Assignment6

Donghyun Kang

Q1

(a) The results produced by the submitted models were compared to Netflix's internal Cinematch algorithm. The objective function or criterion function was "Root Mean Square Error" (RMES). Only if a model improved Cinematch more than 10% did it become a candidate for a winner of the open call. (Bell et al., p. 24)

(b) It was "nearest neighbors" method. This basically computed the "weighted average rating of similar items by the same user." (Bell et al., p. 25) According to this method, the similarities among movies were measured used metrics such as Pearson correlation or cosine similarity, and the degrees of similarities were used as weights to predict the ratings on movies.

(c) The core source of improvement lies in the fact that the combination (or averaging out) the independent model is better than a single model. The authors implied that the same principle was applicable when it comes to forming a better team, beyond as a modeling strategy.

Q2

(a)
- username scvgood2gosir
- friend key: 1407513_1dGN5QfybPrmoiYDgPRwGC1S4SEYj0Ze

(b) I chose problem 6, sum square of difference problem.

```
ssd <- function(input_seq){
  sum_square <- (sum(input_seq))^2
  squares_sum  <- sum(input_seq^(2))
  result <- sum_square - squares_sum

  result
}
print (ssd(seq(1:100)))
```

```
## [1] 25164150
```

(c) I would like to achieve "As Easy As Pi", "Prime Obsession", and "State Of The Art." First, I decided to choose "As Easy As Pi" award given that the name of the website is "Euler project." Euler's formula, $e^{ix} = cos(x) + isin(x)$ evaluates to $e^{ix} + 1 = 0$ at $x = \pi$. Second, I select "Prime Obsession" because the prime numbers are lonely entities; they cannot be divided evenly by other numbers except for 1 and themselves. I want them to be less lonely. Lastly, I chose "State Of The Art" award. because I felt that the early problems were not enough challenging for me.

Q3

(a) I chose "Website contact finding" task.

(b) The reward for doing this is $0.10. This bascially means that the participants will get the reward after they find the contact information including email addresses and phone numbers and submit it.

(c) There are three qualifications; 1) HIT approval rate (%) is greater than 95; 2) Total approved HITs is greater than 100; 3) Masters has been granted.

(d) According to the decription, the allotted time to complete the task is 10 minutes. I guess I can do this 20-25 times in an hour, which can be translated into $2 - $2.5 per hour.

(e) This task will be expired in 7 days from Nov 13th, 2018, which means that it will be expired on Nov 20th, 2018.

(f) Well, the cost will depend on how much people are willing to do the same task repeately but if we assume that they would do the task for an hour and spend 3 minutes per task on average, it would cost the HIT creator $2 million. However, if we suppose that each person would do this only one time, it would cost $100,000.

Q4

(a) I registered for a Kaggle with my google account.

(b)

- The title of competition that I decided to choose is "Google Analytics Customer Revenue Prediction: Predict how much GStore customers will spend." This competition is sponsored by Rstudio together with Google Cloud. Rstudio not only provides the free Integrated Development Environment for individuals but also a commercial license with more functionalities that can be scaled up. Google Cloud provides a cloud computing platform in which offer a high-speed distributing computing and a reliable data sharing. The submitted models in which predict the natural log of revenue for an individual customer are evaluated with "Root Mean Square Error" (RMES), which was utilized by Netflix Prize.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

- There are two types of prizes. First, the 1st, 2nd, and 3rd place participants on the private leaderboard will receive $12,000, $8,000, and $5,000, respectively. The participants can get the winning money regardless of the programming languages they use. The second type of prize targets the participants who build high-performing models using R. Those participants' models will be reviewed by the Sponsor; the 1st, 2nd, and 3rd place participants will win $10,000, $7,000, and $3,000, respectively. A participant can get both prizes. For example, if the 1st place participant in the leaderboard develops a model using R, and volunteer to get the model verified by the Sponsor, that participant can $22,000 in total.
- As for the honor codes, the participating teams are not allowed to use multiple accounts and their codes are only supposed to be shared within a team, not outside of a team. (The number of members in a team should be less than or equal to 8.) And the rules for this competition state that the participants are not supposed to conduct unlawful submissions which include the violation of various rights such as intellectual property rights, trademark, patent, etc.
- The participants are allowed to submit their codes only five times per day and they can select two submissions to get judged in the end. The teams can be merged by the team leader but a new team's submission count has to be within their range as the date of merge date.
- The start date was Sep 13,2018; the deadline for merger and entry is Nov 23, 2018; the competition ends on Nov 30, 2018 11:50 PM UTC.

(c) This competition is sponsored by Google Cloud and Rstudio. The competition description states that the sponsor hopes to get results that can be applied to better operate their stores (GStore). I think Rstudio is not going to directly utilize the winning result because they seem to have a different intention, promoting Rstudio and R as a tool for data analytics.