**Report for Matching Celebrities**

ECE 4950

Yizhou Sun (ys775)

1. Decision for normalize data

According to the result of Assignment 4-5, we know that normalized dataset has a better performance that un-normalized dataset. I decide to normalized data then training the classifier, although , from the data itself we can see that there is no feature whose value is much larger or smaller than others. And by training using normalized, un-normalized data, there is not much differences.

1. Feature selection

Since the features have various attributes such as hair color, etc. We need to select the features that is useful to prediction. Scikit-learn provides some tools for feature selection such as univariate feature selection, recursive feature elimination.

Recursive feature elimination:

Univariate feature selection:

1. Algorithms
2. Decision Tree

fast, 0.71 with selectKBest RFE not much differences. Also Boosting.

Max depth and min leaf size

1. Gaussian Naïve Bayes/Multinomial Naïve Bayes

Not much choices, 0.53

1. Logistic Regression

0.56

1. k-NN

Large dataset not a good idea!

1. SVM

Slow hard to converge, 0.77 not good

1. Boosting
2. Bagging
3. Neural Network

0.78