Project CheckPoint 2

Shishi Jiang (u1583346)
Chuyue Wu (u1590131)

1. Rationale for Dataset Transition

During the Exploratory Data Analysis (EDA) phase using the standard Million Song Subset, we encountered absence of target variable (Genre Labels) issues. We discussed it during Checkpoint 1 and found a method to solve it below.

Our primary objective requires the supervised classification of musical genres. Upon detailed inspection, the standard 10,000-song subset, while containing rich feature data, lacks the specific genre tag required for supervised learning. After revisiting the official Million Song Dataset documentation and download page, we discovered an additional genre dataset provided by the authors. This auxiliary file contains around 60k pre-processed tracks with associated genre tags. The dataset creators selected popular human-curated artist-level tags from MusicBrainz, grouped them into 10 broad genres, and assigned each artist's tracks accordingly.

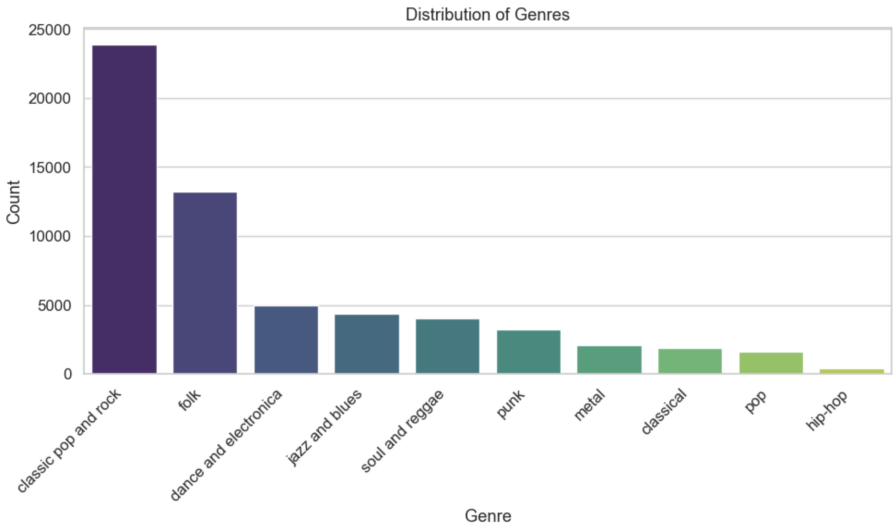
Our first idea was to keep using the 10k subset and simply merge these genre labels into it. However, when we actually attempted the merge, we found that the overlap between the 10k subset and the 60k official genre dataset was very small. Only less than 1k rows could be matched to a genre. The remaining majority of rows had to be dropped, producing a dataset that was too small and sparse. We think the reason is that both datasets are randomly generated from the 1m full Million Song Dataset, resulting in only a small overlap.

To avoid training on such a tiny and biased sample, we decided to transition entirely to the 60k genre dataset. This dataset is already cleaned, and it natively includes the genre labels we require. It also has all the predictors that we originally intended to use in the 10k subset. It allowed us to pursue our original goal of supervised genre classification and keep a much larger number of labeled tracks.

2. Dataset Analysis and Preprocessing

2.1 Class Distribution and Imbalance

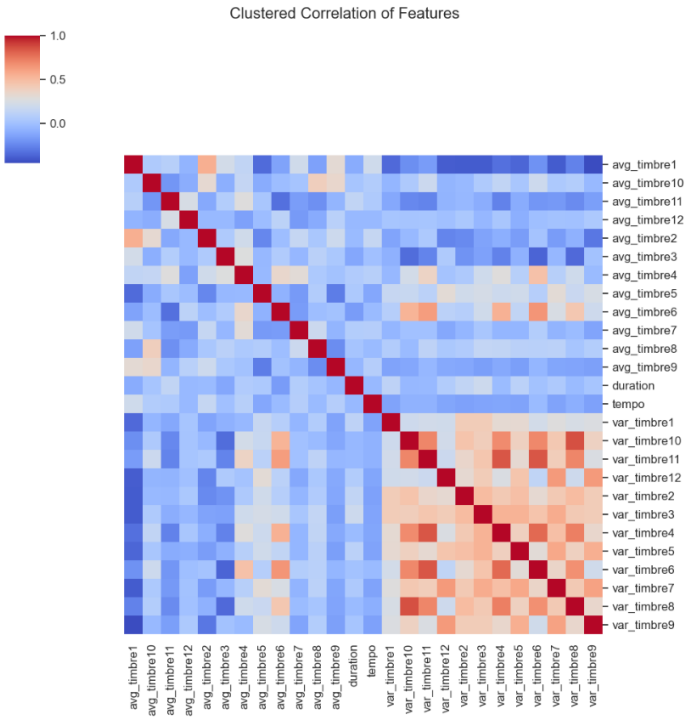
The original dataset exhibits a severe class imbalance. The genre distribution is heavily skewed towards "Classic Pop and Rock" (23,895 samples), while minority genres like "Hip-Hop" contain as few as 434 samples. This imbalance poses a significant challenge, as models tend to bias predictions toward the majority class to minimize global error.

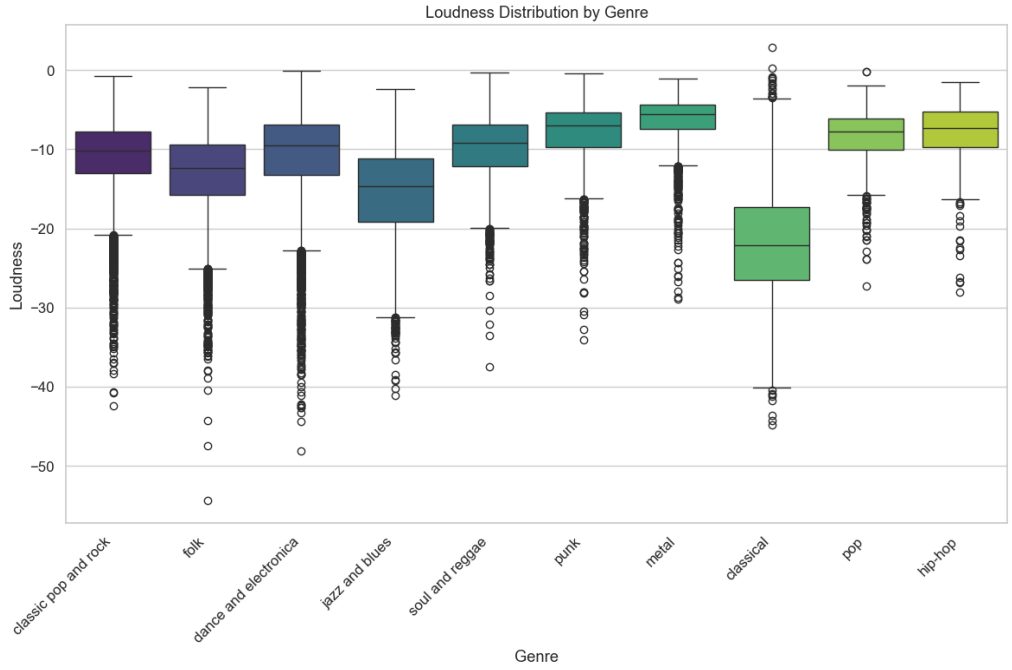


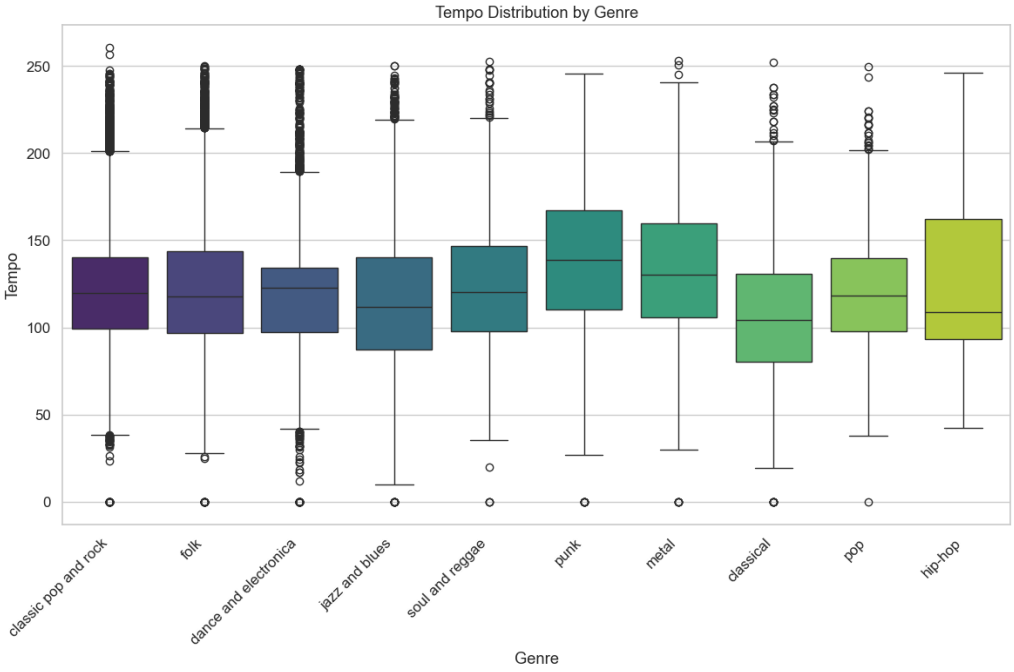
2.2 Feature Analysis

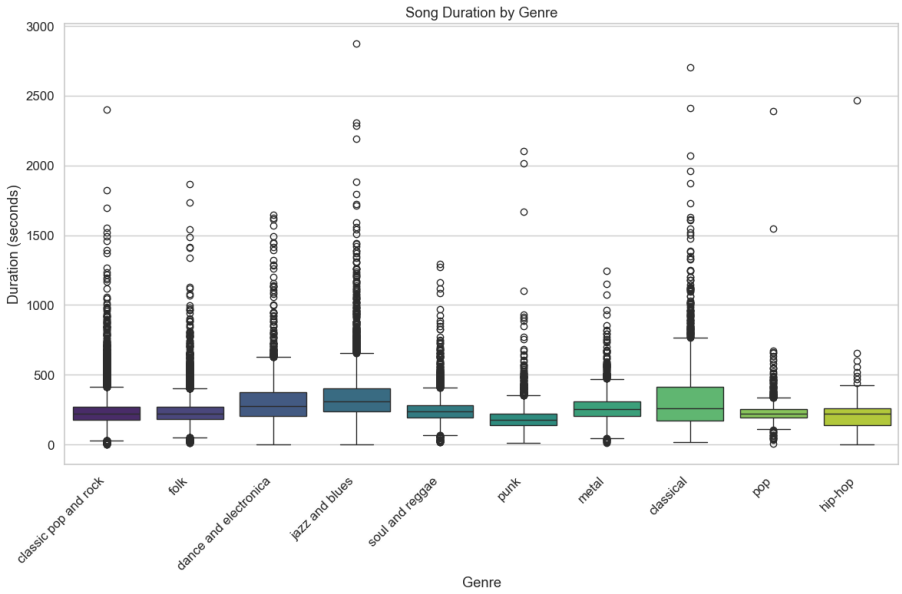
We analyzed the relationships between key audio features to understand their discriminative power. Correlation Analysis: We found a 0.95 correlation between loudness and avg_timbre1. While highly correlated, our result showed that retaining both features improved model performance, suggesting avg_timbre1 captures timbral nuances distinct from pure loudness.

```
Original Logistic Regression Accuracy: 0.5698
Accuracy without avg_timbre1: 0.5636
Result: KEEP it. The feature adds valuable signal despite correlation.
```




Loudness Distribution by Genre
Loudness
0
−10
−20
−30
−40
−50
classic pop and rock
folk
dance and electronica
jazz and blues
soul and reggae
punk
metal
classical
pop
hip-hop
Genre


Tempo Distribution by Genre
Tempo
250
200
150
100
50
0
classic pop and rock
folk
dance and electronica
jazz and blues
soul and reggae
punk
metal
classical
pop
hip-hop
Genre



Below we separate our tasks for model training. Shishi Jiang trained in Logistic Regression and Decision Tree. Chuyue Wu trained in Random Forest and SVM.

3. Logistic Regression and Decision Tree Models from Shishi Jiang

The data was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve class ratios. All features were standardized using StandardScaler to ensure models like Logistic Regression were not biased by variable magnitudes.

I trained Logistic Regression, Decision Tree on the original dataset.

Result: Logistic Regression achieved ~57% accuracy, Decision Tree achieved ~43%.

Analysis: The confusion matrix revealed that the models performed well on "Classic Pop and Rock" but failed significantly on minority classes. Genres with distinctive acoustic signatures—such as metal, classical, and classic pop/rock—achieve higher prediction accuracy because their spectral timbre profiles, loudness levels, and duration distributions cluster tightly in feature space.In contrast, genres like pop (Specifically, recall for "Pop" was near 0%), soul/reggae, and hip-hop are harder to classify because they exhibit significant feature overlap with other genres, greater within-genre heterogeneity, and, in the case of hip-hop, limited sample size.This leads to diffuse feature distributions, which reduces separability and lowers the model’s recall for those categories.

```
Logistic Regression Test Accuracy: 0.5698
                      precision    recall  f1-score   support

 classic pop and rock       0.55      0.80      0.65      3584
            classical       0.69      0.69      0.69       281
dance and electronica       0.58      0.40      0.47       741
                 folk       0.59      0.53      0.56      1979
              hip-hop       0.27      0.05      0.08        65
       jazz and blues       0.61      0.33      0.43       650
                metal       0.72      0.62      0.66       315
                  pop       0.00      0.00      0.00       243
                punk        0.65      0.36      0.46       480
      soul and reggae       0.42      0.19      0.26       602

             accuracy                           0.57      8940
            macro avg       0.51      0.40      0.43      8940
         weighted avg       0.56      0.57      0.54      8940
```

The multinomial Logistic Regression model achieved a test accuracy of 0.5698 and a macro-averaged F1 score of 0.4265, indicating moderate overall performance with substantial variation across genres. As shown in the classification report, the model performed well on classical (F1 = 0.69), metal (F1 = 0.66), and classic pop & rock (F1 = 0.65), suggesting that these genres possess more distinct acoustic patterns that can be separated linearly in the feature space. Mid-range results were observed for folk (F1 = 0.56), dance/electronica (F1 = 0.47), punk (F1 = 0.46), and jazz & blues (F1 = 0.43). However, performance dropped significantly for hip-hop (F1 = 0.08) and pop (F1 = 0.00), where the model failed to correctly classify most instances.

```
Decision Tree Accuracy: 0.4391
                      precision    recall  f1-score   support

 classic pop and rock       0.54      0.53      0.53      3584
            classical       0.55      0.57      0.56       281
dance and electronica       0.33      0.34      0.34       741
                 folk       0.45      0.46      0.46      1979
              hip-hop       0.10      0.09      0.10        65
       jazz and blues       0.34      0.35      0.34       650
                metal       0.49      0.47      0.48       315
                  pop       0.07      0.07      0.07       243
                punk        0.33      0.33      0.33       480
      soul and reggae       0.25      0.24      0.25       602

             accuracy                           0.44      8940
            macro avg       0.35      0.34      0.35      8940
         weighted avg       0.44      0.44      0.44      8940
```

The Decision Tree model achieved a test accuracy of 0.4391 and a macro F1 score of 0.3452, which is notably lower than Logistic Regression (accuracy 0.5698, F1_macro 0.4265). Performance was moderate for high-support genres such as classic pop & rock (F1 = 0.53), classical (F1 = 0.56), folk (F1 = 0.46) and metal (F1 = 0.48), but substantially weaker for minority or acoustically overlapping genres — particularly

hip-hop (F1 = 0.10), pop (F1 = 0.07), and soul/reggae (F1 = 0.25). The drop in metrics suggests that a single unpruned tree tends to overfit dominant patterns while failing to generalize complex boundaries across classes. Overall, Decision Tree serves as a baseline non-linear model, but its standalone performance is insufficient, motivating the use of ensemble methods (Random Forest, XGBoost) to capture richer decision structures.


```
Logistic Regression Validation Accuracy: 0.5600
Decision Tree Validation Accuracy: 0.4326
```

4. Random Forest and SVM Models from Chuyue Wu

First, I splitted the 60k genre data into an 80/20 train-test split. I performed 3-fold cross-validation on the training porting for both random forest and SVM models, to find the best hyperparameters.

For the random forest model, I experimented with the number of trees (n_estimators), depth of trees (max_depth), minimum samples required for a split (min_samples_split), and minimum samples per leaf (min_samples_leaf). After 3-fold training, the best-performing model happens at 400 trees, unlimited depth, min_samples_split = 2, min_samples_leaf = 1.

On the held-out 20% test set, the model achieved an overall accuracy of 0.61. While the accuracy is moderate, the model displays strong performance on classical (F1 score: 0.76) and metal (F1 score: 0.70). In contrast, genres with fewer samples like hip-hop and pop suffer from noticeably low recall rate. Even if I added “class_weight="balanced" in the RandomForestClassifier setting, the result doesn’t change much. This confirms that class imbalance remains a limitation.

However, this is not only about class imbalance. Classical has 1874 samples in the data and achieves highest F1-score 0.76. Pop has 1617 samples in the data but archives lowest F1-score 0.07. This indicates that classification performance is not solely determined by sample size. Instead, it strongly depends on the intrinsic separability of each genre. These relatively high scores indicate that random forest captures

distinctive acoustic patterns in these genres. For example, classical music often has clear dynamic range and timbre patterns, commonly associated with instruments such as piano and strings. And metal music typically has strong rhythmic and specific timbral signatures like electric guitar and drum. Conversely, genres such as pop and hip-hop exhibit highly overlapping timbre patterns, causing significantly lower separability.

Random Forest on Test Set

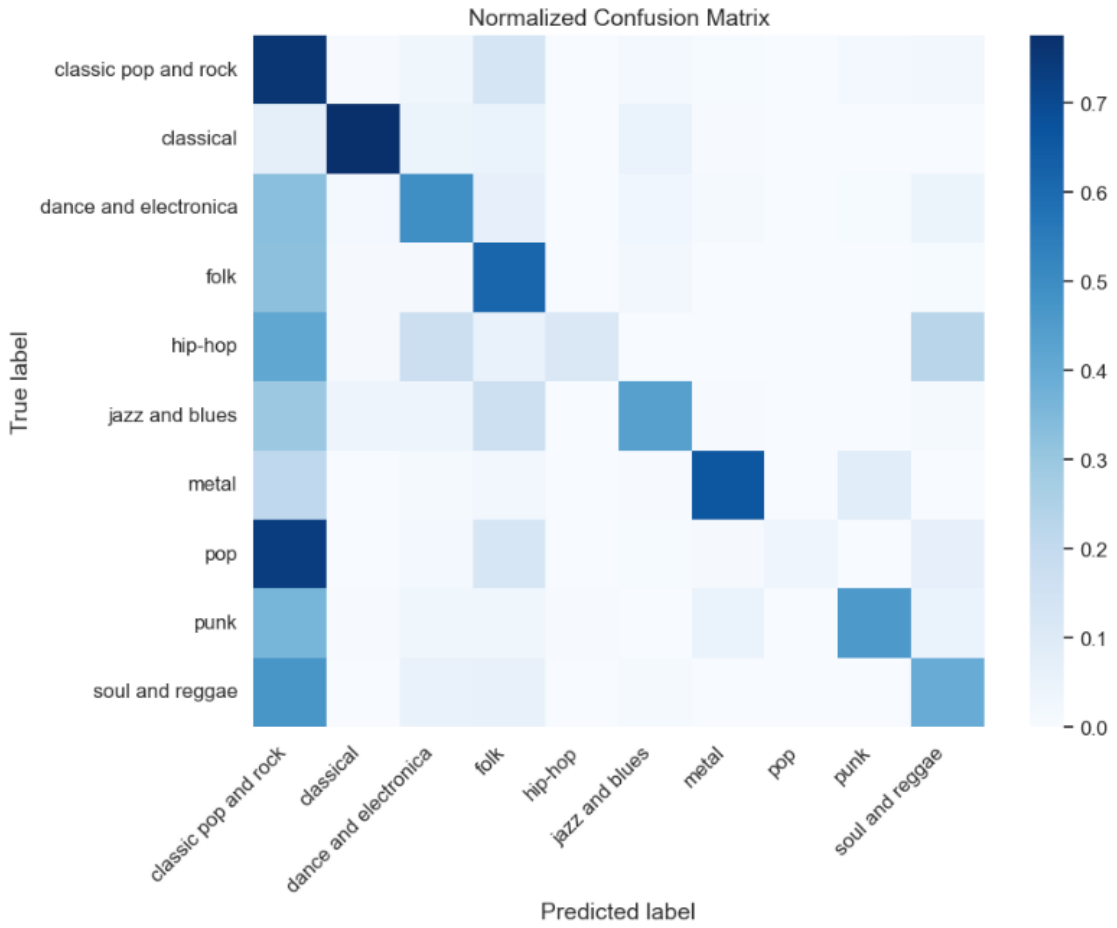
	precision	recall	f1-score	support
classic pop and rock	0.60	0.76	0.67	4779
classical	0.74	0.78	0.76	375
dance and electronica	0.60	0.49	0.54	987
folk	0.62	0.61	0.62	2638
hip-hop	0.56	0.11	0.19	87
jazz and blues	0.64	0.44	0.52	867
metal	0.74	0.66	0.70	421
pop	0.80	0.04	0.07	323
punk	0.68	0.46	0.55	640
soul and reggae	0.54	0.40	0.46	803
accuracy			0.61	11920
macro avg	0.65	0.47	0.51	11920
weighted avg	0.62	0.61	0.60	11920

(Random Forest Result on Test Set)

```
label_data.groupby('genre').size().sort_values(ascending=False)
```
✓ 0.0s

```
genre
classic pop and rock     23895
folk                     13192
dance and electronica     4935
jazz and blues            4334
soul and reggae           4016
punk                      3200
metal                     2103
classical                 1874
pop                       1617
hip-hop                    434
dtype: int64
```

(Genre Sample Count)



(Random Forest Confusion Matrix Result)

For the SVM Model, I tuned the following hyperparameters using 3-fold cross-validation:

- C: Controls the penalty for misclassification.
- gamma: Defines how far the influence of a training point reaches.
- kernel='rbf': Enables nonlinear decision boundaries suitable for curved feature spaces.
- max_iter=2000: Sets the maximum optimization iterations to prevent excessive runtime on large datasets.

After training, the best SVM model happens at {'clf__C': 10, 'clf__gamma': 'scale', 'clf__kernel': 'rbf', 'clf__max_iter': 2000} with an overall test accuracy of 0.46. This is much lower than random forest accuracy of 0.61. Similar to the Random Forest results, classical (F1 = 0.78) and metal (F1 = 0.71) remain the best-performing genres, while pop, hip-hop, and punk continue to show the lowest recall and F1-scores. In other words, although absolute performance drops, the relative difficulty of each genre remains consistent across both models.

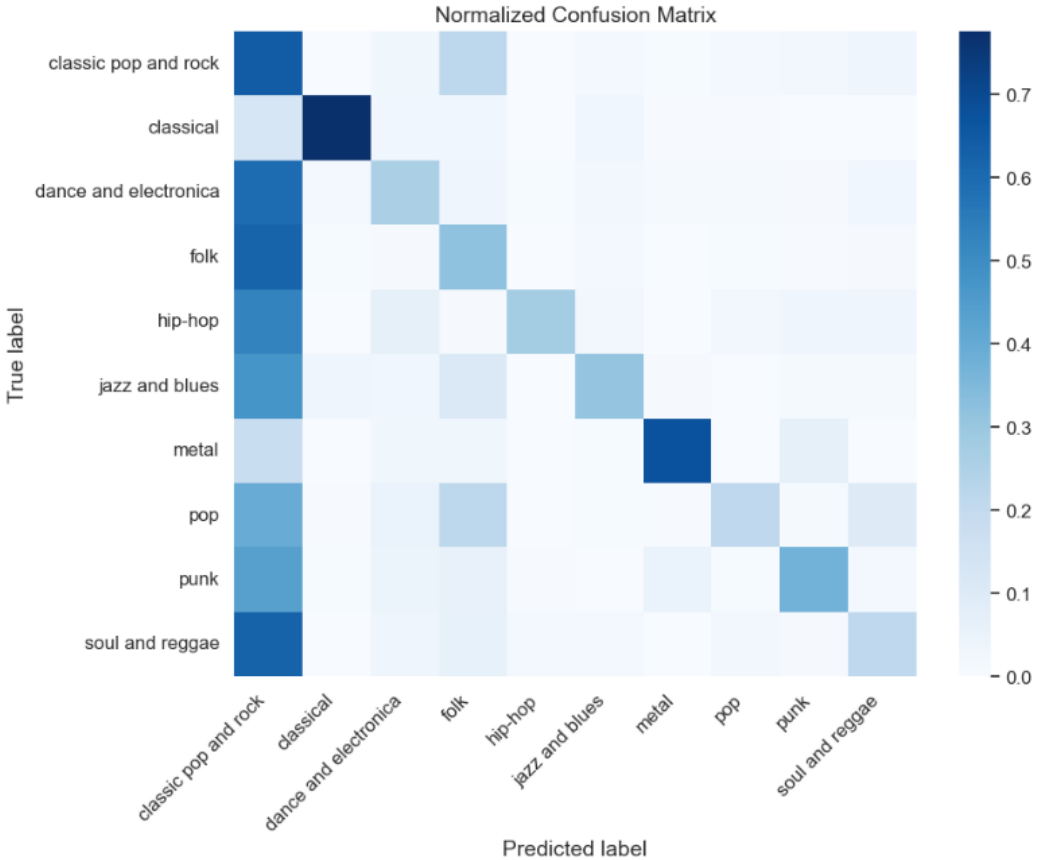
Overall, SVM performs worse than random forest mainly due to the structural difference between two models. SVM with an RBF kernel is not very suitable for large-scale, multiclass, and highly nonlinear data. In our data, the timbre based features involve complex interactions that tree-based models capture

more naturally. Additionally, SVM is more sensitive to class imbalance and overlapping feature distributions. But random forest constructs flexible hierarchical splits, making it better suited for our music classification task.

SVM on Test Set

	precision	recall	f1-score	support
classic pop and rock	0.45	0.64	0.53	4779
classical	0.78	0.78	0.78	375
dance and electronica	0.45	0.26	0.33	987
folk	0.38	0.32	0.35	2638
hip-hop	0.39	0.28	0.32	87
jazz and blues	0.60	0.31	0.41	867
metal	0.76	0.67	0.71	421
pop	0.32	0.21	0.25	323
punk	0.55	0.38	0.44	640
soul and reggae	0.36	0.21	0.26	803
accuracy			0.46	11920
macro avg	0.50	0.41	0.44	11920
weighted avg	0.46	0.46	0.45	11920

(SVM Result on Test Set)



(SVM Confusion Matrix Result)

5. Next Steps:

To further improve the classification performance and address current limitations such as class imbalance and overlapping feature distributions, several directions can be explored:

1. Experiment with more advanced models
   Extend beyond baseline Logistic Regression and Decision Tree by testing XGBoost, LightGBM, and potentially Ensemble Stacking models.These gradient-boosting frameworks can capture nonlinear interactions in timbre and loudness features more effectively and may increase recall on underperforming genres such as Pop, Soul/Reggae, and Hip-hop.
2. Handle class imbalance using resampling
   Apply SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic samples for minority genres (e.g., Hip-hop, Pop).This can help prevent models from being biased toward majority classes like Classic Pop/Rock and Folk.
3. Add SHAP analysis
   To better understand how individual features contribute to genre predictions, we will conduct a SHAP (SHapley Additive Explanations) analysis on our best-performing models (e.g., Random Forest, LightGBM).