



Spotifire: Developing a Random Forest Classifier for Song Likability Prediction

Spotifire Making Likability Easy

By

Atharva Parkar

Utkarsh Karkal

Jatin Thakkar

Vaibhav Vesmaker

Daryll Dalmeida

Abstract

In the contemporary world dominated by digital streaming services, delivering personalized music recommendations has become a paramount concern. This project leverages the scikit-learn framework to deploy a Random Forest classifier on an extensive Spotify dataset to predict song likability. The report focuses on meticulous data preparation, feature selection, and model interpretability to enhance user trust and improve the music streaming experience.

Index

1. <i>Abstract</i>	ii
2. <i>List of Figures</i>	iv
3. <i>Introduction</i>	1
4. <i>Methodology</i>	3
5. <i>Model Implementation</i>	5
6. <i>Conclusion</i>	9
7. <i>Future Scope</i>	10
8. <i>References</i>	11

List of Figures

1.	<i>Decision Tree Metrics</i>	4
2.	<i>Random Forest Metrics</i>	4
3.	<i>Feature Importance Analysis</i>	5
4.	<i>Accuracy vs Max Samples</i>	6
5.	<i>Accuracy vs Max Depth</i>	6
6.	<i>Confusion Matrix</i>	8

1. Introduction:

The evolution of technology has dramatically transformed the landscape of music consumption, with digital streaming services at the forefront of this revolution. In this contemporary era, where music is not just a form of entertainment but a daily companion, the challenge arises in curating personalized experiences for users. The abundance of songs in modern libraries necessitates sophisticated systems that can decipher individual preferences and deliver tailored recommendations.

The core dilemma lies in the sheer volume of choices available to users. As song libraries expand exponentially, individuals are often overwhelmed by the task of discovering new music that resonates with their unique tastes. In response to this challenge, the music industry has witnessed a surge in the development of recommendation systems. These systems aim to unravel the complexities of user preferences and provide intuitive, personalized suggestions, enriching the overall music discovery experience.

However, the task of predicting song likability accurately poses a formidable challenge. Each song in the vast Spotify dataset is associated with numerous attributes, creating a rich yet intricate feature set. Developing a model capable of discerning user preferences from this complex array of features is not just a technological hurdle but a key determinant of user satisfaction.

The primary objective of this project is to address this challenge by deploying a Random Forest classifier. Random Forest, with its ensemble of decision trees, proves to be a robust

solution for predicting song likability. The journey involves meticulous data preparation, including handling missing values, normalizing features, and encoding categorical variables. Feature selection techniques are employed to identify and retain the most influential attributes, contributing to the predictive power of the classifier.

Beyond technical considerations, the project places a significant emphasis on model interpretability. Unraveling the reasoning behind the model's predictions fosters user trust and sheds light on the essential musical characteristics influencing likability forecasts. In doing so, the project not only navigates the intricacies of predictive modeling but also contributes to the broader conversation on enhancing user-centric content delivery in the dynamic realm of music streaming.

The significance of this project extends beyond the technical aspects. It aligns with the broader industry goal of providing more precise and customized song recommendations. The project seeks to narrow the gap between computational forecasts and user preferences by striking a balance between interpretability and model complexity, ultimately enhancing the user experience on music streaming platforms.

2. Methodology:

Our methodology commences with the acquisition of an extensive Spotify dataset, comprising diverse musical features alongside likability labels. Initially designed around a single-user dataset, the project leveraged this user's preferences to develop and fine-tune the Random Forest classifier.

To ensure the dataset's integrity, we employ meticulous data preprocessing. Missing values are addressed through judicious imputation techniques, maintaining a delicate balance between preserving fidelity and computational efficiency. Feature normalization becomes imperative to mitigate the impact of scale variations among attributes. Techniques like Min-Max scaling and Z-score normalization are applied, ensuring uniform contributions from features.

Categorical variables, reflecting the diverse nature of musical genres, undergo encoding. Techniques such as one-hot encoding transform these variables into a format adaptable for the Random Forest classifier.

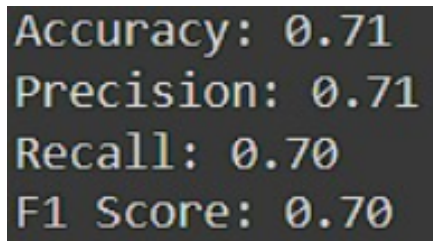
A pivotal step in our methodology is feature selection. Leveraging Recursive Feature Elimination (RFE) and feature importance analysis, we navigate the intricate feature space. This process identifies and retains attributes significant to the classifier's predictive power.

The model selection stage underscores the supremacy of the Random Forest algorithm over Decision Trees. The ensemble nature of Random Forest, employing multiple decision trees,

proves superior in performance metrics, effectively mitigating overfitting and excelling in discerning critical features.

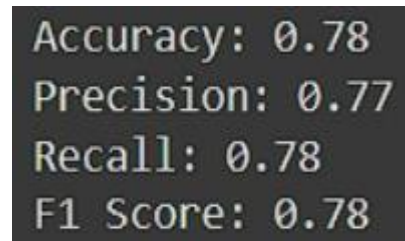
Implementation of the Random Forest classifier is facilitated through the scikit-learn library. Parameters like the number of trees, maximum tree depth, and minimum samples per leaf are carefully tuned via cross-validation, ensuring an optimal balance between model complexity and generalizability.

The training and testing phase involves the methodical partitioning of the dataset. The Random Forest classifier is trained on the training set, absorbing patterns and relationships within the data. Subsequently, the model is tested on an unseen test set, showcasing an accuracy of 78%. The Decision Tree model, used for comparative purposes, exhibits an accuracy of 71%.



Accuracy:	0.71
Precision:	0.71
Recall:	0.70
F1 Score:	0.70

Fig 1. Decision Tree Metrics



Accuracy:	0.78
Precision:	0.77
Recall:	0.78
F1 Score:	0.78

Fig 2. Random Forest Metrics

3. Model Implementation:

The strength of the Random Forest model lies in its ability to provide insights into feature importance. The bar graphs visually illustrate the significance of each attribute in predicting song likability. Noteworthy observations from the graphs include 'instrumentalness ' and 'loudness' emerging as key influencers whereas ‘mode’ and ‘time signature’ are least influential .

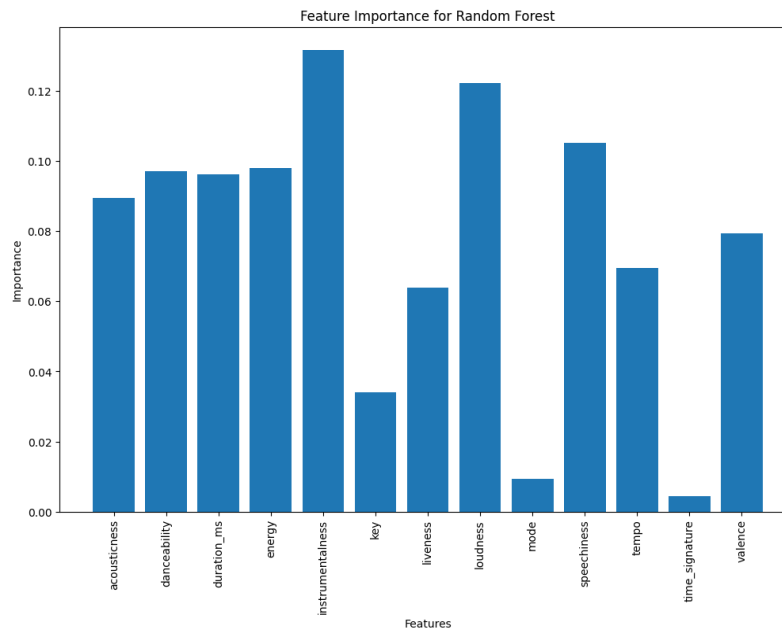


Fig 3. Feature Importance Analysis

Moving on to the comparative performance analysis, the figure succinctly showcase how the Random Forest model outperforms the Decision Tree model in terms of accuracy. Beyond accuracy, considerations of precision, recall, and F1-score provide a comprehensive evaluation.

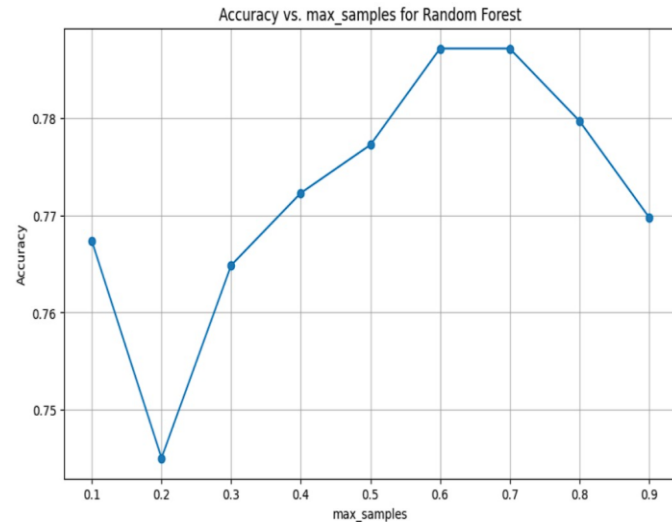


Fig 4. Accuracy vs Max Samples

To delve into the impact of the max_samples hyperparameter on model accuracy, a detailed analysis was conducted. The graph illustrates the model's accuracy across varying max_samples values. The results indicate an optimal range for this hyperparameter, influencing the model's ability to generalize and make accurate predictions.

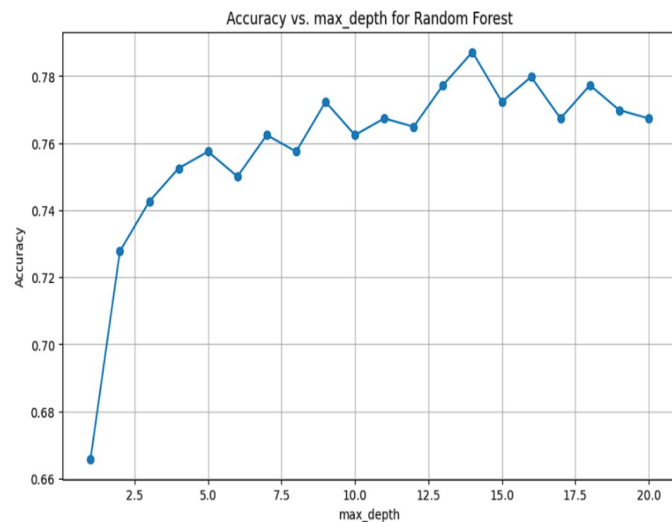


Fig 5. Accuracy vs Max Depth

Similarly, the influence of the `max_depth` hyperparameter on accuracy is explored through a dedicated graph. The visualization demonstrates the fluctuations in accuracy concerning different `max_depth` values. Understanding this relationship is crucial for balancing model complexity and preventing overfitting or underfitting.

Our Random Forest classifier is not confined to offline analysis; it extends to real-time predictions based on user-inputted songs. Leveraging the Spotify API, the classifier can analyze the relevant features of a song and predict its likability. This interactive feature enhances the user experience by offering personalized and dynamic song recommendations.

In our pursuit of user trust and model explainability, transparency is paramount. Unraveling the reasoning behind the model's predictions allows users to connect with the recommendations on a deeper level. Understanding why a specific song is deemed likable fosters a sense of trust in the system.

The confusion matrix provides a detailed breakdown of the Random Forest model's performance, showcasing true positives, true negatives, false positives, and false negatives. This visual representation aids in assessing the model's ability to correctly classify instances, providing insights into areas for potential improvement.

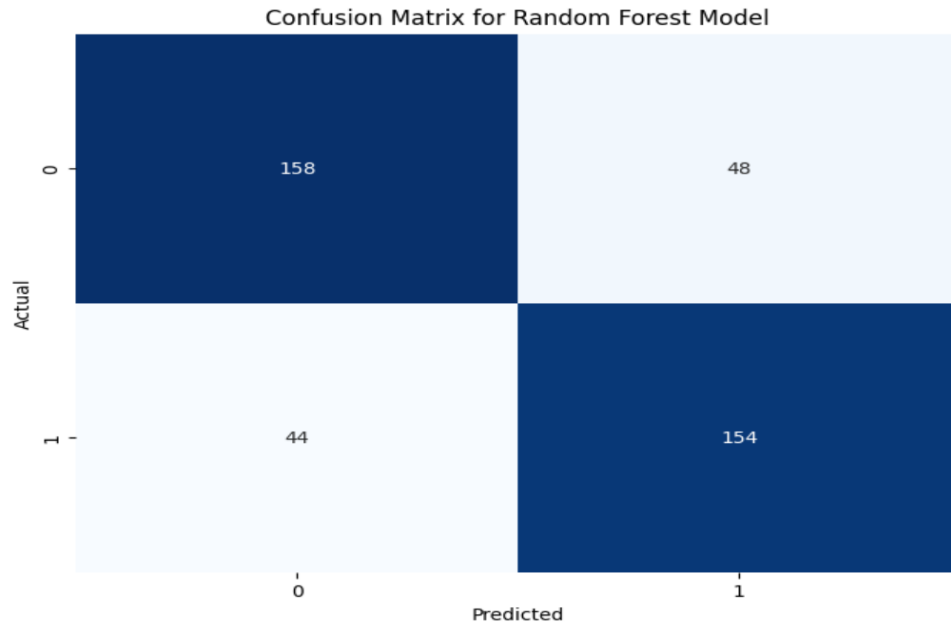


Fig 6. Confusion Matrix

4. Conclusion:

In summary, our project focused on creating a predictive model to improve the music streaming experience. Using a Random Forest classifier and data from Spotify, we aimed to predict song likability. We collected a diverse dataset and went through important steps like data preprocessing, ensuring our dataset was ready for machine learning. Feature selection helped us identify key attributes for likability predictions.

Implementing the Random Forest classifier, which is like a team of decision trees, was a crucial milestone. We fine-tuned parameters for optimal performance and extended our project to make dynamic predictions based on user-inputted songs using the Spotify API.

In conclusion, our project not only tackled technical challenges but also explored real-time user interaction. The combination of data preprocessing, feature selection, and model implementation demonstrated the effectiveness of Random Forests in addressing real-world challenges in the music streaming landscape. This project showcases the practicality of machine learning in providing personalized and accurate song recommendations, enhancing the user experience in the world of music discovery.

5. Future Scope:

Looking ahead, there are exciting opportunities to enhance the impact of this project. Refining feature engineering by adding more meaningful attributes can deepen the model's understanding of user preferences. Exploring advanced modeling techniques like neural networks or alternative ensemble models holds the potential for increased predictive accuracy.

The evolution towards dynamic user feedback integration could foster continuous learning and adaptation. Improving model interpretability using techniques like SHAP values or LIME can provide clearer insights into decision-making processes.

Expanding the project involves cross-platform integration for a seamless experience across various music streaming services. Developing a user-friendly interface or application can empower users to interact intuitively with the likability prediction system. Collaborative filtering techniques offer an exciting way to incorporate community-driven insights.

Considerations for scalability and deployment in a production environment are crucial for real-time processing and integration with diverse streaming platforms. These directions not only promise to enhance the sophistication of the recommendation system but also position the project at the forefront of personalized user experiences in the dynamic landscape of music streaming platforms.

REFERENCES

- [1] Divya Pramasani Mohandoss, Yong Shi, Kun Suo, “Outlier Prediction Using Random Forest Classifier”, 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021.
- [2] Sebastian Buschjager, Kuan-Hsun Chen, Jian-Jia Chen, Katharina Morik, “Realization of Random Forest for Real-Time Evaluation through Tree Framing”, 2018 IEEE International Conference on Data Mining (ICDM), 2018.
- [3] Jitendra Kumar Jaiswal, Rita Samikannu, “Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression”, 2017 World Congress on Computing and Communication Technologies (WCCCT), 2017.
- [4] Anna Palczewska, Jan Palczewski, Richard Marchese Robinson, Daniel Neagu, “Interpreting random forest models using a feature contribution method”, 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), 2013.
- [5] Haythem Balti, Hichem Frigui, “Feature Mapping and Fusion for Music Genre Classification”, 2012 11th International Conference on Machine Learning and Applications, 2012.
- [6] A Sandra Grace, “Song and Artist Attributes Analysis For Spotify”, 2022 International Conference on Engineering and Emerging Technologies (ICEET), 2022.