Executive Summary – SPOT-POP

Objective:

The main objective of this project is to develop a predictive model that can classify the popularity of Spotify tracks based on their audio features. By analyzing a dataset containing various attributes of Spotify tracks, we aim to identify key factors that contribute to a track's popularity and create a reliable predictive system.

Data:

The data for this project is sourced from Kaggle, which can be found here. The dataset contains over 100,000 song entries and 20 different features for each entry, making it a suitable dataset or data exploration, cleaning, and training.

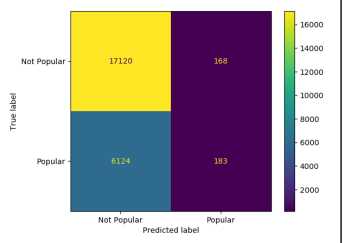
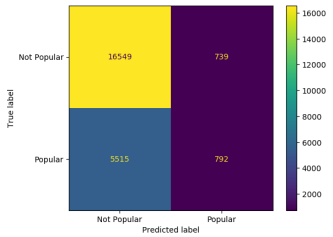
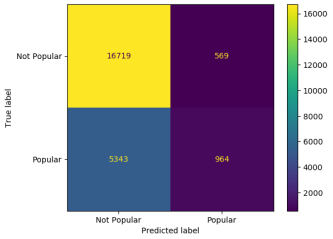
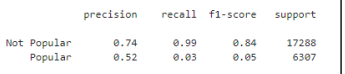
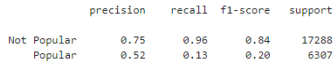
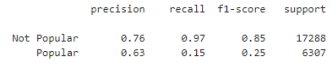
Primary Variables and Features:

After extensive data exploration and cleaning we chose the following features as our input (*Loudness, explicit, danceability, time_signature, energy, tempo, key, liveness valence, mode, speechiness, duration, acousticness, instrumentalness)* and the *(popularity decision)* as the output of the model.

Modeling Approach:

We are leveraging a classification modeling approach to predict whether a song entry will be “popular” or “not popular”. To prepare the data for this approach we categorized the popularity data to being popular and not popular. We are using sklearn library and we are splitting our data with the stratify option to maintain a similar value count between our splits in different categories with the test size being 30%. Logistic Regression, KNN and Random Forest classification approaches are used.

Model Performances

	LogR	KNN (K=30)	RF
Accuracy	73.33%	73.49%	74.94%
Conf Matrix			
Metrics			

While the models have performed well on determining “Not Popular” songs, the model is under performing on predicting “Popular” songs, which can be attributed to the smaller portion of popular data.

Next Steps

Future work includes better feature selection schemes and working with GridSearchCV to find the best fit for each model.