**Week 9 Writeup - Select the Winning Model**

**Ryan O’Hara, Jack Metzger**

**All Models**

In order to maximize generalization and prediction accuracy, we thoroughly assessed nine models in Week 9 that included various combinations of genre encoding, metadata, and GloVe-based text embeddings. First, there are naive baseline algorithm, from separate embedding models, metadata, and genre to combinations of these.

There are final approaches in 9 models, which has been assigned in distance strategy with cosine based model with KNN works, assigning 45% of the influence to embeddings, 10% to genre, and 45% to metadata.

This balance was the best configuration in terms of accuracy and bias-variance equilibrium since it effectively reduced feature-type dominance and dimensional imbalance, resulting in the lowest validation error and consistent test performance.

**Winning Model**

The model with all features and weighted distance was selected which was assigned 45% of the influence to embeddings, 10% to genre, and 45% to metadata. By using the best configuration in terms of accuracy and bias-variance equilibrium since it effectively reduced feature-type dominance and dimensional imbalance, resulting in the lowest validation error and consistent test performance.

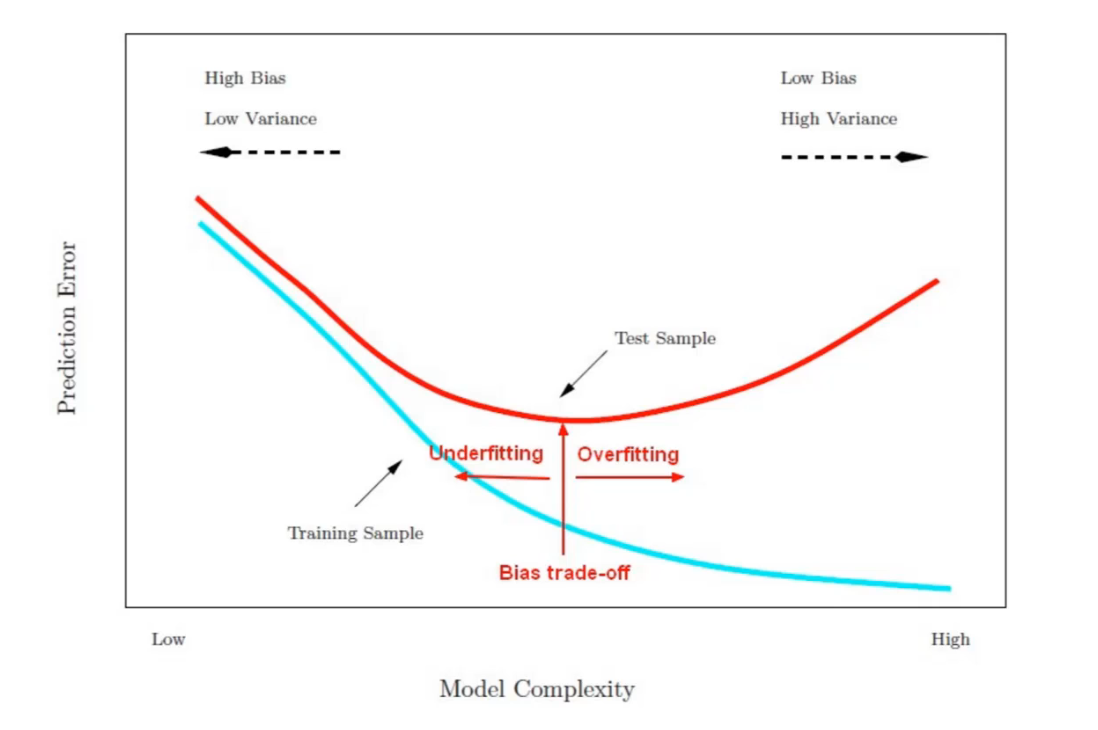
To guarantee a fair contribution across a range of feature blocks, these weights were transformed into per-feature scaling constants such as using one-hot genre encodings, and glove-twitter-200 text embeddings within a cosine-based KNN Models.

By achieving improved validation accuracy, maintaining low test error, and striking a strong bias-variance balance, the weighted concatenated matrix that resulted allowed the model to provide dependable generalization and prevent any one feature set from dominating the similarity calculation.

This new model with the genre weightings ended up being the winning model, clocking in at a new high of 60% accuracy. We expected this new weighting to take care of some problems specifically with the country and electronic genres, and we were delighted to see much improved accuracy for those types of songs.

I think in future iterations it would be interesting to see if we could only include this variable for songs in those two genres, as it very much helped the accuracy there, however I think other genres experience slight downgrades in the quality of their recommendations.

**Bias-Variance Trade off**

****

For our project there is not a traditional way to show a chart of the bias variance tradeoff although we can still analyze this theory with the complexity of our models. When comparing our different recommendation models, we can clearly see how the bias-variance tradeoff plays out across model complexity. The earlier models, such as Week 1 (FastText lyrics-only) and Week 4 (Twitter lyrics-only), are relatively simple. These models rely only on lyrical content and have high bias, meaning they are likely to underfit and miss important variation in song characteristics outside of lyrics. We saw this with the recommendations given for songs like country, and heavy metal. However, because of their simplicity, they exhibit lower variance and are less prone to overfitting, making them more stable but less accurate overall.

Week 8 models introduce more complexity by incorporating all available features except genre. This creates a more balanced tradeoff: the model captures broader musical features like tempo and energy, which reduces bias, while maintaining a moderate level of variance. This model builds on this by adding custom weightings that emphasize lyrical content. This increases the model's complexity and variance, as it starts to fine-tune recommendations based on more nuanced user preferences, but it can also increase the risk of overfitting to lyric-driven patterns.

Week 9 models are the most complex, using all available data including genre and metadata. It has the lowest bias because it leverages the full richness of the dataset. However, this also leads to the highest variance - particularly if the model becomes overly reliant on categorical features like genre. This model is likely to perform best on a large, diverse dataset like our combined group data, but care must be taken to ensure it generalizes well.

Overall, our progression from simpler to more complex models reflects a gradual shift from high-bias, low-variance to low-bias, high-variance. Understanding this tradeoff helps us evaluate which model performs best in practice - not just on training data, but in delivering accurate, meaningful recommendations across users.

**Model Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Number** | **KNN (n=)** | **Accuracy** | **Description** |
| Model 1 | 5 | \*The accuracy for the beginning models was not calculated as it wasn’t determined how we would do it with our recommendation model | Lyrics-only model using FastText embeddings. Tested with varying n-gram sizes (n=5, 15, 50). |
| Model 2 | 15 | “We evaluated our model in a more general sense as we are not currently clustering or using any sort of labels in this iteration of our model building process. We did this by gathering the top 5 recommended songs for many different input songs. While we have included some interesting examples below, there are a few general observations that can be made from this model thus far.” |  |
| Model 3 | 50 | “while there are certainly recommendations that make a lot of sense, you can tell that there is a reason songs cannot be suggested purely based on the merit of their lyrics.” - *Refer to Week 5 Writeup* |  |
| Model 4 | 5 | The manual accuracy score for this model was 49 ± 5, reflecting a 44–54% relevant recommendation accuracy range after accounting for subjectivity in scoring. | Lyrics-only model using Twitter embeddings, same n-gram tuning as Model 1. |
| Model 5 | 15 | - | - |
| Model 6 | 50 | - | - |
| Model 7 | 5 | 43/100 | Used all features except genre to build a content-based recommendation model. |
| Model 8 | 5 | 49/100 | Same as Model 7, but with adjusted feature weights—lyrics were weighted more heavily. |
| Model 9 | 5 | 55/100 | - |
| Model 10 | 6 | 60/100 | Used all available data, including genre and other metadata, for full-feature modeling. Because we used all data, K=6 was essentially K=5 because the first recommendation was always the original song |