**Week 10 Writeup - Data-Centric AI**

**Ryan O’Hara, Jack Metzger**

**Broad Project Review**

To mitigate the impact of anomalies and minimize label noise, we excluded tracks with popularity scores below the 10th percentile and above the 90th percentile.

Secondly, to normalize their distributions and enhance conformity with model assumptions, transform "Acoustic" and “BPM” columns using the square root to achieve a more normal looking distribution. There is a new variable called "power," which is the multiplication of energy and loud. This makes it easy for us to capture the relationship between two variables. Camelot was coded to represent sin and cos values for plotting it on a unit circle. This allowed us to include Camelot as a numerical variable vs a categorical variable. All numeric parameters (other than the lyric vectors) were normalized to fit a scale from 0 to 1.

We removed genre labels with low occurrences to simplify model dimensions. Once reduced, genre tags were one-hot encoded for use in the model.

We then incorporated each of the three datatypes (lyric vectors, music metadata, and genre labels) into one KNN model, where lyrics contributed 45%, metadata contributed 45% and genre labels contributed 10% of the cosine-distance. Our best model thus far has achieved an accuracy of 60% when recommending songs.

**Comparison with Week 9**

In Week 9, we weighed GloVe lyric embeddings, music metadata, and genre one-hot encoding to improve a cosine-distance based KNN recommender. We weighted the distance as: 45% lyrics, 45% music, and 10% genre, which improved song recommendations by integrating genre-specific data and attained 60% (+/- 5% accuracy) relevant recommendation accuracy.

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.

**Model Improvements: Week 10 Data Refinement**

To further enhance our model’s performance, we implemented a series of relatively minor data adjustments this week. These changes did not involve the creation of new features or major transformations—those were addressed in earlier phases. Instead, we focused on refining the dataset to improve predictive accuracy and model robustness.

*1. Removal of Underrepresented Genres*

We removed the genres: folk (112 songs), dance (111 songs), and rnb (86 songs) from the dataset due to their relatively low representation compared to some of the more dominant genres such as pop (546 songs), indie (539 songs), and rock (539 songs). Our hypothesis was that these underrepresented genres may introduce unnecessary noise and class imbalance, which can hinder the model’s ability to generalize. By excluding them, we aim to reduce overfitting on small sample sizes and improve the model's focus on genres with enough data to learn meaningful patterns.

**Table 1:** Table depicting the counts of our major genres.

*2. Introduction of Artist Metadata*

We added the “Artist” column into the metadata to encourage the model to recommend songs by the same artist.. The idea is rooted in user behavior. listeners often enjoy multiple songs from the same artist due to consistent musical style, tone, and lyrical themes. Including artist information should help the model capture and reflect these artist-specific patterns in recommendations.

*3. Addition of Subgenre Metadata*

We incorporated subgenre into the metadata to provide the model with additional context, intended for niche tracks. In some cases, the primary genre alone does not accurately represent the song’s true sound or audience. Adding subgenre may allow the model to provide better recommendations for these niche songs/genres.

**Comparison with Week 9:**

When comparing our Week 10 model with the changes described above to our week 9 model, we found that our week 9 model performed better. Our Week 10 model achieved an accuracy of 58% (+/- 5%) where our Week 9 model achieved 60% (+/- 5% accuracy).

Although the accuracy difference falls within the margin of error (+/- 5%), meaning the results are not statistically significant, Week 9 achieved a higher point estimate on the same testing set. As a result, we are choosing Week 9’s model as our preferred version for deployment.

We also considered the cost-benefit tradeoff of the additional complexity introduced in Week 10. Despite our efforts to fine-tune the data: filtering low-representation genres and incorporating artist and subgenre metadata; the added dimensionality and potential noise did not yield measurable gains. The added dimensions and noise in Week 10 are not worth the tradeoff of achieving a -2% accuracy, even if it falls within the error of Week 9.

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |
| Model 11 | 2 | 58/100 | Removed genres: folk (112), dance (111), and rnb (86). Added “Artist” column into the metadata, and added subgenre into the model.  Maintained the 45% lyrics, 45% metadata, 10% genre cosine-distance weighting from model 10. |

**Table 2:** Table showing all the different model versions, their accuracies, and the specific changes made for each one. Model 11 represents this week’s model. Model 10 is our “Winning Model”.

**Overall Insights**

It is important to note that the errors of our selected model are all-encompassing in our overall dataset, no longer separated into test, train, and validation. We did this because, as a recommendation model, we prefer to have as many possible songs to choose from in our recommendation pool as possible. Also, because we don’t need to be concerned with overfitting, the usual benefits of splitting data no longer apply. The only change we made to adjust for this was to technically take the second recommendation, as the first recommendation will always be the input song.

Overall, we as a group are comfortable with the 42% error we have received at this stage in the model building process. While we are nearing the end of this process for the scope of this course - in a real job, we would have many more weeks and resources to further iterate on our model. We have experienced massive growth in accuracy in a relatively short amount of time and given more time to expand on our current weightings and feature spaces, we feel that we could continue making substantial improvements.

One last note, specific to our project, our error is a very subjective measurement that, given more time spent in the evaluation process, can be further refined. As we are currently more focused on actual model development rather than evaluation, we are limited to grading ~100 songs a week. Given more time and resources to grade our recommendations, we would be able to pick up on more qualitative nuances that would further improve our model. From grading thus far we have been able to pick up on small quirks that are present only in certain types of songs, and more time spent on grading in the future would allow us to notice more non-quantitative characteristics.