Exploring the 'Most Streamed Spotify Songs of 2023' dataset to find streaming trends and developing a supervised regression model to predict the streams for a song based on selected features.

*By Christian, Sean, and Martin*

Results

**Feature importance:**

A screen shot of a computer

Description automatically generated A screenshot of a computer code

Description automatically generated

**Optimizing the hyperparameters with GridSearchCV (took forever)**

A screen shot of a computer

Description automatically generated

A graph with a line and a red line

Description automatically generated

Conclusion

We found that there is a strong correlation between the number of streams a song gets, and the total amount of playlists its added to. Furthermore, a weaker correlation between the “release\_date” and “total\_streams” was also spotted.

After optimizing the Gradient boosting regressor we ended up with a model that when tested had a R2 of 0.811

For the other features in the dataset no significant correlation was found in relation to predicting streams.

Acknowledgements

🥰 ChatGPT 🥰

Introduction

We aim to create a model predicting the number of streams a song is likely to receive using a dataset that includes artist information, release details, and audio features. Our goal is to explore how these features influence a song’s popularity and build a regression model for stream prediction.

We want to see if it is possible to see how popular a song will be. If possible, could this be beneficial for the music industry?

Materials and methods

We have chosen a dataset of the most streamed songs in 2023. The list doesn’t only contain songs that are released in the year 2023. The list also contains songs that have been released going back to the 1960’s and thereby provides us with a little broader variety of songs.

After cleaning the initial dataset we had a look at a heatmap for the many features, to see if any might have correlation. We notice a weak to medium correlation between the valence\_% and energy\_% as well as danceability\_%, but the only strong correlation is with streams and total playlists. This is cool because we want to predict the streams feature!

A screenshot of a computer

Description automatically generated

This is a regression problem, so we picked five regressors to train on, that where then evaluated on their R2 value.

|  |  |  |
| --- | --- | --- |
| Name | R2 (rounded) | MSE |
| GradientBoosting | **0.81** | 4.73e+16 |
| RandomForrest | **0.80** | 4.90+e16 |
| DecisionTree | **0.61** | 9.63e+16 |
| Linear | **0.58** | 1.02e+17 |
| SVR | **0.44** | 1.37e+17 |

*Table 1: Result of testing each model*

We picked the Gradient boosting regressor to capture the correlation between a song's characteristics and its streaming figures. We opted for this method due to its resilience to outliers and its capacity to manage non-linear associations. To optimize the model's performance, we applied Grid Search for hyperparameter tuning. See more in results.

Literature Cited

Géron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media.

Datasource: https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023/data

**BUILDING A MODEL THAT CAN PREDICT SONG STREAMS**