage trust and raise questions around fairness or manipulation, especially if there are financial incentives involved (pay by stream business models).

Comment, et al. “Bias and Variance in Machine Learning.” *GeeksforGeeks*, 12 July 2025, www.geeksforgeeks.org/machine-learning/bias-vs-variance-in-machine-learning/.

OpenAI. "Origins of Major Music Genres: Rap, Classical, Blues, Country, Jazz, Rock, Indie." *ChatGPT*, 27 July 2025, [chat.openai.com](http://chat.openai.com).