How Weather Affects Music Preferences: Analyzing the Correlation Between Top Spotify Songs and UK Weather Data

Guettabi, Tima
fg6@williams.edu
CS and Psychology
&
Bah, Maymouna
mb33@williams.edu
CS and Africana Studies

1. Introduction

In the field of data-driven exploration, this project delves into the connection between the top monthly Spotify songs in the UK and the weather conditions from January 2000 to June 2020. Our goal is to discover potential correlations between atmospheric nuances and musical preferences, seeking to expand the understanding of what factors contribute to the flow of popular songs.

The motivation for this endeavor emerges from the recognition that music, as a dynamic expression, is not solely influenced by personal taste but also by external factors. While existing literature on music preferences has extensively explored the impact of personality traits (Vella and Mills, 2017), political orientation (Christenson and Peterson, 1988), culture, socio-economics, and friendship (Liu et al., 2018), the specific influence of weather conditions remains a relatively unexplored domain. By undertaking this project, we aim to bridge this gap, adding a new layer to the intricate web of factors shaping the musical landscape.

Our hypothesis posits that individuals exhibit distinct music preferences corresponding to the seasons, influenced by both seasonal affective states and atmospheric conditions. During the winter season, which is characterized by gloomy weather and a potential increase in seasonal depression, individuals are expected to gravitate towards melancholic and introspective genres, seeking solace in the emotional resonance of sad music. Conversely, as spring heralds the end of winter and the onset of brighter days, people are hypothesized to shift towards more joyful and uplifting music, reflecting the positive impact of changing weather on mood. In the summer, with the prevalence of sunny days and outdoor activities, individuals are predicted to favor upbeat dance music, aligning with the energetic and celebratory ambiance of the season. As autumn arrives and the weather transitions, the hypothesis suggests a preference for more romantic and reflective music, mirroring the contemplative mood associated with fall. This hypothesis aims to capture the emotional dynamics that shape seasonal music preferences throughout the year.

To achieve our goal, we merged two datasets. The first, extracted from Chart2000, encompasses the top 50 Spotify songs each month in the UK (Chart2000). This dataset serves as a representative chronicle of collective musical preferences over two decades. The second dataset, sourced from Kaggle ("UK Weather by Month") (Button, 2023), provides a detailed account of weather conditions, including temperature, rainfall, and sunlight duration, during the same period.

Understanding the potential correlations between weather patterns and musical trends contributes not only to the empirical understanding of music preferences but also to the broader discourse on how environmental factors shape cultural phenomena. This analysis aligns with the growing body of interdisciplinary research, where data illuminates the hidden connections between seemingly disparate elements.

We will delve deeper into our hypothesis and goal in subsequent sections. Our goal is to create a model that predicts a classification output for whether or not a song is popular (in the top 50), and the model uses various weather and song features to predict this. Our hypothesis specifically states that individuals exhibit distinct music preferences that correspond to the seasons. Essentially, their musical choices throughout the year are shaped and influenced by both seasonal affective states and atmospheric conditions. In the upcoming sections, we will also detail the methodologies applied, present the results of our exploration, and engage in a comprehensive discussion about the profound findings that have emerged from this fascinating intersection of meteorology and musicology.

2. Preliminaries

In our initial modeling attempts, we started with **Logistic Regression** as a basic algorithm to predict song popularity based on various song features and weather conditions. Logistic Regression, known for its simplicity and interpretability in binary classification tasks, was chosen for its ease of understanding. However, as our preliminary analyses unfolded, it became apparent that the relationship between the features and the target variable was more intricate and potentially nonlinear.

Before we explain our shift in strategies, Let's delve into the technicalities of logistic regression. Logistic regression is particularly useful when dealing with binary classification problems, as is the case in our analysis where we aim to predict whether a song makes it to the top 50 or not. The logistic regression model transforms a linear combination of input features into a probability score using the sigmoid function.

For a given example with input features $x_1, \ldots, x_d$, the logistic regression prediction is given by:

$$\hat{p} = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d)}}$$

where $\theta_0, \theta_1, \ldots, \theta_d$ are the learned parameters. The logistic regression model then classifies an example as belonging to the positive class (e.g., a song making it to the top 50) if $\hat{p}$ is greater than a chosen threshold. For our analysis, we set $\hat{p} = 0.5$. The choice of a 0.5 threshold in logistic regression is common as it strikes a balance in binary classification, since the model estimates the probability of an instance belonging to the positive class (e.g., class 1), producing a probability score between 0 and 1.

Applying a 0.5 threshold means that if the predicted probability is 0.5 or higher, the instance is classified as the positive class. Otherwise, it is classified as the negative class (popularity class 0). This specific threshold is preferred because it minimizes the misclassification rate, assuming equal importance for false positives and false negatives.

The parameters θ are learned by maximizing the likelihood function, which measures how well the model explains the observed data. The optimization process involves finding the values of θ that maximize the likelihood, often performed using iterative methods like gradient descent.

To avoid overfitting, a regularization term is commonly incorporated. In our case, we employ the L2 penalty, adding a term to the loss function that penalizes large parameter values. The overall function being minimized is expressed as:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\right) + \lambda\sum_{j=1}^{d}\theta_j^2$$

where y_i is the actual class label for the i-th example, $\hat{p}_i$ is the predicted probability, and λ is the regularization hyperparameter tuned using a validation set. This logistic regression framework enables us to model the probability of a song being in the top 50 based on the associated weather conditions.

The inherent limitations of Logistic Regression, particularly its challenge in handling complex interactions and non-linear patterns, prompted a reassessment of our modeling approach. Our objective was to construct a model capable of capturing the nuanced dependencies within our dataset, encompassing an array of song characteristics and weather parameters. To meet this objective, we transitioned to the application of Random Forests.

The Random Forest algorithm, renowned for its effectiveness, belongs to the category of bagged (Bootstrap Aggregating) algorithms. In essence, a Random Forest comprises multiple decision trees, each obtained through bagging, with a slight modification in the splitting criteria. To comprehend its operation, consider a dataset $[D]$ containing n rows and d features.

The algorithm unfolds as follows:

1. **Data Sampling:** Multiple datasets $[D]_1, [D]_2, \ldots, [D]_m$ are sampled, each of size n, from $[D]$ with replacement. This process, known as bootstrapping, creates diverse subsets for training individual trees.

2. **Tree Training:** For each sampled dataset $[D]_i$, a full decision tree $h[D]_i$ is trained. Here's the key modification: before each split, a random subset of k (where $k \leq d$) features is chosen without replacement, and only these features are considered for the split. This introduces variability in the tree-building process.

3. **Aggregation:** The final model results from aggregating predictions from the m individual trees. In classification tasks, this aggregation often involves taking the mode, while for regression tasks, it entails calculating the mean. This aggregation process ensures a robust and generalized predictive model.

This approach, incorporating elements of randomness in both data sampling and feature selection, contributes to the strength of Random Forests. The ensemble of diverse decision trees, each trained on a different subset, collectively enhances the algorithm's ability to handle complex relationships within the data, mitigate overfitting, and provide accurate predictions for new, unseen instances.

For context, a decision tree is a hierarchical structure that makes decisions based on a set of conditions. In our case, these conditions would include features like song danceability and sunlight. Decision trees are adept at capturing non-linear relationships and interactions within complex datasets. However, a single decision tree may suffer from high variance and instability. Random Forests address this issue by aggregating the predictions of multiple trees, yielding a more reliable and accurate outcome.

Random Forests offer several advantages tailored to our specific use case. First, they effectively handle high-dimensional datasets, accommodating the numerous features derived from both song-related and weather-related information. Second, the ensemble nature of Random Forests allows them to capture intricate patterns and dependencies in the data, essential for discerning the multiple factors influencing song popularity.

Given the complexity of our dataset, Random Forests emerged as a suitable choice, striking a balance between flexibility and stability. Their ability to provide insights into feature importance enhances interpretability, offering a clearer understanding of the influential factors contributing to song popularity. In essence, Random Forests serve as a robust and versatile tool for predicting song popularity in our project, leveraging ensemble learning to navigate the complexities of our dataset.

In our modeling process, the hyperparameters of the Random Forest model, such as the number of trees and tree depth, were carefully tuned to optimize predictive performance. The best performance we have obtained corresponds to the following parameters: n_estimators=3000, max_depth=100, max_features=sqrt.

(we imported the Random Forest Model from the scikit-learn library (Pedregosa et al., 2011))

3. Baseline

Initially, as we embarked on understanding the relationship between weather conditions and song popularity, we planned on using a logistic regression model. For that model, our baseline aimed to reflect a scenario of random guessing, where the probability of predicting song popularity is akin to a 50/50 chance. However, as we will discuss further in a later section, we discovered that Logistic Regression performed poorly against this baseline, so we opted for a more sophisticated approach.

To better understand this relationship, we then chose to implement a Random Forest model. For this model we implemented Logistic Regression with gradient clipping as our baseline model for comparison. The rationale behind choosing logistic regression as a baseline lies in its simplicity, interpretability, and its widespread use in binary classification problems. The addition of gradient clipping addresses specific challenges encountered during the training process. Logistic Regression minimizes a loss function using gradient descent.

Gradient clipping ensures controlled parameter updates, promoting stable optimization and preventing extreme changes in the model parameters.

Despite its simplicity, Logistic Regression serves as an effective baseline for comparison. It allows us to establish a foundation for understanding how well more complex models, like Random Forests, perform in comparison. Logistic Regression, although linear, can capture basic relationships between features and the target variable.

The inclusion of gradient clipping addresses potential challenges in dealing with intricate relationships between weather features and song popularity. Gradient clipping is crucial to prevent issues like exploding gradients during training which we did face while training our model. Logistic Regression, minimizing a loss function through gradient descent, benefits from gradient clipping to ensure controlled parameter updates and stable optimization.

This baseline is chosen to address intricate relationships between features and target variables, aligning with our goal of establishing a stable baseline for comparison with more complex models. Subsequent sections will delve into the outcomes of these baseline models, guiding our modeling choices. This baseline choice enables a clear comparison with more sophisticated models, shedding light on the additional predictive power gained from the complexity of ensemble models like Random Forests.

4. Data

In this section, we outline the data sources utilized in our project and the pre-processing steps undertaken to ensure the quality and relevance of the datasets. We used seaborn (Waskom, 2021) and MATLAB (MATLAB, 2010) libraries to plot and visualize our data. We also used the pandas library for data manipulation (McKinney et al., 2010).

4.1 Weather Data

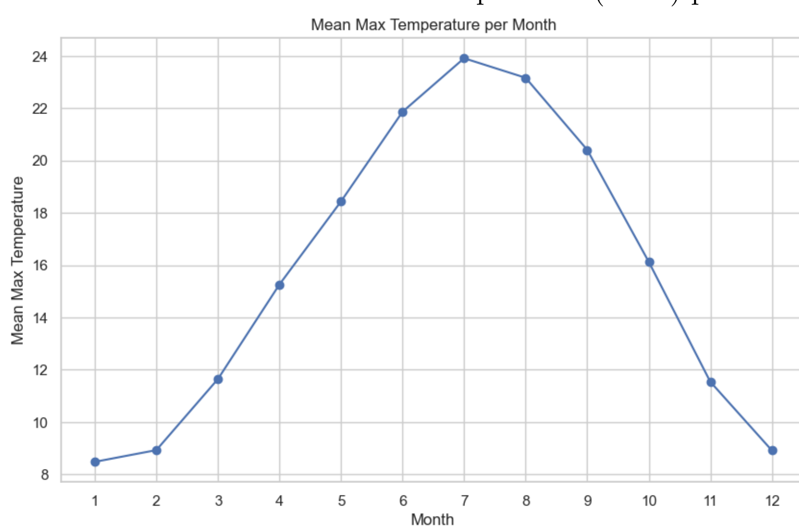
The weather data was obtained from the 'met-office-weather-month.csv' file, which was imported into a Pandas DataFrame named `weather_data`. To narrow the focus to relevant years for our analysis (2000 to 2020), we filtered the data accordingly. Recognizing 'Heathrow' as a representative station for UK weather, we isolated data related to this station, dropping the 'station' and 'decade' columns. The resulting DataFrame, denoted as `weather_data`, provides a detailed account of weather conditions in the UK, specifically near London. The resulting data set has the following columns:

- tmax: Mean daily maximum temperature (°C)
- tmin: Mean daily minimum temperature (°C)
- af: Number of days of air frost
- rain: Total rainfall (mm)
- sun: Total sunshine duration (hours)

Figure 1: Head of the weather_data Dataset

	year	month	tmax	tmin	af	rain	sun
1126	2000	1	8.6	2.4	7.0	16.5	78.6
1127	2000	2	10.4	3.8	5.0	62.2	102.5
1128	2000	3	12.1	4.9	1.0	16.0	120.4
1129	2000	4	12.9	5.4	0.0	99.6	135.8
1130	2000	5	18.0	9.6	0.0	87.2	202.9
...	...	...	...	...	...	...	...
1367	2020	2	11.1	4.3	1.0	99.8	62.0
1368	2020	3	12.0	3.9	1.0	42.8	148.0
1369	2020	4	18.2	6.5	1.0	38.2	235.4
1370	2020	5	21.1	9.1	0.0	2.0	308.6
1371	2020	6	22.5	12.6	0.0	54.0	174.9

Figure 2: Visualization of the mean Max Temperature (tmax) per month in 2017



4.2 Song Data

The song data was sourced from the 'chart2000-songmonth-0-3-0070.csv' file and imported into a DataFrame named `song_data`. Irrelevant columns such as 'indicativerevenue', and country-specific columns ('us', 'uk', 'de', 'fr', 'ca', 'au') were dropped. Missing values were identified and subsequently removed by dropping any rows with NaN entries. The 'month' column which had originally both the year and month (MM/YYYY format) was handled by splitting it into separate 'month' and 'year' columns for consistency and merging compatibility. To ensure seamless merging of the song and weather datasets, a dictionary was created to map month names to numerical values, and the 'month' column was updated accordingly. To ensure alignment between the song and weather datasets, we truncated the song data to end in June 2020, removing the last 150 rows to match the temporal scope of the weather data.

Figure 3: Head of the song_data Dataset

	position	artist	song	month	year
0	1	Rob Thomas & Santana	Smooth	1	2000
1	2	Christina Aguilera	What A Girl Wants	1	2000
2	3	Savage Garden	I Knew I Loved You	1	2000
3	4	Celine Dion	That's The Way It Is	1	2000
4	5	Eiffel 65	Blue (Da Ba Dee)	1	2000
...	...	...	...	...	...
12445	46	Surf Mesa & Emilee	ILY (I Love You Baby)	9	2020
12446	47	Lil Mosey	Blueberry Faygo	9	2020
12447	48	Kane Brown	Cool Again	9	2020
12448	49	BLACKPINK & Selena Gomez	Ice Cream	9	2020
12449	50	Future & Drake	Life Is Good	9	2020

4.3 Merging Datasets

The weather data `song_data` and song data `weather_data` were merged based on the 'year' and 'month' columns, creating a comprehensive dataset. Next, for our analysis, we obtained song features ('Danceability', 'Energy', 'Key', 'Loudness', 'Speechiness', 'Acousticness', 'Instrumentalness', 'Liveness', 'Valence', 'Tempo') using Exportify (link). Then the merged weather-song dataset was further combined with the song_features dataset based on the 'song' column. To ensure consistency in text matching, the 'song' column was converted to lowercase. The song features are described as follows:

- Danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- Energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.
- Key: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation .
- Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.
- Speechiness: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audiobook, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music.Values below 0.33 most likely represent music and other non-speech-like tracks.

- Acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

- Instrumentalness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

- Liveness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.

- Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

- Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.

To read more about what each song feature means, please refer to this link.

4.4 Ranking and Popularity Determination:

Initially, the classification approach aimed to categorize the top 50 songs of each month. Thus, we added a popularity column to our dataset. The top ten songs per month were labeled with popularity = 1, and the remaining 40 with popularity = 0. This method highlighted a significant challenge – class imbalance. The unequal distribution of popular and less popular songs affected the model's learning. (10 popular songs vs 40 unpopular songs per month). To address the class imbalance, our strategy shifted. The focus narrowed to the extremes, keeping only the top ten songs (popularity = 1) and the last ten songs (popularity = 0) each month. This refined approach tackled class imbalance effectively (2427 data points with popularity = 0 and 2288 with popularity = 1). Songs in between the top ten and last ten for each month were dropped, creating a more balanced dataset for improved model performance.

4.5 Saving Merged Data

The resultant merged DataFrame, representing the amalgamation of weather data, the top 10 songs, the last 10 songs per month, and each song's features, was saved to a CSV file named `top10bestandworst2.csv`. This file serves as the input data for our subsequent analyses and machine-learning modeling. We later drop the columns: "Spotify ID", "Artist IDs", "Genres", "Artist Name(s)" from the dataframe as they are not useful to our machine learning model training process.

5. Training And Validation Of Models

In alignment with best practices for dataset partitioning, we opted for an 80-10-10 split among the training, validation, and testing sets. Specifically, we organized our data into three distinct temporal intervals:

- **Training Set (80%):** The years spanning from 2000 to 2016 were allocated to the training set. This substantial portion of the dataset provided the foundation for the models to learn patterns and relationships.
- **Validation Set (10%):** The years 2017 to 2018 were designated for the validation set. This intermediary set served as a crucial checkpoint during the training process, allowing us to fine-tune model parameters and identify the optimal depth for the decision trees.
- **Testing Set (10%):** The most recent years, from 2019 to 2020, were reserved for the testing set. This independent dataset served as the final evaluation ground for assessing the generalization performance of our trained models.

The rationale behind this temporal split lies in ensuring that the models are trained on historical data, validated on a subsequent period to gauge adaptability, and ultimately tested on the most recent data to assess real-world applicability. We want to assess the model's capacity to generalize and perform well across different time periods. The temporal split of data into training, validation, and testing sets is designed to ensure that the model is exposed to historical data during training, allowing it to learn patterns and trends. The validation set, representing a subsequent period, serves as a check on the model's adaptability by assessing its performance on data it hasn't seen before. Finally, the testing set, comprised of the most recent data, provides a real-world evaluation of the model's applicability to new and unseen information.

This strategic division of data into training, validation, and testing sets adheres to the 80-10-10 principle, facilitating a robust evaluation of our models' performance. The subsequent sections will delve into the specifics of the training methodologies, validation procedures, and the outcomes derived from this comprehensive approach.

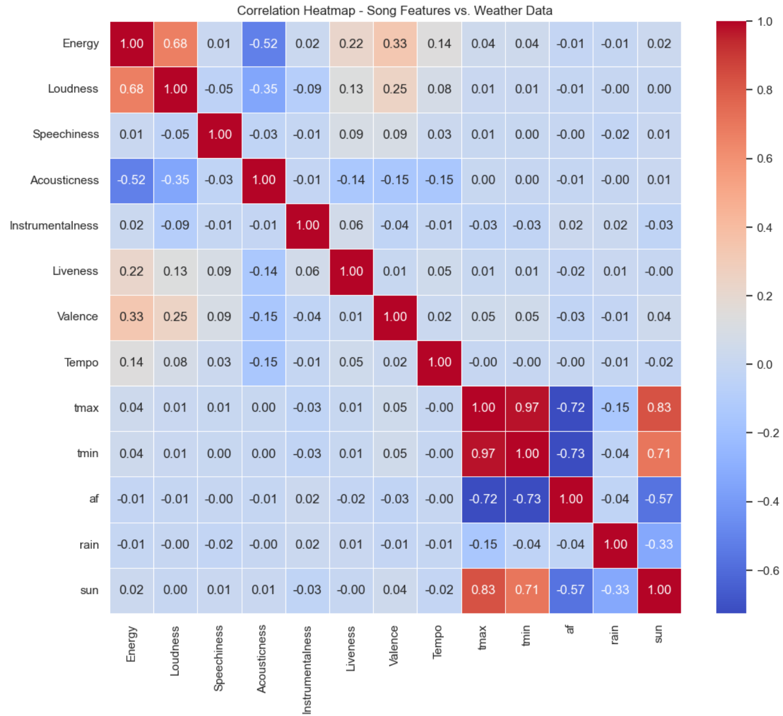
6. Justification of Model choice using Visualization:

For this project, our two potential models were a Logistic Regression model with Gradient Clipping and a Random Forest model. After looking at various visualizations and comparing models against the baseline, we opted to use the Random Forest Model.

6.1 Correlation Heatmap

Examining the figure 4, it becomes evident that the correlation between song features and weather data is either weak or virtually nonexistent. Choosing Random Forest over Linear Regression is justified when dealing with data characterized by weak correlations between features. Linear Regression assumes a linear relationship between the input features and

Figure 4: Correlation Heatmap - Song Features vs. Weather Data



the target variable. If the data exhibits weak or non-linear correlations, Linear Regression might struggle to capture the underlying patterns, leading to suboptimal predictions.

On the other hand, Random Forests are well-suited for capturing complex, non-linear relationships in the data. The ensemble nature of Random Forests, which aggregates predictions from multiple decision trees, allows them to adapt to intricate patterns and dependencies. Each tree in the ensemble can focus on different aspects of the data, collectively contributing to a more comprehensive understanding of the underlying relationships.

In this scenario where feature correlations are weak or non-linear, Random Forests offer us flexibility and robustness, making them a favorable choice. They can better discern subtle interactions and dependencies that might be overlooked by a linear model like Linear Regression, leading to more accurate predictions in such cases.

6.2 Comparing Model Against Baseline

In comparing the performance of our Logistic Regression model against the established baseline of random guessing, we observe clear indications of its limitations:

1. Random Baseline
    - Validation Accuracy: 0.52165
    - Test Accuracy: 0.42775
2. Logistic Regression Model

- Validation Accuracy: 0.49134
- Test Accuracy: 0.46820

The random baseline, reflecting a scenario where predictions are made by chance, provides a baseline accuracy of approximately 52.2% for validation and 42.8% for the test set.

In stark contrast, the Logistic Regression model, designed to capture relationships between features and song popularity, falls short in both validation and test accuracy. The model achieved a validation accuracy of 49.1% and a test accuracy of 46.8%, indicating that it performs marginally better than random guessing.

The Logistic Regression model, despite its simplicity, struggles to surpass the predictive capacity of a basic random guess. This suggests that the linear nature of Logistic Regression may not capture the complexities inherent in the relationships between weather conditions and song popularity. The underperformance against a random baseline highlights the necessity of employing more sophisticated models. Ensemble models like Random Forests, known for their ability to capture intricate relationships and non-linear patterns, become crucial in improving predictive accuracy.

In the subsequent sections, we explore the outcomes of the Random Forest model to understand how they address the shortcomings of Logistic Regression and offer superior predictive capabilities in our context.

7. Results

To assess our model's performance, we compare their accuracy metrics with our previously established baseline model. In our case, the baseline is akin to randomly guessing a song popularity, where the probability of correctly predicting whether a song is popular or not is essentially a coin flip. Looking at the different models we trained, let's see how their performance is compared to the baseline.

7.1 Model Performance:

For the Random Forest model, the baseline being used is the Logistic regression with gradient clipping model.

Validation accuracy is a metric used to assess the performance of a machine learning model on a set of data that it hasn't seen during the training process. On the other hand, test accuracy is a measure of a model's performance on a completely independent data set that it has never encountered before. This provides an indication of how well the model generalizes to new, unseen data.

1. **Logistic Regression:**
    - Validation Accuracy: 0.49134
    - Test Accuracy: 0.46820

2. **Random Forest:**

   - Validation Accuracy: 0.64935
   - Test Accuracy: 0.71098

7.2 Comparison

For the random forest model, the validation accuracy was 0.64935, which is significantly better than the validation accuracy of the logistic regression model. The Random Forest model, with a validation accuracy of 64.9%, demonstrates a more substantial improvement over logistic regression. This indicates that the model has learned patterns in the data and is making predictions that are significantly better than the baseline. The test accuracy of this model was 0.71098, which is considerably better than the test accuracy of the Logistic Regression model. The Random Forest model's test accuracy of 71.1% is considerably higher than the baseline, showcasing its greater ability to generalize and make accurate predictions on new and unseen data.

7.3 Interpretation

A model with higher validation accuracy is generally considered more robust, indicating that it is learning meaningful patterns from the training data. Additionally, the test accuracy, is a metric that is crucial for evaluating how well a model can make accurate predictions on new, previously unseen data. A higher test accuracy implies better generalization.

In our case, the Random Forest model, with its significantly better validation and test accuracy compared to Logistic Regression, suggests that it is more adept at capturing the underlying patterns in the data and making accurate predictions on both seen and unseen instances. These metrics provide insights into the effectiveness of our models in predicting song popularity based on the chosen features and weather conditions.

7.4 Implications

The performance improvement, especially with the Random Forest model, indicates that there are patterns and relationships in the data, whether or not they are weak patterns, that contribute to predicting song popularity. This lends credence to the idea that factors beyond chance, such as intrinsic song characteristics, play some sort of a role in shaping musical preferences.

Thus, the Random Forest model shows a significant enhancement over the baseline, indicating its ability to better discern patterns in the data. When using the Logistic Regression model as a baseline, we can then say that the Random Forest model is a "success."

In the subsequent sections, we will further explore the insights gained from our models and discuss the broader implications of these findings in the context of cultural phenomena and the influence of environmental factors on music preferences.

8. Ablation Study

8.1 Class imbalance

The ablation study focused on addressing class imbalance by refining the dataset to only include the top ten and last ten songs per month, effectively creating a more balanced representation for training. The top 50 songs of each month were labeled, with the top 10 songs assigned a label of popularity=1, and the remaining 40 songs labeled with popularity=0. This method resulted in a significant class imbalance, with only 10 popular songs compared to approximately 40 unpopular songs per month. To fix this, we concentrated solely on the extremes—retaining the top ten songs (popularity = 1) and the last ten songs (popularity=0) each month. Hence, songs falling between the top ten and last ten for each month were excluded, contributing to the overall balance of the dataset.

Before this adjustment, the Random Forest model exhibited the following performance metrics with class imbalance:

- **Validation Accuracy:** 0.63722
- **Test Accuracy:** 0.17341

After addressing the class imbalance using the refined approach, the model performances were as follows:

- **Validation Accuracy:** 0.64935
- **Test Accuracy:** 0.71098

8.1.1 Observations:

The Random Forest model showed improvements in both validation and test accuracy after addressing the class imbalance. This indicates that the refined data set, focusing on extremes, positively influenced the Random Forest's ability to capture complex patterns and relationships within the data.

8.1.2 Random Forest Adaptability

The success of the Random Forest model in improving accuracy after addressing class imbalances is intricately tied to its ensemble learning architecture. Ensemble learning, a fundamental characteristic of Random Forests, involves combining predictions from multiple individual models, often referred to as decision trees. This ensemble approach imparts robustness and adaptability to the model.

One of the inherent strengths of Random Forests lies in their ability to capture complex, non-linear patterns within the data. In the context of predicting song popularity, where the relationships between various features and the target variable can be intricate and non-linear, Random Forests shine. The ensemble of decision trees collaboratively works to discern and model these nuanced dependencies, providing a more comprehensive understanding of the underlying patterns.

The refined data set, which specifically focuses on extremes of song popularity (top ten and last ten songs per month), aligns well with the ensemble nature of Random Forests. By concentrating on these extremes, the model is exposed to a diverse set of scenarios, allowing it to learn and adapt to the intricate relationships that define songs at the popularity extremes. This aligns with the philosophy of ensemble learning, where each decision tree contributes its insights, collectively forming a more robust and accurate prediction

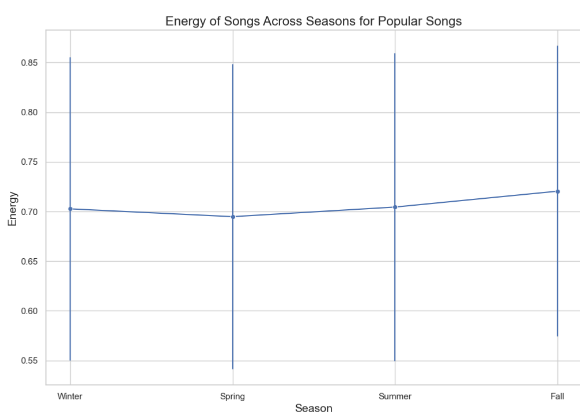
8.1.3 Overall Implications

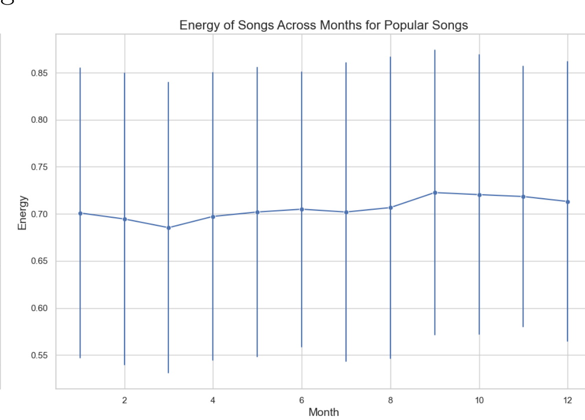
The shift in strategy, from considering the top 50 songs to concentrating on the extremes of popularity, played a crucial role in mitigating the class imbalance challenge. By creating a more balanced dataset, the Random Forest model could better generalize and make accurate predictions on unseen data, leading to the observed improvement in test accuracy. This adjustment not only addressed a specific challenge in the dataset but also leveraged the strengths of Random Forests in handling imbalances and complex relationships within the data.

8.2 Feature Engineering:

We aimed to create additional features that might capture important patterns in our data. Thus, we added a 'season' column to our DataFrame based on the 'month' column, and then applied one-hot encoding to create dummy variables for the 'season' column. This only helped improve our Random Forest model performance by 1%, and as the Feature Importance Graph shows (Figure 5), the 'season' columns features were the least important of all features.

However, in the visualization of the data, trends are easier to see when seen through the seasons' lens rather than the month lens as the figures below show:



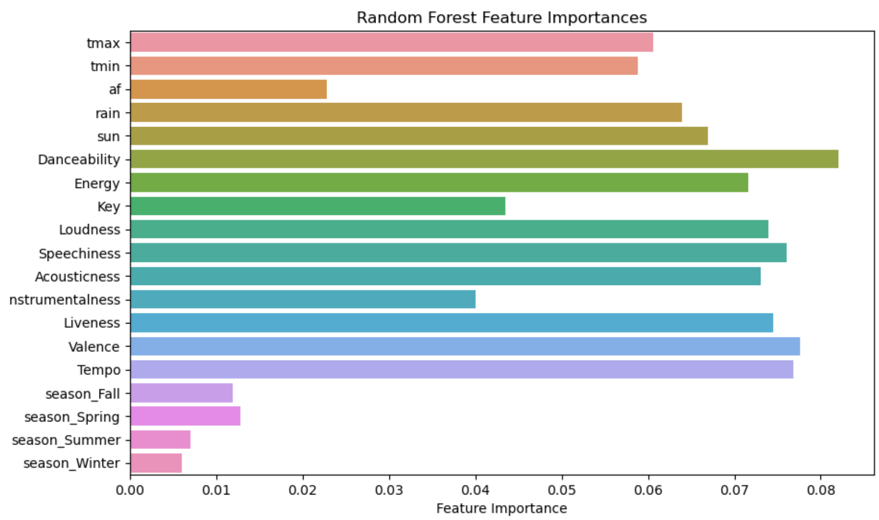


9. Discussion

The Random Forest model outperformed Logistic Regression, demonstrating its ability to capture the intricate relationships within the dataset. This aligns with our justification for choosing Random Forests in the presence of weak or non-linear correlations.

Additionally, as shown below, the Random Forest model identified the importance of various features in predicting song popularity. The feature importance, listed in descending order, are as follows:

Figure 5:



tmax (Mean daily maximum temperature)	0.06065317
tmin (Mean daily minimum temperature)	0.05882446
af (Number of days of air frost)	0.02275582
rain (Total rainfall)	0.06392285
sun (Total sunshine duration)	0.06694227
Danceability	0.08206818
Energy	0.07159498
Key	0.04349029
Loudness	0.07391056
Speechiness	0.07607553
Acousticness	0.07306664
Instrumentalness	0.04002637
Liveness	0.07448881
Valence	0.0776029
Tempo	0.07685957
Fall Season	0.01194391
Spring Season	0.01278183
Summer Season	0.00703342
Winter Season	0.00595844

9.1 Feature Importance

Danceability and Valence, with feature importances of 0.08206818 and 0.0776029 respectively, emerged as the most influential factors in predicting song popularity, underscoring the significance of musical characteristics related to rhythm and intensity in shaping preferences. These findings highlight the pivotal role that certain intrinsic song features play in determining the appeal of a song. Valence, in the context of music, refers to the musical positiveness or negativity conveyed by a track. It represents the musical mood, with higher valence indicating a more positive and cheerful mood, while lower valence signifies a more negative or melancholic mood. Now, danceability is a musical feature that quantifies how suitable a piece of music is for dancing. It is a measure that reflects the rhythm, tempo, and overall musical elements that contribute to a song's potential for dance. Specifically, danceability considers factors such as beat clarity, regularity, and overall rhythmic qualities that make a song conducive to movement and dance.

In summary, the feature importance suggests that songs with higher danceability and valence values are more likely to be popular, indicating a preference among listeners for music that is not only rhythmically engaging but also emotionally positive and cheerful. These insights provide a nuanced understanding of the intrinsic elements that contribute to the widespread appeal of songs.

9.2 Weather Impact

Now, contrary to our initial hypothesis, weather-related variables (tmax, tmin, af, rain, sun) demonstrated lower importance in predicting song popularity. This implies that while atmospheric conditions may influence music preferences, other factors, particularly intrinsic song features, play a more significant role.

We then hypothesized that a stronger correlation might exist between weather conditions and happy songs compared to sad songs. We conjectured that ”happy” weather could influence people to listen to upbeat tunes, and that the selection of sad songs might be less contingent on weather conditions; As this paper titled ”Here comes the sun: music features of popular songs reflect prevailing weather conditions” (Anglada-Tort et al., 2023) found: ”music features reflecting high intensity and positive emotions were positively associated with daily temperatures and negatively associated with rainfall, whereas music features reflecting low intensity and negative emotions were not related to weather conditions.”

To explore this, we isolated happy and sad songs based on their valence - a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track - with happy songs defined as those with a valence greater than 0.7 and sad songs as those with a valence less than 0.4. Subsequently, we generated correlation maps for both categories. Upon examination, the difference in correlation between weather features and song features appeared to be relatively weak.

Figure 6: Correlation Matrix for Sad songs

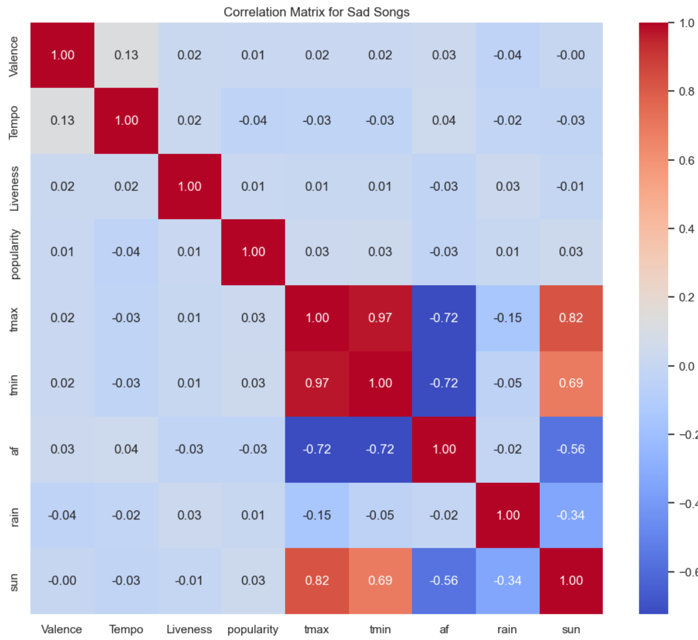
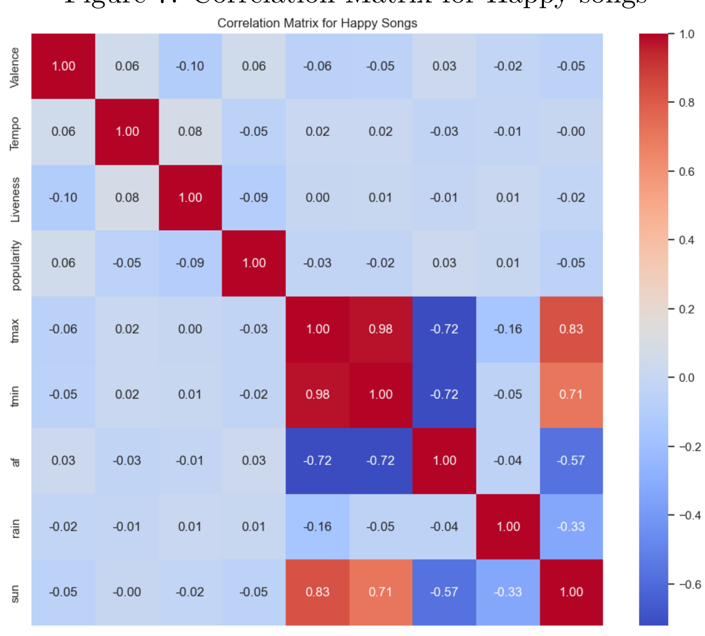


Figure 7: Correlation Matrix for Happy songs



10. Conclusion

Our exploration into the intersection of meteorology and musicology has provided valuable insights into the factors influencing the popularity of songs. While our initial hypothesis posited a clear connection between weather conditions and musical preferences, the results underscore the complexity of this relationship. The predictive nature of weather on popular music is less pronounced than anticipated, with intrinsic song features such as danceability and valence emerging as more influential determinants.

The success of our Random Forest model in outperforming Logistic Regression further emphasizes the intricate and nuanced patterns within the dataset. Danceability and valence, as the standout features, highlight the importance of rhythm, intensity, and emotional tone in shaping listener preferences. This reinforces the idea that, in the realm of popular music, the intrinsic qualities of a song play a pivotal role in determining its appeal, often surpassing the impact of external environmental factors.

10.1 Reflections

Our initial hypothesis posited that there would be a discernible influence of weather conditions on the popularity of songs, with expectations that distinct atmospheric nuances corresponding to different seasons would shape musical preferences. However, the results from our analysis challenge this hypothesis, indicating that the predictive nature of weather on popular music is not as clear-cut as anticipated.

Contrary to our expectations, the feature importance of weather-related variables (tmax, tmin, af, rain, sun) was relatively low compared to intrinsic song features (danceability, valence, energy, etc). This implies that, while weather conditions may have some influence on music preferences, their impact is not as significant as certain intrinsic qualities of songs.

The specific feature importance values highlight that factors like temperature, air frost, rainfall, and sunshine duration do not play a predominant role in determining the popularity of songs in our data set. This suggests that the emotional tone (valence) and danceability of a song, among other intrinsic features, have a more pronounced effect on listener preferences.

In essence, our hypothesis regarding the clear influence of weather conditions on musical preferences did not align with the observed data. The results indicate a more complex interplay of factors, where intrinsic song characteristics take precedence over seasonal variations in weather. The lack of a strong correlation between weather variables and song popularity emphasizes the intricate and multifaceted nature of musical preferences, which may be influenced by a myriad of personal, cultural, and psychological factors that extend beyond the immediate environmental conditions.

10.2 Limitations

While our analysis sheds light on the significant role of certain intrinsic features in predicting song popularity, the exact relationship between weather and music preferences remains elusive and is likely influenced by a combination of factors that our current model may not fully capture.

There are also some limitations of our study that warrant consideration. Firstly, the geographical scope is restricted to the United Kingdom, potentially limiting the generalizability of the findings to other regions with distinct weather patterns and cultural influences. The limited number of weather features incorporated in the analysis may not fully capture the nuanced relationships between meteorological conditions and music preferences. Furthermore, the monthly temporal resolution might oversimplify the dynamics, overlooking finer variations in weather patterns and their potential impact on song popularity. Additionally, external factors like the COVID-19 pandemic, which significantly altered societal behaviors, were not explicitly considered, suggesting the potential influence of unaccounted-for variables on the observed trends. Acknowledging these limitations underscores the need for caution in extrapolating the study's findings and highlights opportunities for future research to address these constraints and enhance the depth and breadth of our understanding of this intriguing intersection of meteorology and musicology.

10.3 Lessons

Our analysis highlights the cautionary lesson that assumptions based on common sense or prevailing beliefs ("Of course people listen to sad songs when it's gloomy outside!") may not always be clearly supported with empirical data. In this case, despite the expectation of a clear influence of weather on music preferences, the data suggests a more nuanced relationship, emphasizing the importance of relying on robust evidence rather than presuppositions when exploring complex interdisciplinary phenomena. With that being said, the interplay of weather and music remains a fascinating area for continued exploration in understanding the multitude of elements that contribute to the ebb and flow of popular songs over time.

References

Manuel Anglada-Tort, Harin Lee, Amanda E Krause, and Adrian C North. Here comes the sun: music features of popular songs reflect prevailing weather conditions. *R. Soc. Open Sci.*, 10(5):221443, May 2023.

Tom Button. UK weather by month (4 stations). `https://www.kaggle.com/code/tombutton/uk-weather-by-month-4-stations`, March 2023. Accessed: 2023-12-16.

Chart2000. Chart2000.Com: Music charts 2000 - 2023. `https://chart2000.com/about.htm`. Accessed: 2023-12-16.

Peter G Christenson and Jon Brian Peterson. Genre and gender in the structure of music preferences. *Communic. Res.*, 15(3):282–301, June 1988.

Meijun Liu, Xiao Hu, and Markus Schedl. The relation of culture, socio-economics, and friendship to music preferences: A large-scale, cross-country study. *PLoS One*, 13(12): e0208186, December 2018.

MATLAB. *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Elizabeth J Vella and Gregory Mills. Personality, uses of music, and music preference: The influence of openness to experience and extraversion. *Psychol. Music*, 45(3):338–354, May 2017.

Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL `https://doi.org/10.21105/joss.03021`.