

Song Popularity

Madeline Gorman

Western Governors University

Table of Contents

A. Project Highlights	3
B. Project Execution	3
C. Data Collection Process	6
Data Selection and Collection:	6
D. Data Extraction and Preparation	8
E. Data Analysis Process	8
E.1 Data Analysis Methods.	8
E.2 Advantages and Limitations of Tools and Techniques	9
E.3 Application of Analytical Methods.....	9
F Data Analysis Results	10
F.1 Model Significance.....	10
F.2 Practical Significance	12
F.3 Overall Success	12
G. Conclusion	13
G.1 Summary of Conclusions.....	13
G.2 Effective Storytelling.....	13
G.3 Recommended Courses of Action	14
References.....	16
Appendix A.....	17
Evidence of Project Completion.....	17

A. Project Highlights

Research Question:

This project addressed the question: “Is there a relationship between song metadata and song popularity?”

Scope of Project:

The scope of this project included data collection, exploration, cleaning, and analysis of an existing Spotify dataset. It also included classifying a song’s popularity as high or low, as well as the development, training, and evaluation of a random forest classifier model.

Tools and Methodology Used:

The methodology used in this project was the CRISP-DM methodology, which consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

To implement the solution, I used Python in a Jupyter Notebook. Using Python, I imported the necessary libraries, including Pandas, Scikit-learn, Matplotlib, and Seaborn, to implement all of the required steps to complete this project.

B. Project Execution

Project Plan:

The projected goals, objectives, and deliverables were the following:

- Goal: The goal of this project is to develop a model that can predict a song's popularity based on its metadata and audio features.
 - Objective 1.1: Identify data that is suitable for the project, perform exploratory analysis, and clean it.
 - Deliverable 1.1.1: The deliverable for this objective is a clean dataset ready for modeling.
 - Objective 1.2: Test, train, and validate the data on a random forest classifier and determine the parameters with the highest accuracy and most balanced confusion matrix.
 - Deliverable 1.2.1: The deliverable for this objective is a model that successfully predicts the popularity of a song depending on its features.
 - Objective 1.3: Develop a written report of the findings and conclusions, including visuals
 - Deliverable 1.3.1: The deliverable for this objective is a complete written report of the findings and conclusions of the project, with visuals to enhance them.

All of these goals, objectives, and deliverables were met, with no issues. They were all clear and easy to follow.

Project Planning Methodology:

The methodology used in this project was the CRISP-DM methodology. It consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. I completed these phases much faster than intended, but I still stuck to them. In this methodology, it's almost always necessary to go back and forth between the phases, which is something I definitely experienced. I went back and forth between the modeling and evaluation phases several times to refine the model.

Projected and Actual Completion of Timeline and Milestones:

Milestone or deliverable	Duration (hours or days)	Projected start date	Anticipated end date	Actual end date
Determine project goals	1 day	1/13/2026	1/13/2026	1/13/2026
Examine and explore potential datasets	1 day	1/14/2026	1/14/2026	1/13/2026
Clean and transform the data	1 day	1/15/2026	1/15/2026	1/13/2026
Develop and implement the model	2 days	1/16/2026	1/16/2026	1/15/2026
Evaluate the results and create written report	2 day	1/17/2026	1/17/2026	1/17/2026

I ended up starting and finishing this project on the anticipated start and end dates. I completed the first three milestones in my plan on the first day, when I originally anticipated

each of those items taking one day each to complete. These items were completed fast due to finding the dataset quickly and discovering it was already mostly complete and clean. The last two milestones took me two days each to complete, which I originally anticipated each would only take one day. The milestone of developing and implementing the model took a day longer than expected because the model needed to run for a long time, and I had to run it multiple times. The milestone of evaluating the results and creating a written report took a day longer than expected because it contained more writing than I could do in one day. Overall, I stuck to the outline rather closely.

C. Data Collection Process

Data Selection and Collection:

I initially planned to collect the dataset by downloading the CSV file from Kaggle, and that is exactly what I did. The process of selecting this dataset consisted of browsing quite a few Spotify datasets on Kaggle to see what kind of data each one had. The one I ended up using was the one I originally planned to use in Task 2.

The data was complete and mostly clean. I only had to remove a few duplicates and impute 3 missing values, but other than that, the data was complete, clean, and ready to use. This is the process I anticipated in the previous task as well. There were no obstacles when collecting this data. The process of downloading it from Kaggle and reading it into a Jupyter Notebook was very straightforward.

I didn't experience any unplanned data governance issues. I took all of the precautions I planned to take in Task 2. These included:

- Data governance precaution – Keep a record of all data cleaning, transforming, and analysis to allow for reusability and reproducibility.
- Privacy precaution – No precaution needed because the dataset does not contain any identifying information.
- Security precaution – Store the data in a password-controlled location to prevent unauthorized access to the data.
- Ethical, legal, and regulatory compliance precaution – Only use the data that is available in the dataset because it is public data that is free to use for educational and research purposes. This ensures compliance with legal, ethical, and regulatory standards.

Following these precautions allowed me to stay in compliance with data governance, privacy, security, ethical, legal, and regulatory standards with no issues.

An advantage of using this dataset was that it was already relatively clean and required minimal additional cleaning. The audio features had the correct data types and were in the right formats. This allowed me to begin working with and analyzing the data quickly. A disadvantage of using this dataset was that it didn't include any outside factors that could contribute to a song's popularity. Factors, including whether an artist is signed to a record label, the number of followers they have, and the marketing efforts they have taken, can also contribute to the success and popularity of a song. The absence of non-audio features limited this project to analyzing a song's popularity solely based on its audio features.

D. Data Extraction and Preparation

I downloaded the CSV file from Kaggle and read it into a Jupyter Notebook using Python's Pandas library. The data was complete and mostly clean. Using Pandas, I observed 1 unnecessary column, 3 null values, and 450 duplicates. I removed the column and all of the duplicates, and imputed the 3 null values with 'Unknown'. The null values were the artist's name, album name, and track name, but all of the numerical values were still present, so I decided to change the missing values to 'Unknown' rather than removing them. The data overall did not require much cleaning, because it was collected in a methodical way. For example, there were exactly 1000 songs collected for each genre. It was a very easy dataset to work with. Using Pandas was appropriate for this dataset because it provided numerous methods to observe different aspects of the data and clean it accordingly.

E. Data Analysis Process

E.1 Data Analysis Methods.

The data analysis method I used was a random forest model. Random tree classifiers are particularly effective when handling data that contains non-linear relationships among variables. The dataset I used contained a multitude of numeric features, making this model a perfect fit. The model was used to predict whether a song has high or low popularity based on its features. It was an appropriate choice for supporting the hypothesis because it not only provided high accuracy, but it also allowed for the use of feature importance. Feature importance is a large

focus of this project, and with it, I was able to identify which song features contribute most to popularity.

E.2 Advantages and Limitations of Tools and Techniques

An advantage of using a random tree classifier is that it's especially effective when handling complex relationships between variables. This dataset contained many numeric variables with complex relationships, making this model particularly beneficial. One limitation of using a random tree classifier is that it is computationally complex. Training the model was computationally intense and took about 2 hours to run.

E.3 Application of Analytical Methods

The first requirement for the random forest model was to define a target variable, which was popularity. I assessed the popularity column in the data and identified that the 50th percentile was at 35. That was where I decided to cut the data in half. This requirement was verified by categorizing popularity as high if it was above 35 and low if it was below 35. The next requirement was to encode the popularity variables. I encoded the popularity categories as 0 for high and 1 for low. This requirement was verified by affirming that the values were being represented by the correct classifications. Another requirement for this model was to use numerical features. The song features included danceability, energy, speechiness, acousticness, instrumentalness, liveness, and valence, and were each represented by a number. This requirement was verified by confirming that each feature had suitable data types before moving on to modeling. The next requirement was to split the data into training and testing subsets. I implemented this step so the model could be trained on one set and tested on the other. The

training set consisted of 80% of the data, while the testing set consisted of 20%. This requirement was verified by implementing the `train_test_split` function. The use of training and testing data verified that the model would be evaluated on unseen data, ensuring its reliability. The last requirement was tuning the hyperparameters to improve the model's performance. Parameters, including the tree's depth and the number of trees, were adjusted to identify which performed best. The final model was trained on the best performing parameters, completing the random forest classifier's implementation.

F. Data Analysis Results

F.1 Model Significance

Model 1 – Baseline Random Forest:

- Model type: Supervised classification model
- Algorithm used: Random Forest Classifier
- Metrics used: Accuracy and confusion matrix
- Benchmark for success: The model will be considered successful if it performs reasonably, with an accuracy score above 50%, and has a decently balanced confusion matrix.
- Conclusion: The baseline model had an accuracy score of 76.27% and the following confusion matrix: $\begin{bmatrix} 8746 & 2330 \\ 3058 & 8576 \end{bmatrix}$. This confusion matrix, along with the accuracy score, made for a great starting place, as both surpassed the benchmark. The accuracy score was significantly above the benchmark of 50%, and the

confusion matrix was decently balanced. The model predicted 8746 true positives and 8576 true negatives, along with 2330 false negatives and 3058 false positives. These results support my research question because the model's ability to predict a song's popularity based on its features indicates that song metadata does contribute to a song's popularity. They also show that the model is successfully predicting song popularity given a song's features.

Model 2 – Final Random Forest:

- Model type: Supervised classification model
- Algorithm used: Random Forest Classifier
- Metrics used: Accuracy and confusion matrix
- Benchmark for success: The model will be considered successful if the accuracy score is 70% or higher, demonstrates either similar or better performance in comparison to the baseline model, and the confusion matrix contains acceptable levels of false positives and negatives.
- Conclusion: The final model had an accuracy score of 76.52% and the following confusion matrix: [[8770 2306] [3026 8608]]. This confusion matrix shows that the model predicted 8770 true positives and 8608 true negatives, along with 2306 false negatives and 3026 false positives. This model surpassed the benchmarks, with an accuracy score above 70%, better performance than the baseline model, and a more balanced confusion matrix. In the final model, the number of true positives and true negatives increased, while the false positives and false negatives decreased, suggesting class balance in predictions. The misclassification errors are relatively

evenly distributed among both classes. These results support my research question because they indicate song metadata contributes to its popularity. This model also successfully predicted a song to be high or low popularity 76.52% of the time.

F.2 Practical Significance

The practical significance of the model was assessed by the model's predictive performance. The benchmarks to determine the model's practical significance were the model's accuracy score being 70% or higher, performing similarly or better than the baseline model, and having a balanced confusion matrix. All three of these benchmarks were met, deeming this model practically significant. This indicates that the model is reliable enough to make decisions about the likelihood of a song becoming popular, given its features.

The clients, music artists, producers, writers, and record label executives, would be able to utilize this model to predict a song's popularity before release. If they created a song and had the measurements of each musical feature, they would be able to utilize the model to predict the song's popularity. This would allow them to allocate more or fewer resources to any given song and its release, depending on the results of the model. They would also be able to see which features hold the most importance and contribute most to a song's popularity. With knowledge of which features are most important, they would be able to incorporate those features into songs they create.

F.3 Overall Success

The determining factor of this project's success was the creation of an effective predictive model and ultimately, identifying whether a song's features contribute to its

popularity. Given the results, this project was a success. The final model proved to be effective and was able to predict a song's popularity correctly 76.52% of the time. The model also produced a list of the most important features, further affirming that song features do contribute to a song's popularity. Overall, this project was successful.

G. Conclusion

G.1 Summary of Conclusions

This project aimed to answer whether song metadata contributes to song popularity, in addition to developing a model that can predict whether a song will be popular or not. The results of the model indicate that song metadata does, in fact, contribute to a song's popularity. The model was able to successfully predict a song's popularity 76.52% of the time. The model also allowed for the use of feature importance, revealing the specific features that contribute the most to popularity. The results of feature importance indicate the relative contribution each feature has to predicting a song's popularity. Ultimately, with this project's success criteria, this model was a success and answered the research question affirmatively.

G.2 Effective Storytelling

Using the Python library Seaborn, I plotted 2 confusion matrices: one for the baseline model results and one for the final model results. They both show the number of true positives, true negatives, false positives, and false negatives the model produced. Comparing these numbers allowed me to visualize whether the model was performing better or not with the selected

hyperparameters. After these plots, I plotted a horizontal bar chart using Python's Matplotlib library to visualize feature importance. This was crucial in my visualizations because it directly relates to the research question. In this project, I aimed to determine if song metadata contributes to song popularity, and the feature importance itself shows the specific song features that contribute the most to a song's popularity. The confusion matrices showed the improved performance of the final model in comparison to the baseline model.

G.3 Recommended Courses of Action

One recommendation I have is that music artists, producers, and writers should focus primarily on the audio features with the highest importance. When making a song, I suggest that they ensure to incorporate the features found to be the most important, as they contribute the most to a song's popularity. The top three features were acousticness, duration, and danceability. Ensuring these features are present in a song will increase the likelihood of it becoming popular. This directly addresses the research question because the three features previously mentioned are the three that contribute most to a song's popularity.

Another recommendation I have is that it would be beneficial for record labels to implement this project's final model. If they created a song and had the measurements of each musical feature, they would be able to utilize the model to predict the song's popularity. This would allow them to allocate more or fewer resources to any given song and its release, depending on the results of the model. This directly addresses the research question because the results of the model indicate that the music features of a song do contribute to the success and popularity of a song. Implementing this model would allow the artists to gauge the potential popularity of a song.

References

No sources were cited.

Appendix A

Evidence of Project Completion

Dataset used: <https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset>