5200 PAC Competition Final Report

Ho Ting Yi (th3019@columbia.edu)

In this competition, I had 11 submissions and got the final score of 14.97042, ranked 433 in the competition.

In this report, I would elaborate how I have done my best submission, what are the previous try, and share what I had learned in the competition,

1. My final submission

file name : 5200 PAC Final Submission_Ho Ting Yi (th3019)

The first phase is data tidying.

First, I used nearZeroVar function to check if there is any outliers. I would exclude the outlier in this step. Also, from the previous submissions I knew that the variable genre would be the key variable to improve the model. Therefore I tidied this column to become more easy to analyze and build models. I used gsub function to remove the punctuation in the genre. For example, remove slash and parenthesis. I used strsplit function to separate the words in the way I wanted.

```
nearZeroVar(songs,saveMetrics = T)
table(songs$time_signature)
fct_count(songs$genre)

songs4 <- songs
songs4$genre <- gsub("[|]|'|,","",songs4$genre)
songs4$genre <- strsplit(songs4$genre,split)
songs4$genre <- gsub("\\[|\\]","",songs4$genre)
```

Second, to focus on the genre variable, I used mutate function to add variables. I mutate dummy variables of certain genre including “pop”,”rock”,”dance” and “rap”. If a song’s genre has these keywords, it could be reflected in the dummy variable column accordingly. I chose these variables because I did explanatory analysis of the data set and found the genre types that showed up the most in the top 300 ranking songs. In this way, I could predict the rating by having these genre type as a predictor.

```
library(dplyr)
songs5<-songs4%>%
  mutate(dummy_pop= if_else(str_detect(genre, "pop"), 1, 0))%>%
  mutate(dummy_rock= if_else(str_detect(genre, "rock"), 1, 0))%>%
  mutate(dummy_dance= if_else(str_detect(genre, "dance"), 1, 0))%>%
  mutate(dummy_rap= if_else(str_detect(genre, "rap"), 1, 0))
songs5
```

Third, I used mice packages to deal with the missing data. I also use sapply to check after I addressed the missing data and found out there were no missing data after doing this.

```
library(mice)
songs6 = mice::complete(mice(songs5,seed = 617))

sapply(songs6, function(x) sum(is.na(x)))
```

Lastly, I remove the categorical function and the id function to make sure my prediction is only including numerical type so that I could do regression with only numerical numbers.

```
library(dplyr)
songs6<-songs6%>%
  select(-id,-performer,-song,-genre)
```

I tidied the score data in the same way so then I could use the score data to test the model afterwards.

The second phase is splitting the data. I used the caret package and the createDataPartition method to split the sample in to the portion 3:7.

```
library(caret)
set.seed(1731)
split = createDataPartition(y = songs6$rating, p = 0.75, list = F,groups = 40)
train5 = songs6[split,]
test5 = songs6[-split,]
```

The third phase is modeling. I used ranger package to run the model using all the numeric variable and the dummy variables I added. I tried 1000 trees and found out the RMSE is the best comparing to the past models that I had done.

```
library(ranger)
set.seed(1731)
forest_ranger = ranger(rating~.,
                       data = train5,
                       num.trees = 1000)
pred_train5 = predict(forest_ranger, data = train5, num.trees = 1000)
rmse_train_forest_ranger5 = sqrt(mean((pred_train5$predictions - train5$rating)^2))
rmse_train_forest_ranger5

pred5 = predict(forest_ranger, data = scoringData4, num.trees = 1000)
```

This is how I conduct my best prediction model.

2. My previous submission

file name : 5200 PAC Previouis Submission_Ho Ting Yi (th3019)

I had done the previous submission also in three phase.

The first phase is data exploration and data tidying. I used str function, table function and tapply function to get a whole picture of what the data looked like. I checked for the missing data. I also used nearZeroVar function to check if there is any outliers.

The second phase I split the data in to train and test data, trying different method as well as different set seeds. I used simple spilt and createDataPartition in the caret package.

The last phase I used all the model frameworks and tuning methods that had been taught in the lecture. I used linear regression, subseting hybridstepwise, tree, ridge, lasso, ranger and randomforest.

The best score I got here is 15.6.

3. My learning

First of all, I realized that exploring the data and tidying the data are the most important things when I have to address data and do prediction. Technical experience is crucial, but think first after I start using all the method to create model would get better outcome. For example, in my previous submissions, I did not think much about the variables, and exclude the messy categorical variable at first. However, when putting myself as a spotify user, I might use the genre to rate the songs because certain genre are much more popular than the other genre. Therefore, I should consider the genre a potential variable that could improve the model.

Second, tidying the data could really increase the efficiency when modeling the data. If I did not phrase the categorical variables into strings and then to dummies, it would be much more difficult to do all the modeling.

Third, in this competition, I had realized that the world of coding is so large that we always need discussion and improve with classmates. Discussions could lead to more creative ideas. Also, debugging also needs someone else's help because it would be a time that I am so stuck in the errors and could not do anything. Moreover, the forum

and the sharing githubs are also good resources to self learn and improve myself. Learning coding should not be restricted to the lectures, I should do more self learning and always make myself to think as many solutions as I can.

In conclusion, I have learned that always remember to think more rather than just code as much functions as I know. I regard data tidying to be an important part to be more efficient when modeling. Last, self learning is always important when coding, because there's always a better way to solve the problem!

Thank you for reading my assignment!
I really enjoy the class this semester. Thank you, Prof. Lei Yu, for answering all of our questions so patiently and clearly. I love your lecture so much.
Thank you Jocelyn, for helping me in the TA appointment so that I could get clear to the concepts that I was so confused with.