1.

In this paper, authors describe their own experience of participating the Netflix Prize open call contest. This contest was sponsored by Netflix and participant's mission is to train their model based on more than 100 million ratings of movies so that their model could precisely predict holdout set of 3 million ratings.[1] The criterion function of this contest is to have the highest improvement in root mean squared error (RMSE). The RMSE could be represented as follows:

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

Y hat here is the prediction value calculated by the model and y is the actual rating. The final value is rounded for four decimal places. The cutoff for this contest is to improve the Netflix's internal algorithm, Cinematch, by 10% or more.

At the beginning of the Netflix Prize contest, the most commonly used method is the 'Nearest Neighbors'. With this method, the predicted rating for an item would be 'a weighted average rating of similar items by the same user'[2]. However, the drawbacks of this method are also obvious. It's rather hard to choose the similarity metric and highly correlated neighbors

[1] Bell, Robert M., Yehuda Koren, and Chris Volinsky, "All Together Now: A Perspective on the Netflix Prize," Chance, 2010, 23 (1), P24

[2] Same paper, P25

of a movie could lead to ‘double counted’ problem. What’s more, this method could hardly be used for movies with few or no close neighbors.[3]

The model of authors is a combination of more than 100 prediction sets. The reason why authors combined so many models is that an implement of a new model would improve the accuracy of the whole model as long as this new one’s correlations with other parts are not high.[4]

Reference

Bell, Robert M., Yehuda Koren, and Chris Volinsky, “All Together Now: A Perspective on the Netflix Prize," Chance, 2010, 23 (1), 24 - 29

[3] Same paper, P26

[4] Same paper, P28

2.

(a)

My user name is:“mikepeng” and my friend key is:1410648_pPJQPuI6ydb89E7MoSvnXPq90MhxQ3Za

(b)

I have choosen the problem 1:“Multiples of 3 and 5”. I used python to solve this problem and the code is listed as follows:

```
In [1]: rv = 0
        for i in range(1000):
            if i % 3 == 0 or i % 5 == 0:
                rv += i
        rv
```

```
Out[1]: 233168
```

(c)

For me, the three awards that I would most aspire to achieving are “High Five”,“One In A Hundred” and “Big Game Hunter”.

“High Five” means that I have solved the five most recent problems. By achieving this award, I could keep pace with times and learn most recent skills to solve problems.

“One In A Hundred” means that I’m among the first hundred to solve a problem. By achieving this award, I have made my own effort to a cutting edge problem and I hope latter people could be inspired by my code in the thread part of that problem.

“Big Game Hunter” means that I have solved 25 of the 50 hardest problems. By achieving this award, I have proved my ability of coding to some extent and I’ll be proud of my progress.

3.

a) The project I selected is called “Clean Up How-To Questions” by “Content Research”. This project ask people to view a short question asked by a reader of an online how-to guide, and rewrite it to be clear, concise, and free from spelling and grammar errors.

b) The payment structure of this project only consists of one part. People will get $0.05 for each question they have cleaned up. There is no extra bonus in this project.

c) There is only 1 qualification required in this experiment. People who want to be a question editor must have been granted by the “content Research” (the author). Therefore, before you start your job, you must submit your request to the author.

d) The allotted time for this task is 20 minutes. Considering that I have an average English level, I think I could clean up 3 questions in an hour. The implied hourly rate for this task is $0.15/hour.

e) This experiment will expire on 11/23/2018, 07:03AM.(After 5 days from 11/18/2018)

f) Considering that each person who takes this task could get $0.05, therefore, the maximization of this project would cost the HIT experiment creator if 1 million people participated in the task is 50 thousand dollars.

4.

a) My account name is: Mike Peng.

b) There are 13 open competitions in Kaggle now and I'd like to choose the first one for further description.

The title of this competition is "Two Sigma: Using News to Predict Stock Movements". The sponsor of this competition is Two Sigma Investments. This company is a scientifically driven investment company. Their employees follow principles of technology and innovation as much as principles of investment management. They involved machine learning and distributed skills in their investment and risk management evaluations.

In this competition, the participant needs to calculate a signed confidence value for each asset in the dataset. This value represents the participant's attitude towards this asset, while positive value stands for positive return in the future and vice versa. What's more, the absolute value of the confidence value also represents the confidence you possess toward your prediction. The submission score will be calculated as the mean divided by the standard deviation of x_t, where x_t is the sum of every asset's confidence value multiplied by the 10-day market-adjusted leading return for day t. In another word, the more accuracy your prediction is, the higher score you will get.

The total prize for this competition is $100,000. The 1st will get $ 25,000; the 2nd will get $ 20,000; the 3rd will get $ 15,000 and the 4th to 7th will get $ 10,000 each.

As for the honor code of this competition, the Kaggle emphasizes that participants should never "cheating". “Cheating” here means that train your model by using dataset which is not provided by the official organizer or using the provided information in a way that is not intended.

For the timeline of this competition, it starts at 9/25/2018 and ends at 7/15/2019. In this period, it could be subdivided in to a Submission period and a Scoring period. In the Submission period, participants will train their models in Kaggle Kernels. During the Scoring period, participants should select their two best submissions to be scored.

Finally, for the submission instructions, it highlights that all submissions will occur through the Kernels environment. Participants must use the custom python module of Kernel to access the competition data, step through time, collect predictions, and write an appropriate submission file. What’s more, the compute constraints for this competition are: 16GB Memory; 6 Hours total runtime and 4 CPU cores.

c) Considering that the sponsor is an investment company, they could use that wining model to implement their current investment model to increase their prediction accuracy and build a relationship between media data and financial system.