Classification on Spotify songs with machine learning techniques

Helen Zhang

TL;DR

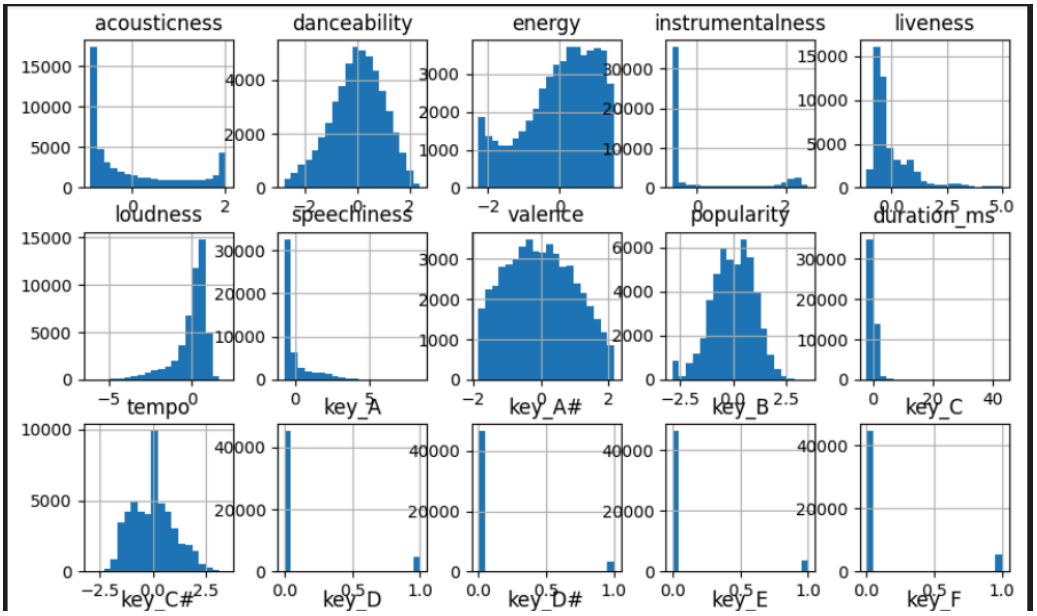
1. **Data Processing**: Import data + clean data + data visualization
2. **Dimension reduction:** PCA/ t-SNE
3. **Models Used**: Logistic Regression / Tree / Random Forest / Adaboost / Neural Network
4. **Best AUC I can get**: 0.9228 (by adaboost)

1. Data Processing

We got the raw data for 50,000 rows and many dimensions.
My training dataset X includes:

- Dimensions did not have missing data: 'acousticness','danceability', 'energy', 'instrumentalness', 'liveness', 'loudness', 'speechiness', 'valence'. These dimensions are normalized.
- duration (get rid of -1), tempo(get rid of "?"), and filled the missing values by the mean of the dataset (filling other values will change the distribution of data). These are not normalized.
- key and mode: changed them to dummy variables, and are not normalized.

Y: label of outcome: music_genre column, changed to categorical variable, not normalized.



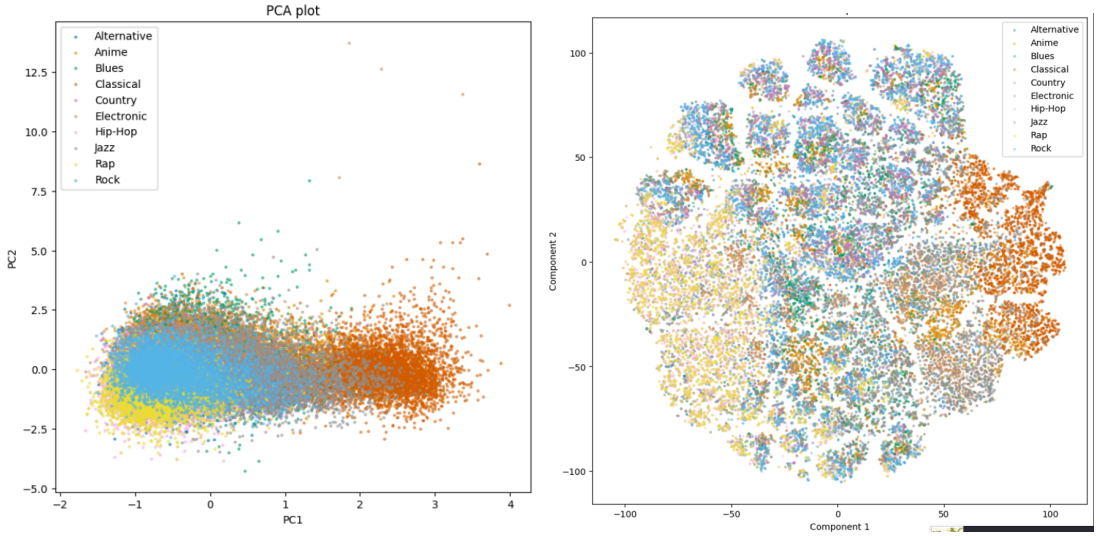
After visualizing the dataset and found that most variables are not normally distributed. Only danceability, valence and popularity are normally distributed.

2. Dimension Reduction

PCA and t-SNE is performed for dimension reduction.
PCA:

- 3 eigenvalues >1: acousticness: 3.706, danceability: 1.347, energy: 1.056
- Total explained variance by these 3 dimensions: about 48% (29%, 10%, 8%)

- 2-D PCA plot: not a good separation between classes. Same for 2D t-SNE.



3. Training

Split test set and training set: used an algorithm to make sure each class has 4500 training and 500 for testing, no leakage.

Method	**Tuned parameter**	**AUC**
Tree	/	0.6851
Random Forest	N_estimators = 200, max_sample = 1, max_feature = 0.2	0.9154
Adaboost	Max_depth = 10, n_estimators = 300, learning rate = 0.01	0.9228
Neural Network	5 hidden layer, 3 ReLU	0.9086

**** Several Notes(very interesting findings to me when I trained the models):**

- These are all the best possible outcomes I got.
- Dimension reduction(PCA and t-SNE) does not significantly increase the outcome of all models, so I choose to stick with the original dataset. I think I should do LDA after reviewing the slides (we have labels for the data) but unfortunately I didn't have time.
- For Adaboost and the Neural network, we want to match the output to genres instead of a random float output. Data needs to be reprocessed so that the output are predicted probabilities of each class, and we will pick the one with greatest probability as the outcome.
- AUC is calculated on a multiclass basis
- Adaboost and neural networks takes longer time to train but have better performance
- The most predictive variable in adaboost model: popularity (model has lowest AUC when this variable is dropped). However popularity does not have the biggest eigenvalue in PCA, which confirms the fact that this data may not be linearly separable.