1. Netflix Prize and Bell, Koren, and Volinsky (2010).

The submissions to the Netflix Prize open call contest would be judged based on their improvement over Netflix's internal algorithm, Cinematch. The measurement for improvement is the root mean squared error (RMSE). The formula for calculating RMSE of an estimator $\hat{\theta}$ regarding an estimated parameter θ is[1]:

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}.$$

According to the Netflix prize rules[2], submissions will be considered as qualified entries if the algorithm is written in English, originally developed and did not require any third-party software/licenses and payment on the part of Netflix. Besides, the submissions might be disqualified if they did not submit both their source code and their algorithm description within one week upon qualifying. All qualified submissions will be reviewed and judeged. Hence, there was no cutoffs regarding the fit quality beyond which a submission would not be judeged.

At the beginning of the Netflix Prize contest, the most commonly used method for predicting ratings was nearest neighbors (Bell, Koren, and Volinsky 2010, p.25). This collaborative filtering method predicts a user's rating for a certain movie using "a weighted average rating of similar items by the same user" (ibid.). Though the way this method works and the results it produces are intuitive and easy to explain to users, this method is problematic in several ways (ibid. P.25-6). To start with, the measurement of similarity is based on arbitrary choice of metric. Moreover, the design of the relative weights does not take into account the effect of the composition of a neighborhood. In addition, for movies without enough similar rated counterparts, the predictions are unreliable.

Bell, Koren, and Volinsky (2010) mention that as long as a model is not highly correlated with the other models, the combination of them will improve the RMSE. Therefore, the characteristic of being not correlated with the other models made one model improve the overall prediction when blended with the other models.

---

[1] See the wikipedia of RMSE https://en.wikipedia.org/wiki/Root-mean-square_deviation

[2] See https://www.netflixprize.com/rules.html or https://www.netflixprize.com/assets/rules.pdf

2. Collaborative problem solving: Project Euler

(a) dt1996: 1408664_htxCIknhTifIr6BodTa0m9HuuTk1Kz87

(b) Problem 1: Multiples of 3 and 5

Problem description: If we list all the natural numbers below 10 that are multiples of 3 or 5, we get 3, 5, 6 and 9. The sum of these multiples is 23. Find the sum of all the multiples of 3 or 5 below 1000.

My codes:

```
In [9]: sum = 0
        for n in range(1000):
            if n % 3 == 0 or n % 5 == 0:
                sum += n

In [11]: sum
Out[11]: 233168
```

My answer is 233168.

(c) The three awards that I would most aspire to achieving are: As Easy As Pi, Fibonacci Fever, One In A Hundred. I like the first two awards because it's fun to choose the problems based on criteria that involve some sort of meaning and randomness simultaneously. I like the last one because it creates a competitive atmosphere.

3. Human computation projects on Amazon Mechanical Turk

(a) I select a task with the title "Classify Receipt". The requester is Scoutlt. This task asks workers to look at a receipt image and identify the business of the receipt.

(b) The payment is $0.03 for classifying one receipt.

(c) The eligible participants are those who (1) have a HIT approval rate (%) is over 97; (2) currently stay in the United States; (3) have total approved HITs more than 1000.

(d) This job takes roughly 20 minutes to finish. The hourly rate is $0.09.

(e) This job expires on November the 25th.

(f) The maximum cost of this project would be 0.03*1,000,000 = 30,000 dollars.

4. Kaggle open calls

(b) One of the open competitions listed on Kaggle is “House Prices: Advanced Regression Techniques”. It is a “getting started competition” provided by Kaggle data scientists and designed for participants with little machine learning background. It aims at helping people who want to get involved in Kaggle to familiarize themselves with the Kaggle’s platform, acquire the basic skills in machine learning, and get to know people in the Kaggle community.

Based on a training dataset with 79 explanatory variables on (almost) all the aspect of residential homes in Ames, Iowa, the participants of this competition are asked to predict the final price for each home in another test set. The submissions are evaluated on the Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted price and the logarithm of the observed sales price.

There is no cash prize for this competition. A participant can only sign up one Kaggle account. The competition allows for sharing codes to all participants on the forums, yet it forbids private sharing of codes or data outside of teams. There is no maximum team size limitation. In addition, the team leaders are allowed to merge their teams with other teams. This competition started from August the 30th, 2016 and runs indefinitely with a rolling leaderboard that invalidates entries after two months.

The maximum number of submission that each team can make per day is 5. The team can choose up to 2 final submissions to be reviewed and judged by Kaggle. The submission file need to include a header and have the following format:

Id,SalePrice

1461,169000.1

1462,187724.1233

1463,175221

etc.

(d) This competition works mainly as a practice for machine-learning beginners who want to get involved in Kaggle and it won’t result in an absolute winner under a rolling time line. Therefore, the answers might not be used for anything. But it is also possible that Kaggle cooperates with some real estate companies and uses the high-quality solutions to set prices for homes.