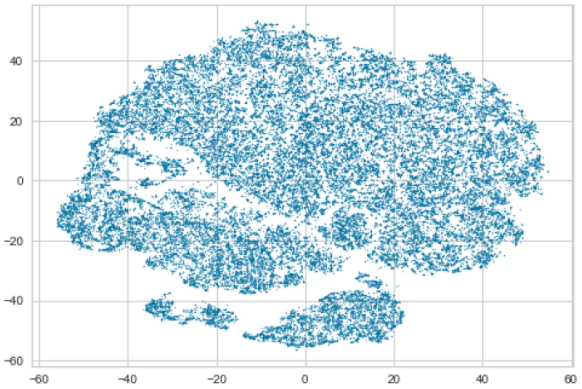
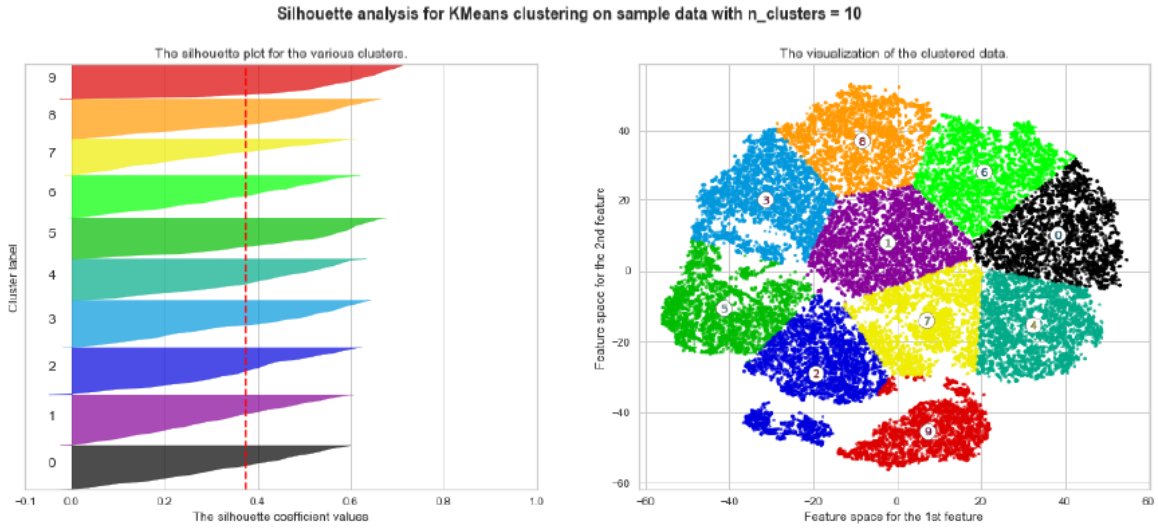
Machine Learning Project Report

Jiangjie Bian jb6942

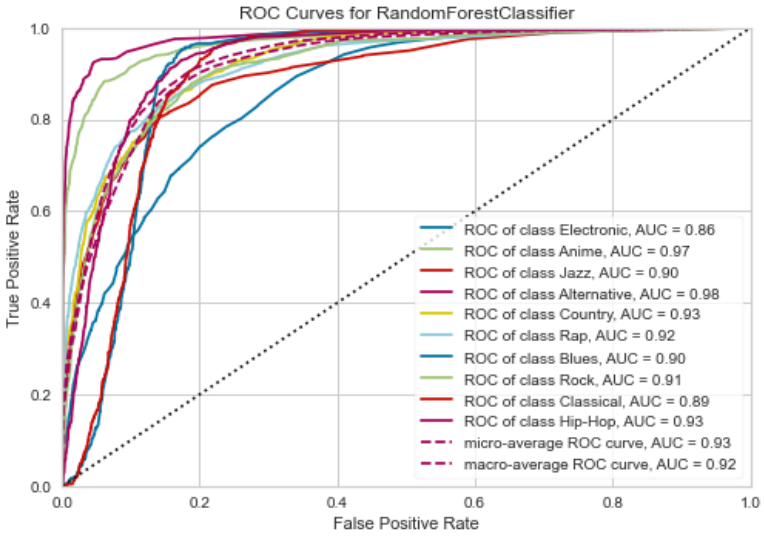
Firstly, I processed the music dataset: I removed the linguistic properties of artist and song; I dropped 5 rows of data containing the ‘nan’ value; I replaced all the ‘?’ value in column ‘tempo’ with 0.0 and converted the whole column from string to integer; I converted string values in column ‘key’ to numerical values and dummy coded categorical value in column ‘mode’; dealing with missing value or categorical or string format value helps me to better utilize these values to build the classification model. I also normalized all the original numerical features excluding acoustic features. Normalization helps to change the values of numeric columns in the dataset to use a common scale. After processing the data, I used the t-SNE dimensionality reduction method on the numerical data for the 2D solution which is shown below. The t-SNE preserves the local data structure and yield a better distribution of data. Then, I used the kMeans clustering method to cluster the t-SNE 2D solution into 10 clusters and measure the silhouette value of the clustering. The kMeans yields the optimal position of each cluster center and the silhouette value helps to justify the clustering make sense. The visualization of the genres as clusters is shown below.



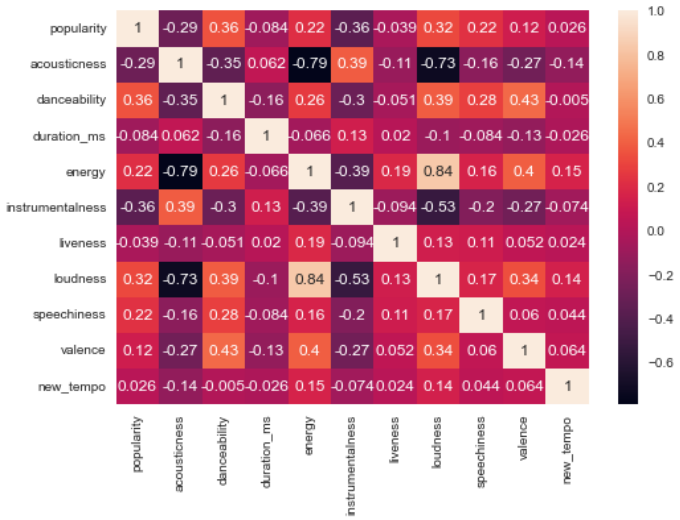


Comparing clusters with the music genre, I noticed that although the t-SNE and kMeans give clear clustering, the clusters are still slightly biased: a few pieces of music with the same genres are in different clusters. After clustering which provide me with the data with their respective

cluster, I added the corresponding cluster number as a new attribute for the music into the data set. I think this is the most important factor which provides the model with another good predictor generated by the overall characteristics of music. Finally, I used the random forest classification method on the dataset, calculating the accuracy and AUC and plotting the multiclass ROC by using the yellowbrick. The random forest is a sophisticated enough model to yield the classification without overfitting. The accuracy and AUC help to measure the performance of the model. The ROCAUC in yellowbrick helps to plot multiclass AUC without transforming category labels of genres into numerical labels. The ROC graph is shown below.



An interesting observation is that, among all music attributes, the energy of the music is strongly positively correlated with loudness and is strongly negatively correlated with acousticness. The visualization of the correlation matrix is shown below.



My Final AUC is 0.92