Predictive Analysis Competition Report

Sheng Luo

Columbia University SPS

Applied Analytics Frameworks & Methods I

12/2/2022

This is how I'm thinking about this. I explored what makes some songs popular at the ideal time. This dataset categorizes well-known songs based on aural characteristics like volume and rhythm. The objective is to create a model that predicts ratings based on the acoustic properties of the songs provided in scoringData.csv using the dataset of well-known songs. The statistic is RMSE, which will be used to analyze the given data (root mean square error). The model is better the lower the RMSE. I imported the dataset first, then eliminated the extra characters from the genre values. Using the genres present in the training and scoring data, we separated the genres into vectors and created dummy variables from the columns of the genre vectors. To guarantee that both sets of data had the same number of columns, the training data and the scoring data were coded. I will divide the dataset into two modules for train and test when classification and data cleaning are finished for most regression prediction analyses. Even if certain features cannot be utilized because they over-fit the training set or exhibit multicollinearity with one another, feature selection typically goes beyond baselines, adjustments, and ensembles, allowing you to employ as many features as you can.

For example, in this contest, one of the features we can use to promote RMSE is 'genre.' I had to populate the genre dummy variables because I wanted to use the

ranger model later. the variables in ranger I picked the most frequently occurring time_signature, tempo, energy, song, track_duration, adult_standards, album_rock, etc. The data source is the train dataset, and the number of trees is 2000. I submitted this final model, and the RMSE run through the ranger model is 14.81134.

Of course, besides trying the ranger model, I also tried using the mice package to eliminate the missing value, but it didn't work very well either. The RMSE of the tuned gbm model could be more satisfactory at 16.3037. The RMSE of the gbm model is also 16.46424, which is even higher than that of the tuned gbm. I also used the RMSE of the tuned random forest, which is 16.43234. The RMSE of the Ipred model is the highest at 18.72334. The RMSE of the Tuned Tree model is 16.46605. All The categorical genre-processed ranger model provides the lowest RMSE, the better model among all the models I have tried. Therefore, the ranger model worked the best.

The good thing I did in this prediction contest is that I tried many models and got the results to compare. The downside is that I should have cleaned and classified the data before the models. I didn't filter the appropriate variables in the model. So next time, I will select only some impact variables. I will also try again with the XGboost model because this is also the best prediction model. It may have a lower RMSE, plus the previous data manipulation. This is where I could have done a better job. I will try again with this method or model.