SONG POPULARITY PREDICTION

INTRODUCTION

Goal:

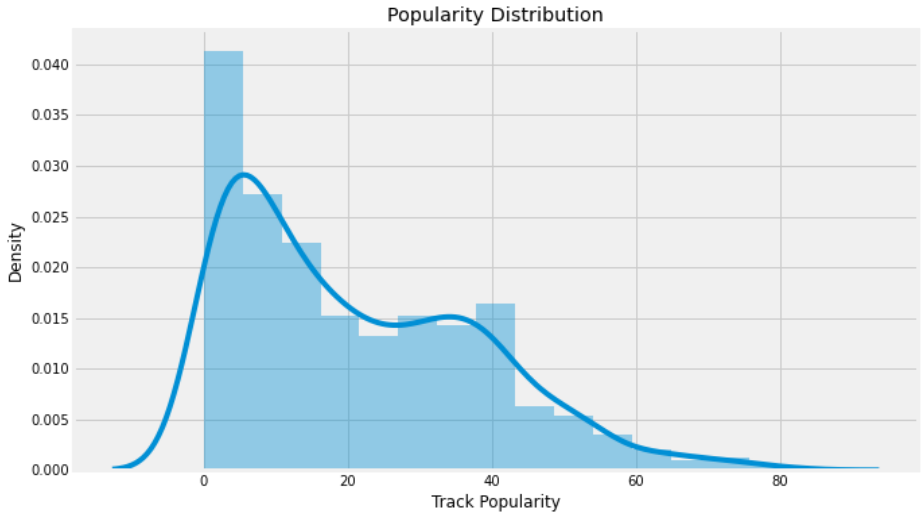
Build a predictive classification model and identify factors/music features that make a song popular.

DATA

- Music tracks featured on South by Southwest 2022
  Official playlist.
  - Focus on artists at a relatively similar career stage;
  - SXSW spans over several days and is completely decentralized;
  - The songs are similarly promoted through the SXSW website and the official Spotify playlist

SONG FEATURES

- Acousticness (ranges from 0 to 1)
- Danceability (ranges from 0 to 1)
- Energy (ranges from 0 to 1)
- Instrumentalness (ranges from 0 to 1)
- Valence (ranges from 0 to 1)
- Track Popularity (ranges from 0 to 100)
- Liveness (Ranges from 0 to 1)
- Loudness (ranges from -60 to 0)
- Speechiness (
- Mode (major (1) or minor (0))
- Key (ranges from 0 to 11)
- Artist Popularity (ranges from 0 to 100)
- Release Date (in yyyy-mm-dd format)



WORKFLOW

Data Acquisition

- Spotify API

Data Preprocessing And Feature Engineering

- Created a target variable using ‘Track Popularity’.
- Dropped tracked released before 2019
- Dropped ‘Artist Popularity’
- Dummified caterogical variables

Modeling

- KNN
- Logistic Regression
- Random Forest
- XGBoost
- Naïve Bayes
- Optimized using grid search

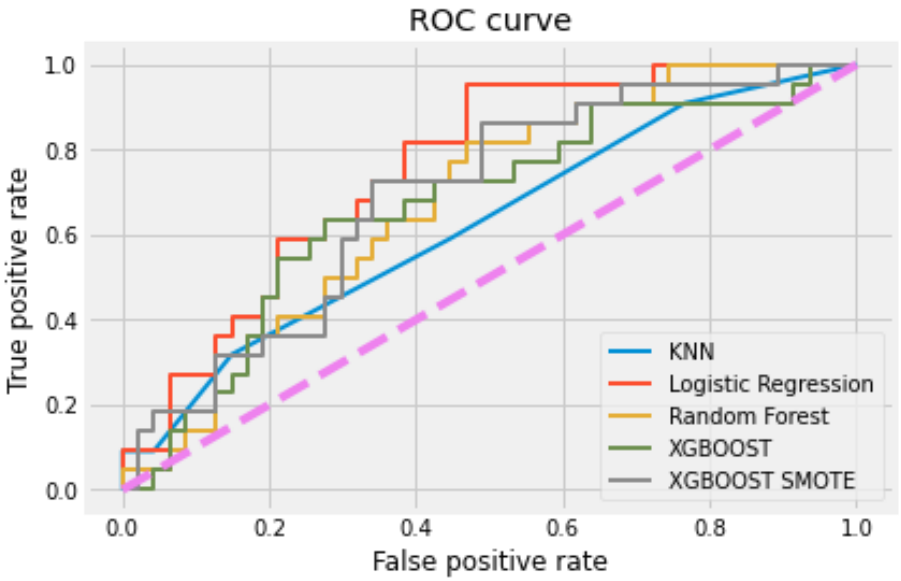
Addressing class imbalance

- Oversampling
- SMOTE

Analysis of feature importance

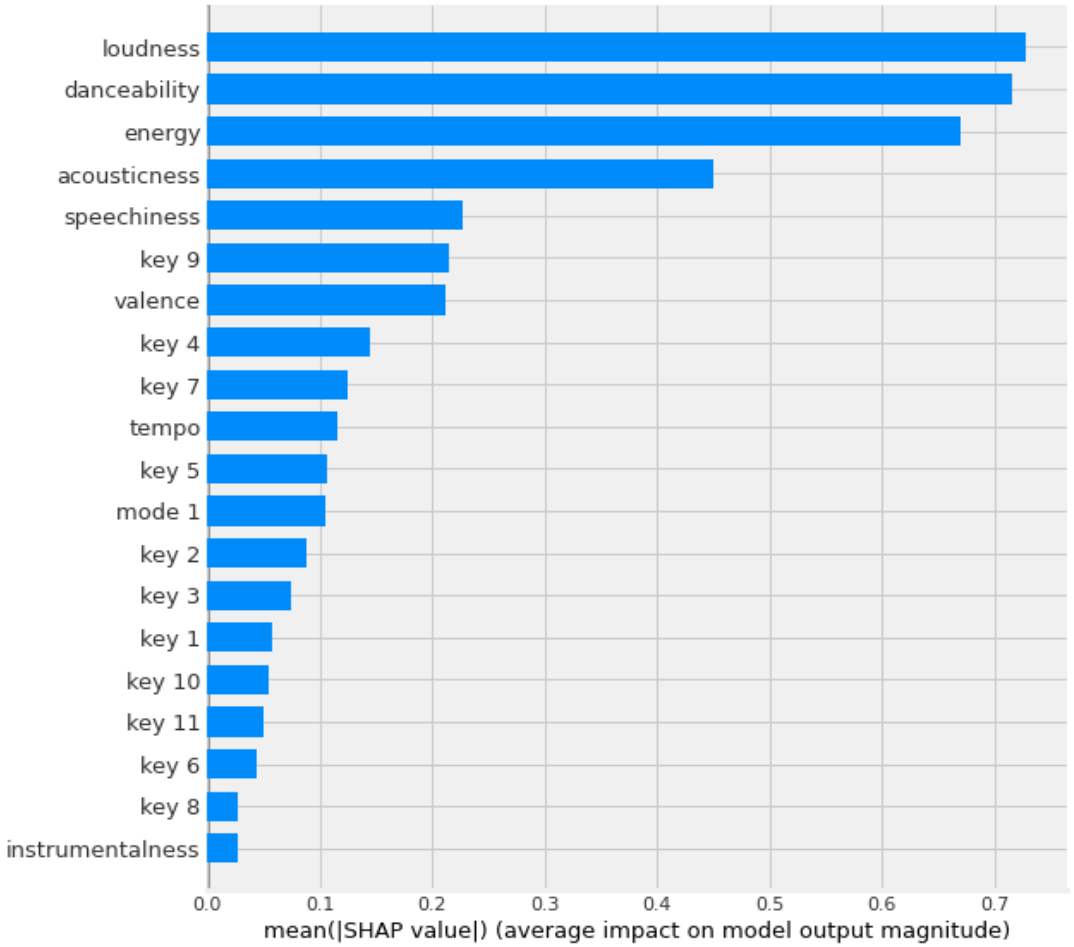
- Logistic Regression coefficients
- Shapley values

RESULTS



- KNN:
  - ROC AUC score = 0.6238
  - F1 score = 0.3889
- Logistic Regression:
  - ROC AUC score = 0.75828
  - F1 score = 0.4444
- Random Forest (Oversampled):
  - ROC AUC score= 0.6809
  - F1 score = 0.3889
- XGBOOST:
  - ROC AUC score = 0.6692
  - F1 score = 0.4126
- XGBOOST (SMOTE):
  - ROC AUC score = 0.6915
  - F1 score = 0.5926

FEATURE IMPORTANCE



FUTURE WORK

- Explore larger datasets
- Proxy for popularity
- Hyperparameter tuning