Analysing Taylor Swift's Music: Exploring Album Popularity, Song Characteristics, and Predicting Top Hits

Gabriella Neilon[a]

[a]*Stellenbosch University*

Keywords: Taylor Swift, Random Forest Algorithm, Song Prediction

1. Background

This report focuses on Taylor Swift, one of the most influential and acclaimed artists of our time. With a remarkable career spanning across country and pop genres, Taylor Swift has left an indelible mark on the music industry, garnering 12 Grammy Awards and a massive fan base.The main objective of this analysis is to predict the top 5 singles from her latest album, *Midnights (3am Edition)*, which was released in October 2022. By analysing a combination of factors, including the album's overall popularity and the distinctive characteristics of each song, I aim to identify the tracks that have the potential make a significant impact on the charts. By delving into Taylor Swift's musical journey and employing predictive modeling techniques, this analysis offers insights and predictions regarding the potential success of her latest album's singles.

2. Data & Methodology

Data

The dataset used in this analysis is the "Taylor Swift Spotify" dataset obtained from *Kaggle*. It includes various attributes for each song, such as the release date, song length, popularity (represented as a percentage based on Spotify's algorithm), danceability (a measure of how suitable a track is for dancing), acousticness, energy (a measure of intensity and activity), instrumentalness (indicating the presence of vocals in the song), liveness (the probability of the song being recorded with a live

*Corresponding author: Gabriella Neilon
Email address: 22581340@sun.ac.za (Gabriella Neilon)

audience), loudness (the volume level of the music), speechiness (presence of spoken words in the track), valence (a measure of the song's emotional tone), and tempo (beats per minute).

To simplify the analysis, only the deluxe editions and "Taylor's Version" albums are selected. "Taylor's Version" refers to the music that Taylor Swift owns and has re-recorded. This selection helps to minimise potential bias in the data, considering that Taylor Swift's popularity has likely increased over time, and these albums contain more songs. By including these versions, the analysis levels the playing field among all the albums. Additionally, a target variable called "hit" was engineered using binary encoding. A song is coded as 1 if it was classified as a hit in the respective albums, based on the popularity scores, and 0 if not.

Methodology

In order to predict her next top hits on her newest album, I employed the *random forest algorithm*. The *random forest algorithm* was chosen to make the "next hit prediction" as it provides accurate predictions and efficiently handles large datasets.The strength of this model lies in its "social cognition" of individual decision trees that work together. Each decision tree is created by looking at different features or characteristics of the data and finding the best split at each step. When you want to make a prediction using the Random Forest, you ask all the decision trees for their opinions, and then take a majority vote. Each tree gets one vote, and the majority prediction becomes the final prediction of the Random Forest. This way, the Random Forest combines the strengths and insights of multiple decision trees to make a more accurate prediction.Even if some trees are wrong, the majority of the correct trees guide the collection to the right direction

The *random forest algorithm* consists of an ensemble of decision trees. The ensemble, called a 'forest,' is trained using bagging or bootstrap aggregating. Bagging is a technique that improves the accuracy of machine learning algorithms by combining their predictions. The algorithm makes predictions by averaging or taking the mean of the outputs from multiple decision trees. Increasing the number of trees enhances the precision of the predictions. Random forest overcomes the limitations of a single decision tree algorithm. It mitigates overfitting issues and improves accuracy. Decision trees are built based on entropy and information gain. Entropy measures uncertainty, while information gain quantifies the reduction in uncertainty of the target variable given independent variables. A higher information gain indicates a greater reduction in uncertainty (entropy). Entropy and information gain are crucial for splitting branches, a vital step in building decision trees. The key characteristic of the *random forest algorithm* lies in the random selection of root and splitting nodes. The random forest utilises "bagging" to generate predictions. Bagging involves using different samples of training data rather than a single sample. Each decision tree produces different outputs based on the training data fed into the *random forest algorithm*. These outputs are ranked, and the highest-ranked output becomes the final prediction.

Descriptive Statistics

In Figure 2.1, I present a comprehensive visual representation of Taylor Swift's discography, showcasing the popularity of each album and the emotions evoked by their most popular songs. To capture the emotional essence, I conducted feature engineering by categorizing songs as either "Happy" or "Sad" based on their valence score. Songs with a valence score below 0.5 are classified as "Sad," while those equal to or above 0.5 are deemed "Happy."

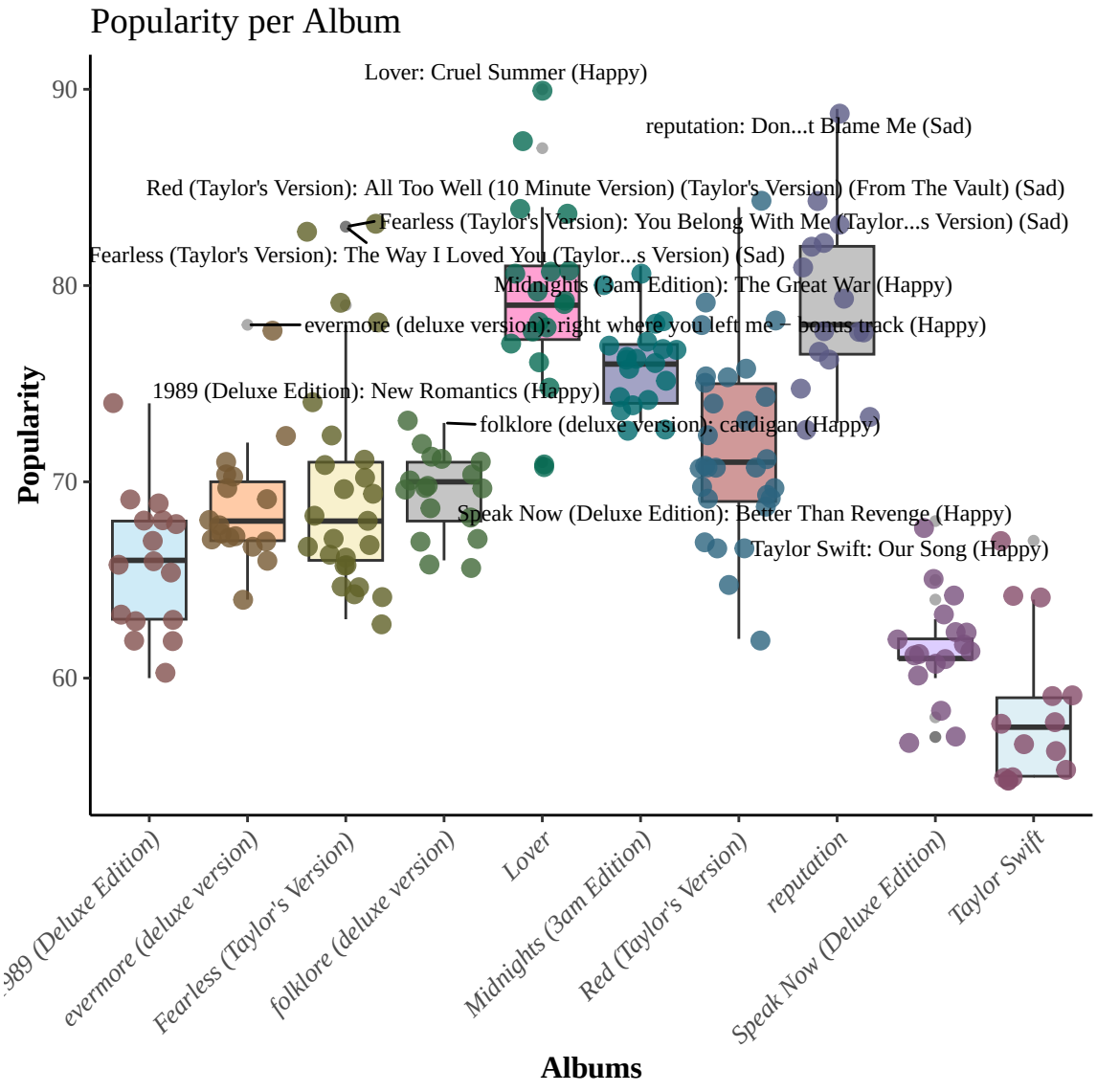

Figure 2.1: Popularity per Album

The results reveal that Taylor Swift's most popular album, as indicated by its overall popularity and the emotions it elicits, is "reputation". This is followed closely by the album "Lover" and her latest

release, “Midnights (3am Edition)”. On the other hand, her earlier albums, such as “Taylor Swift” and “Speak Now,” exhibit comparatively lower popularity. This observation suggests that popularity is influenced by the release date, with Taylor Swift’s later albums enjoying greater recognition, possibly reflecting her ascent to fame over the years.

The subsequent visual representation (Figure 2.2) explores the shared characteristics found among the top 5 most popular songs from each of Taylor Swift’s albums. By identifying these common traits, I further investigate the composition of the albums using these influential characteristics. This analysis is presented in Figure 2.5.

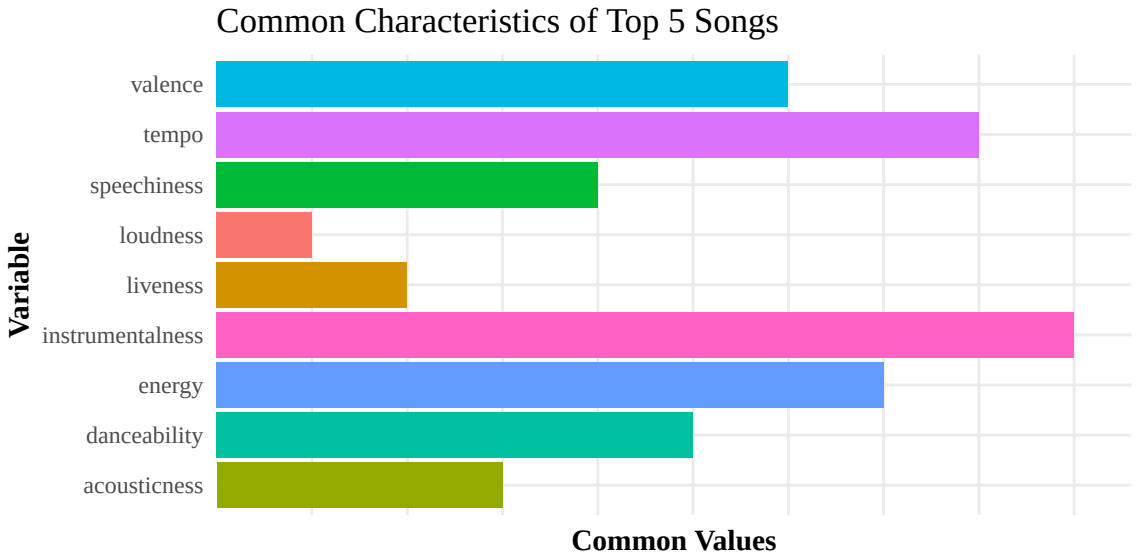

Figure 2.2: Common Characteristics of Top 5 Songs

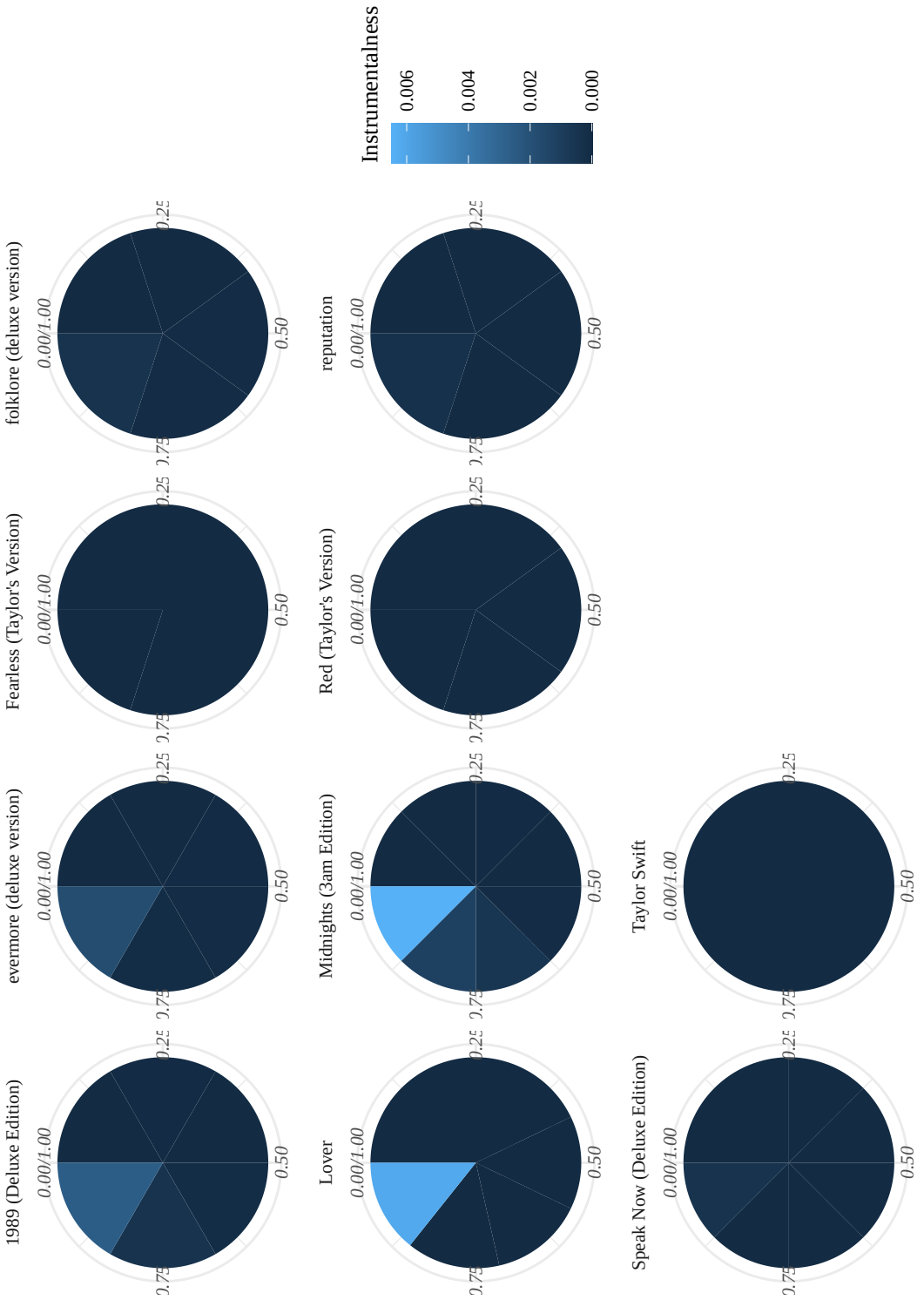

Figure 2.3: Common Characteristics of Top 5 Songs Concentration

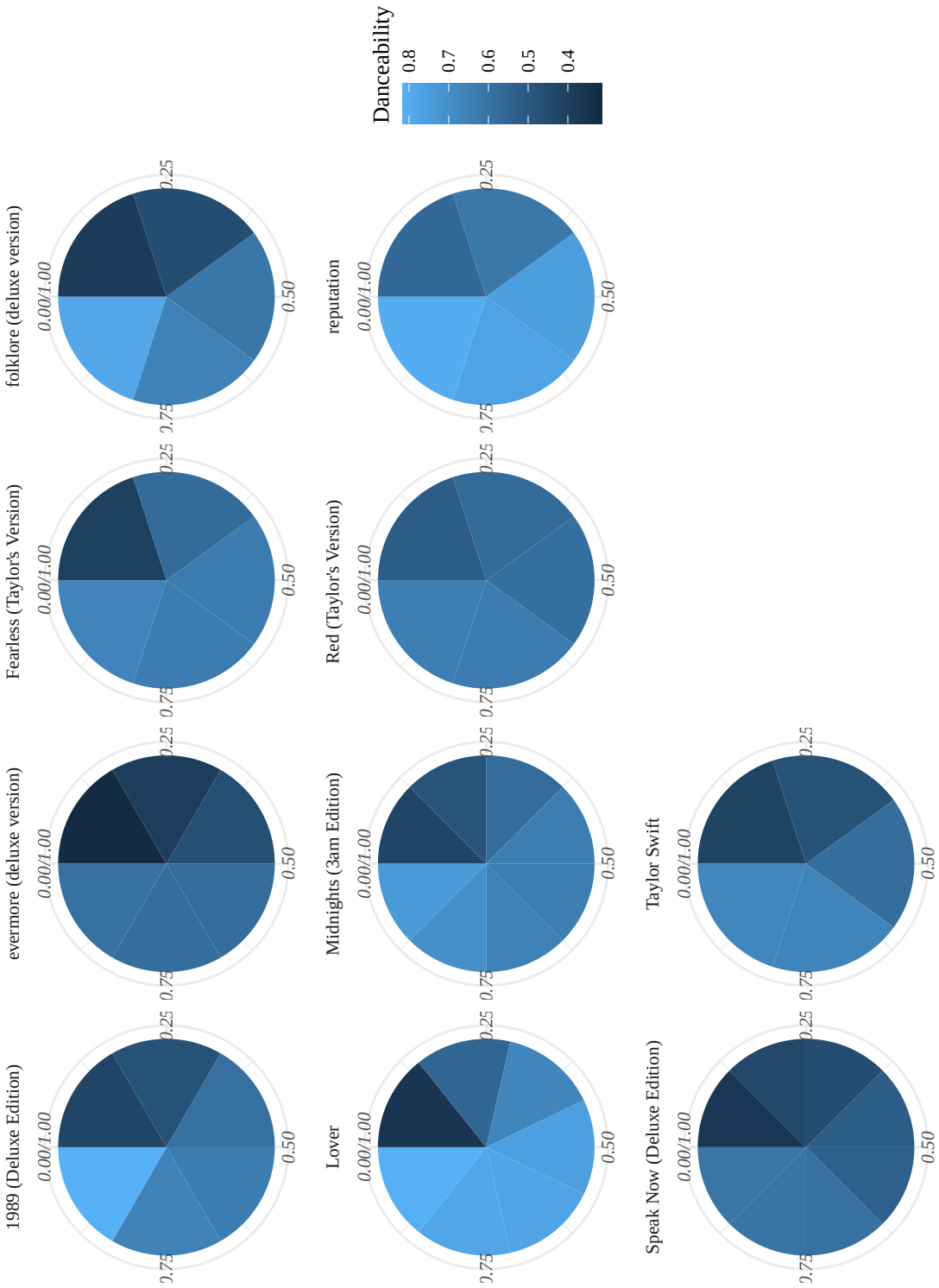

Figure 2.4: Common Characteristics of Top 5 Songs Concentration

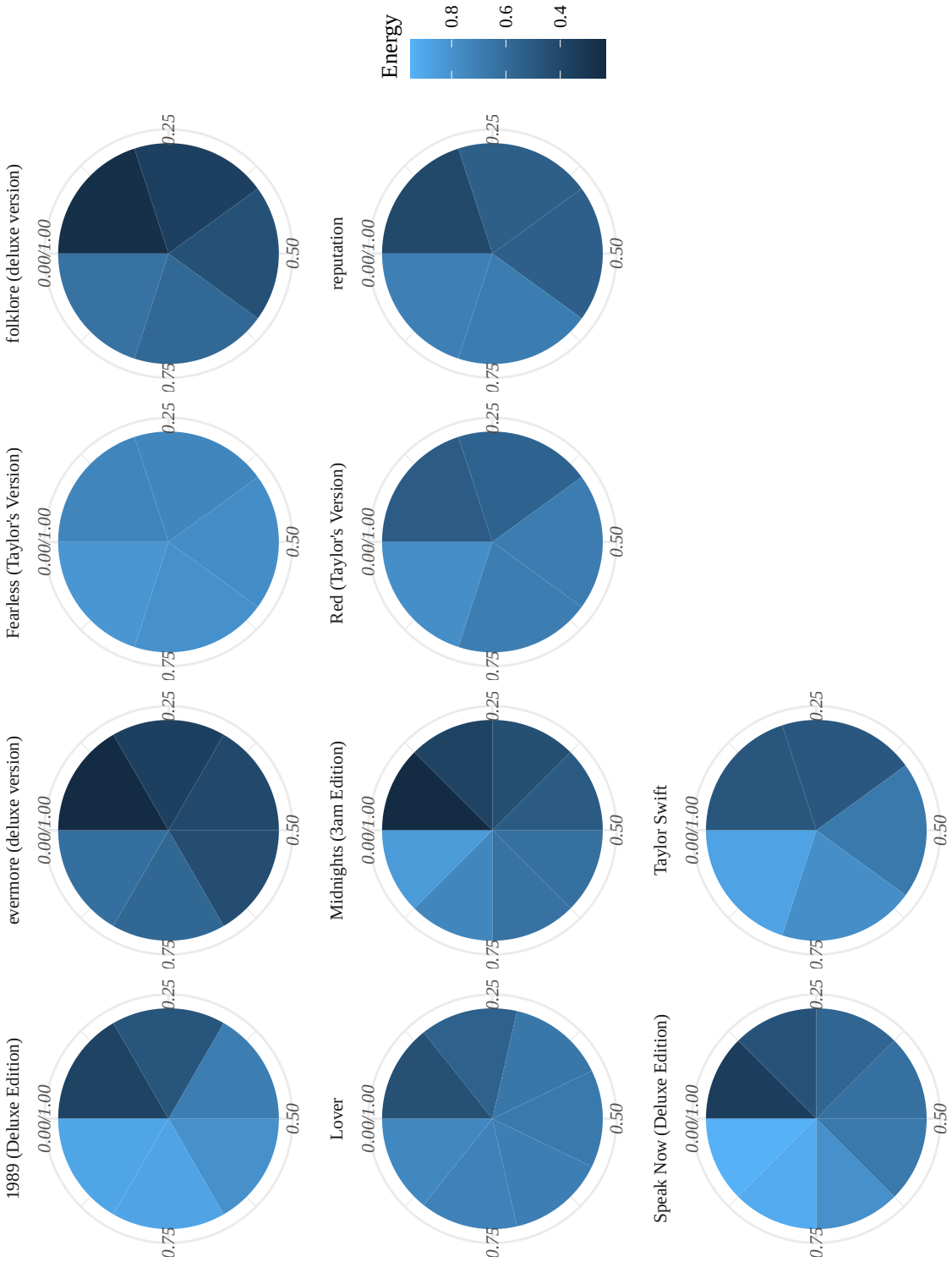

Figure 2.5: Common Characteristics of Top 5 Songs Concentration

Generally, Taylor Swift's albums exhibit relatively low levels of *instrumentalness*, except for the albums "Lover" and "Midnights (3am Edition)," which have already established themselves as among her most popular works. This suggests that instrumental elements play a significant role in the appeal and success of these particular albums. In Figure 2.1, it is observed that albums with lower overall popularity tend to exhibit lower levels of *energy*, while those with higher popularity showcase higher levels of *danceability*. These findings highlight the connection between specific musical attributes and the overall reception of Taylor Swift's albums.

3. Results

Baseline Model

This model employs the default parameters, with the test data representing Taylor Swift's latest album, "Midnights (3am Edition)," while the train data encompasses all of her previous works leading up to the release of her newest album.

Table 3.1: Baseline Model

Song Name
The Great War
Would've, Could've, Should've
Maroon
Snow On The Beach (feat. Lana Del Rey)
High Infidelity

Based on this model, the predicted top 5 singles ("hits") for her next releases are anticipated to be *Maroon*, *Snow On The Beach (feat. Lana Del Rey)*, *The Great War*, *Would've, Could've, Should've*, and *Anti-Hero*.

Hypertuned Model

In this model, I aim to find the optimal configuration for predicting the next top 5 "hits" by exploring different combinations of hyperparameters based on the lowest Root Mean Square Error (RMSE). From this, at each split, the model randomly selects 10 variables to determine the best split. Secondly, I ensure that each terminal node in the decision trees contains at least one observation. Furthermore, I use sampling with replacement, allowing each bootstrap sample used for training individual trees to contain duplicate observations. Additionally, I train each tree on a randomly sampled subset of the

training data, comprising 50% of the observations. This random sampling process contributes to the creation of diverse trees and helps reduce correlation among them. Finally, I set a random seed of 123 to ensure the reproducibility of the results. This ensures that when the model is re-run, the same random seed is used, resulting in consistent outcomes. After tuning the hyperparameters, the resulting "best model" achieves a prediction error of 0.171, as indicated by the RMSE. This performance metric assesses how accurately the model predicts the next top 5 hits based on the chosen configuration of parameters.

Table 3.2: Hypertuned Model

Song Name
Snow On The Beach (feat. Lana Del Rey)
Maroon
Would've, Could've, Should've
The Great War
Anti-Hero

After fine-tuning the hyperparameters, the resulting "best model" achieves a prediction error of 0.171, as indicated by the RMSE. The RMSE serves as a performance metric, assessing the accuracy of the model in predicting the next top 5 hits based on the selected parameter configuration. Based on this model, the predicted top 5 singles ("hits") for her next releases are anticipated to be *Snow On The Beach (feat. Lana Del Rey)*, *Maroon*, *Would've, Could've, Should've*, *The Great War*, and *Anti-Hero*.

Feature Interpretation

In my analysis, I explore two methods to assess the importance of variables: *impurity-based and permutation-based variable importance.*The *impurity-based approach* measures the importance of a feature by evaluating how much the randomness in predictions is reduced when splitting on that feature. It uses the impurity measure, such as entropy, to quantify the disorder or lack of purity in a set of samples. The underlying assumption is that an important feature will lead to more effective splits, reducing impurity and improving prediction accuracy.

On the other hand, *permutation-based importance* also estimates feature importance by evaluating the decrease in model performance when the values of a feature are randomly permuted. This approach provides a direct measure of a feature's contribution to the predictive power of the model. It takes into account the impact of the feature within the entire model, providing valuable insights.

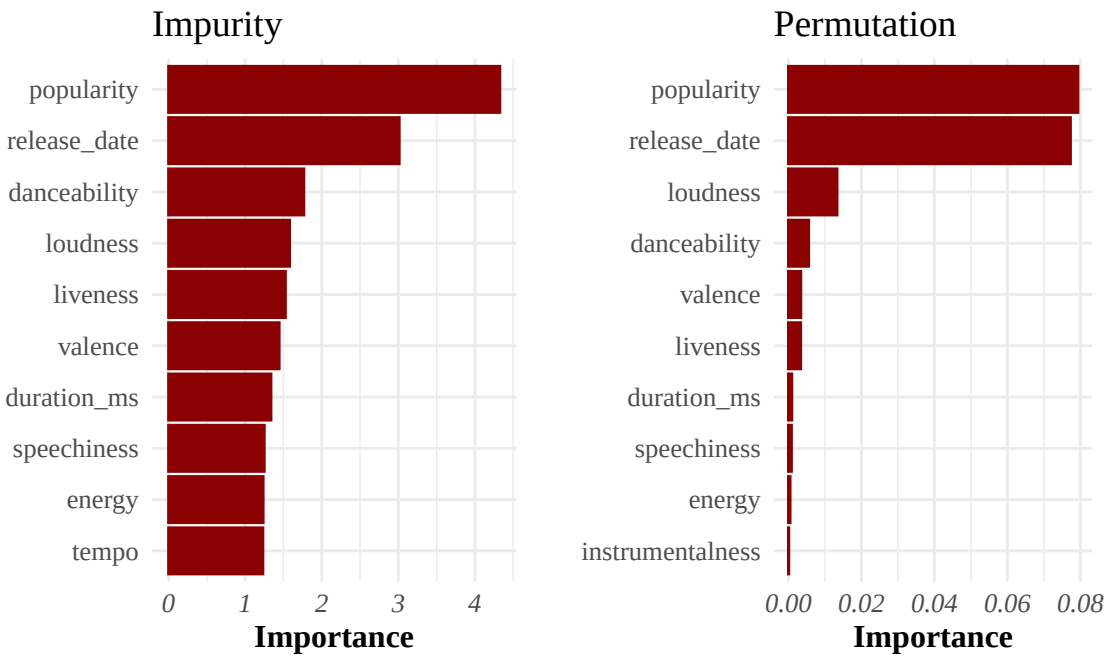

Figure 3.1: Impurity-based and Permutation-based Performance

From both the *impurity-based and permutation-based importance* analyses, certain characteristics stand out as significant. These include *popularity, release date, loudness, and danceability.* These findings align with my previous hypothesis that Taylor Swift's next album's hits will be influenced by her growth in fame over the years. Since "hit" classification is based on popularity, it makes sense that the next hits will be closely linked to the songs' popularity. Additionally, consistent with the descriptive statistics from Figure 2.5, danceability emerges as a key feature in her most popular albums.

Model Performance

The area under the ROC curve (AUC) is a widely used metric for assessing the overall performance of the model. It quantifies the model's ability to assign higher scores to positive instances than negative instances. An AUC value of 0.5 suggests a random classifier, meaning the model performs no better than random guessing. Conversely, an AUC value of 1 indicates a perfect classifier, where the model perfectly distinguishes between positive and negative instances.

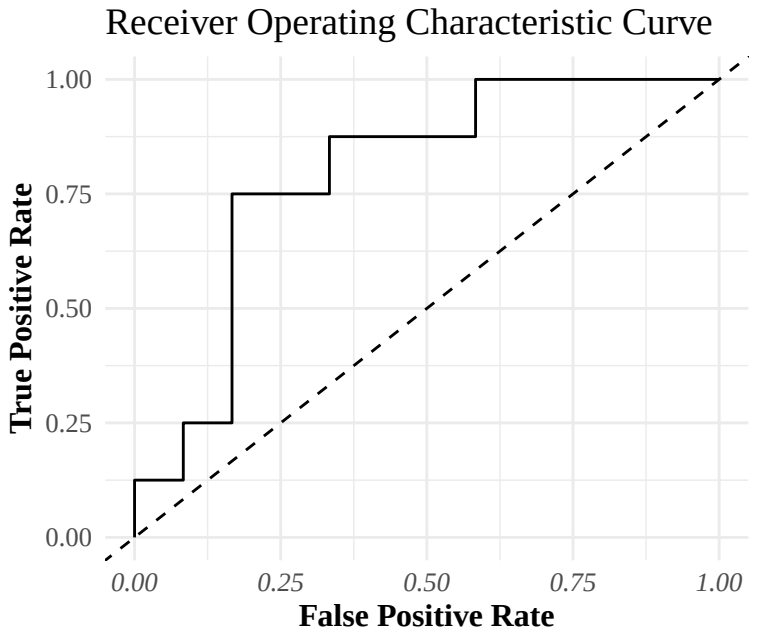

Figure 3.2: Receiver Operating Characteristic Curve

The hypertuned *best model* achieved an AUC value of 0.7917, which indicates that its performance surpasses that of a random classifier. This means that the model is highly proficient in making accurate predictions and effectively distinguishing between positive and negative instances in the classification task at hand. The high AUC score demonstrates the model's strong discriminatory power and its ability to provide reliable classifications. Consequently, we can conclude that the model is an effective classifier.

4. Conclusion

Based on the analysis performed, it can be concluded that the model has shown success in predicting the next top 5 singles from Taylor Swift's discography. The evaluation of the model's performance using the ROC curve revealed that it outperforms a random classifier, indicating its capability to make accurate predictions and serve as a reliable classifier. Although there is a slight difference in the ordering of the songs between the baseline model and the hypertuned model, both models predict the same top 5 songs. It is noteworthy that the song "Anti-Hero" was indeed Taylor Swift's first single from her newest album, which highlights the model's ability to capture upcoming hits. Furthermore, while the remaining songs predicted by the model have not been officially announced as singles, their inclusion in the top 5 aligns with the preferences of the devoted Taylor Swift fan community, known as "Swiftie". As a dedicated "Swiftie" myself, I can affirm that these songs hold high favourability among "Swiftie" enthusiasts based on their online discussions and interactions.