Song Popularity

Fion Ho, Joyce Mok, Hung Nguyen, Kaili Nguyen, Rithika Reddy, Isabelle Supandji, Joshua Zhang

- Introduction
- Project Goal
- Our Data
- Pre-Processing
- Our Methods

- Variable Importance
- Performance Evaluation
- Final Insights
- Recommendation
- Our Poster

Upgrade

Stats 140 ▼

Introduction

With 100,000 songs being uploaded to music streaming platforms daily, the music industry is highly competitive. Producing popular songs proves to be extremely difficult in such an over-saturated market.

31%	489M	100M+	$12B
Global Streaming Market Share	Monthly Listeners	Tracks Available	Annual Revenue

Introduction

Project Goal

Our Data

Pre-Processing

Our Methods

Variable Importance

Performance Evaluation

Final Insights

Recommendation

Our Poster

Upgrade

Stats 140 ▼

Project Goal

This project aims to build a highly accurate machine-learning model that can predict a song's popularity based on its characteristics.

To test our hypothesis, we used 4 models - logistic regression, K-nearest neighbors, support vector machine, and random forest and compared the accuracy rates to determine the "best" model for our project

Play

Follow

Upgrade

Stats 140 ▼

Introduction

Project Goal

Our Data

Pre-Processing

Our Methods

Variable Importance

Performance Evaluation

Final Insights

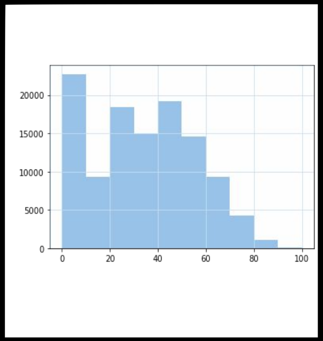
Recommendation

Our Poster

Our Data

Our Kaggle Dataset

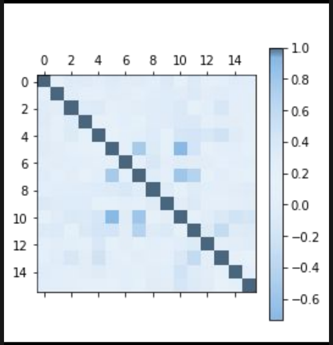
114,000 records with 21 different variables



Popularity Histogram

Our response variable, popularity, is a value between 0 - 100, with 100 being the most popular

Determined by total number of plays and weighed based on recency



Correlation Matrix

Has many use cases including helping us determine need for PCA, check model assumptions, and identify highly correlated variables

Upgrade

Stats 140 ▼

Introduction

Project Goal

Our Data

Pre-Processing

Our Methods

Variable Importance

Performance Evaluation

Final Insights

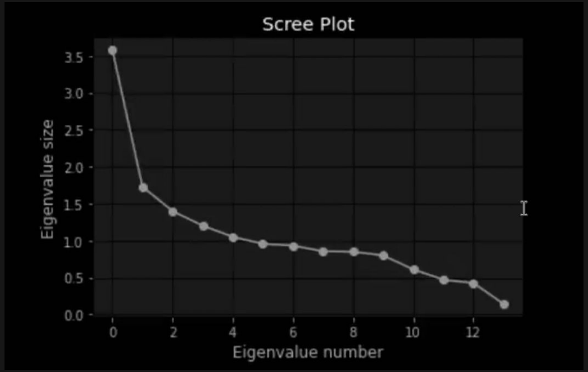
Recommendation

Our Poster

Pre-Processing

01 Removed unused variables

02 Transform popularity and genres into binary variables



03 Principal Component Analysis (PCA)

Possibility that larger models with more predictors will decrease accuracy due to overfitting.

PCA will identify patterns in the data based on correlation to reduce dimensionality

We transformed our training and test sets using the PCA of 2 components since the plot shows us that the first 2 principal components capture most of the information in our data

04 K-Fold Cross Validation

05 Hyper Parameter Tuning

- Introduction
- Project Goal
- Our Data
- Pre-Processing
- Our Methods

- Variable Importance
- Performance Evaluation
- Final Insights
- Recommendation
- Our Poster

Upgrade

Stats 140 ▼

Our Methods

Used 80% of our data to train our 4 learning algorithms and the remaining 20% as the test set

Logistic Regression

Estimates coefficients for each predictor to calculate the probability of the binary outcome for new data

K-Nearest Neighbor (KNN)

Non-parametric approach that finds k number of centroids, assigns point to the cluster, updates the points to new clusters, and repeats

Support Vector Machine (SVM)

Plots each data point in an n-dimension and performs classification to find the best hyper-plane that splits the data

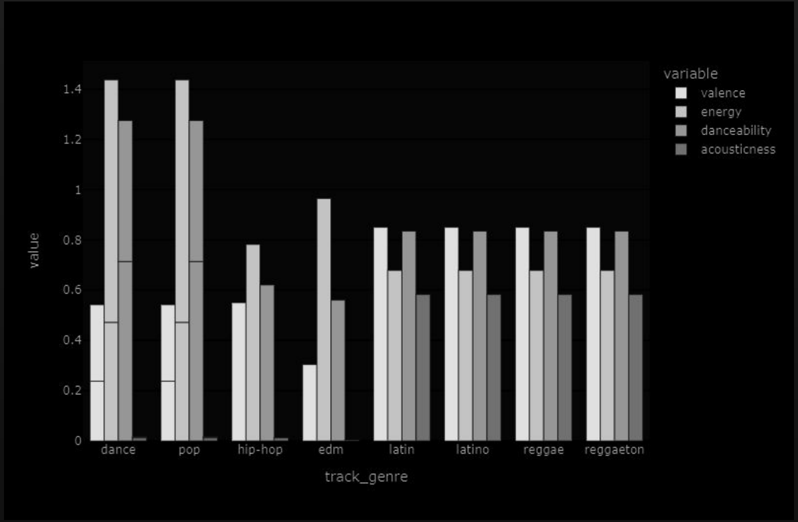
Random Forest

Fits multiple decision trees and uses averaging to improve the accuracy rate

Introduction
Project Goal
Our Data
Pre-Processing
Our Methods
Variable Importance
Performance Evaluation
Final Insights
Recommendation
Our Poster

Variable Importance

Top 10 Genres & Feature Selection



#	Feature Selection
1	Instrumentalness
2	Loudness
3	Speechiness
4	Explicit
5	Valence

- Introduction
- Project Goal
- Our Data
- Pre-Processing
- Our Methods

- Variable Importance
- Performance Evaluation
- Final Insights
- Recommendation
- Our Poster

Upgrade

Stats 140 ▼

Most Accurate
Modeling Technique

Performance Evaluation

	Logistic Regression	KNN	SVM	Random Forest
Accuracy	**0.712**	**0.767**	**0.717**	**0.716**
Precision	0.493	0.587	0.603	0.674
Recall	0.639	0.032	0.046	0.022
F1 Score	**0.061**	**0.612**	**0.087**	**0.042**

Introduction

Project Goal

Our Data

Pre-Processing

Our Methods

Variable Importance

Performance Evaluation

Final Insights

Recommendation

Our Poster

Upgrade

Stats 140

Final Insights

01 K-Nearest Neighbors was the best model

All models demonstrate high accuracy metrics above 70% but when looking at the F1 score, there's a drastic difference between KNN, making it our best model which predicts song popularity with 76.7% accuracy.

02 Instrumentalness, Loudness, Speechiness, and Explicit were the most important predictors

After running a feature selection, we discovered that instrumentalness and speechiness had a strong negative correlation with popularity while loudness and explicit had positive correlations with popularity.

- Introduction
- Project Goal
- Our Data
- Pre-Processing
- Our Methods

- Variable Importance
- Performance Evaluation
- Final Insights
- Recommendation
- Our Poster

Upgrade

Stats 140 ▼

Recommendation

Use Cases

For Artists

Artists can optimize the popularity of their new music based on the important features identified in feature selection. Ideal songs should be **loud, explicit tracks with a strong blend of vocals and music.**

For Music Labels

Increase ROI

Maximize revenue by signing dance, pop, and hip hop artists, which are the most popular genres.

Influence Decision Making

Leverage our KNN model to determine which single to release prior to an album

Optimize Budget

Allocate additional spend to support songs that are projected to be less popular

Introduction

Project Goal

Our Data

Pre-Processing

Our Methods

Variable Importance

Performance Evaluation

Final Insights

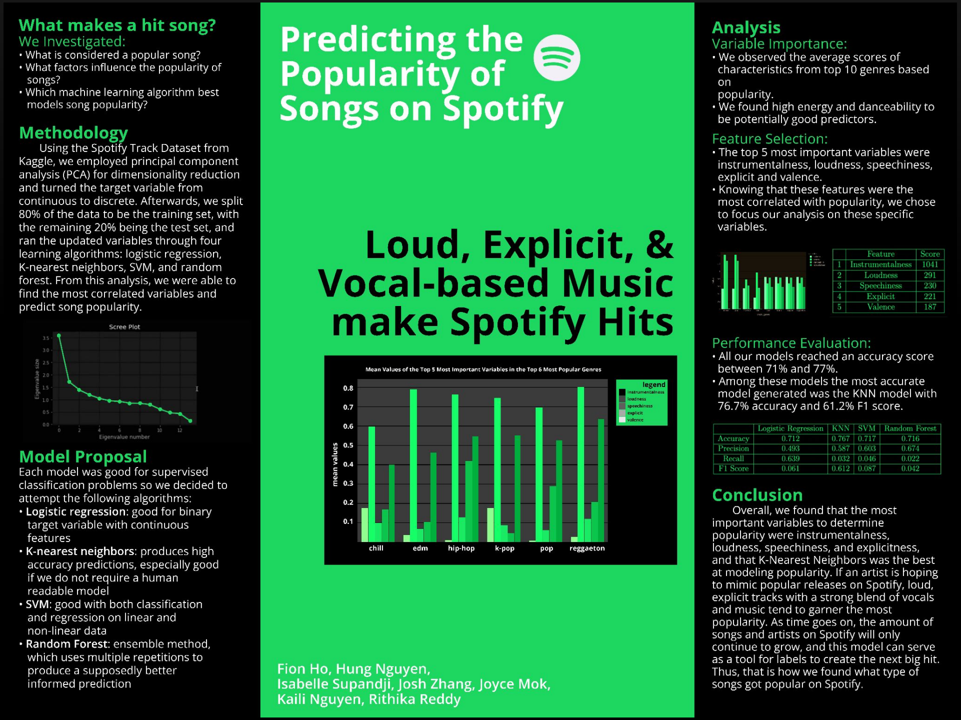
Recommendation

Our Poster

Upgrade

Stats 140

Our Poster



What makes a hit song?

We Investigated:

- What is considered a popular song?
- What factors influence the popularity of songs?
- Which machine learning algorithm best models song popularity?

Methodology

Using the Spotify Track Dataset from Kaggle, we employed principal component analysis (PCA) for dimensionality reduction and turned the target variable from continuous to discrete. Afterwards, we split 80% of the data to be the training set, with the remaining 20% being the test set, and ran the updated variables through four learning algorithms: logistic regression, K-nearest neighbors, SVM, and random forest. From this analysis, we were able to find the most correlated variables and predict song popularity.

Model Proposal

Each model was good for supervised classification problems so we decided to attempt the following algorithms:

- **Logistic regression**: good for binary target variable with continuous features
- **K-nearest neighbors**: produces high accuracy predictions, especially good if we do not require a human readable model
- **SVM**: good with both classification and regression on linear and non-linear data
- **Random Forest**: ensemble method, which uses multiple repetitions to produce a supposedly better informed prediction

Predicting the Popularity of Songs on Spotify

Loud, Explicit, & Vocal-based Music make Spotify Hits

Fion Ho, Hung Nguyen,
Isabelle Supandji, Josh Zhang, Joyce Mok,
Kaili Nguyen, Rithika Reddy

Analysis

Variable Importance:

- We observed the average scores of characteristics from top 10 genres based on popularity.
- We found high energy and danceability to be potentially good predictors.

Feature Selection:

- The top 5 most important variables were instrumentalness, loudness, speechiness, explicit and valence.
- Knowing that these features were the most correlated with popularity, we chose to focus our analysis on these specific variables.

	Feature	Score
1	Instrumentalness	1041
2	Loudness	291
3	Speechiness	230
4	Explicit	221
5	Valence	187

Performance Evaluation:

- All our models reached an accuracy score between 71% and 77%.
- Among these models the most accurate model generated was the KNN model with 76.7% accuracy and 61.2% F1 score.

	Logistic Regression	KNN	SVM	Random Forest
Accuracy	0.712	0.767	0.717	0.716
Precision	0.493	0.587	0.603	0.674
Recall	0.639	0.032	0.046	0.022
F1 Score	0.061	0.612	0.087	0.042

Conclusion

Overall, we found that the most important variables to determine popularity were instrumentalness, loudness, speechiness, and explicitness, and that K-Nearest Neighbors was the best at modeling popularity. If an artist is hoping to mimic popular releases on Spotify, loud, explicit tracks with a strong blend of vocals and music tend to garner the most popularity. As time goes on, the amount of songs and artists on Spotify will only continue to grow, and this model can serve as a tool for labels to create the next big hit. Thus, that is how we found what type of songs got popular on Spotify.

Thank You!
Questions?

thank u, next

Ariana Grande

0:23 -3:25