MIT-GS-Vetr Online Prediction Market

Dhaval Adjodah, Yan Leng, Shi Kai Chong, Peter Krafft,
David Shrier, Alex Pentland

1 Introduction

This document highlights initial findings determined from six rounds of online prediction run over the past 6 months. We obtained a total of 5,122 predictions made by 891 students during the summer rounds, and 10829 predictions from 716 students in the fall rounds with unweighted accuracies ranging from 4.32% error in round 2 to as little as 0.16% error in round 3 and 6. We also beat official futures market forecasts such as in round 1 where our students were less volatile and more accurate than futures market predictions during the Brexit market crash of the SP500 prices. The average (cumulative) error of the collective prediction varied over time but generally converged to the true value. In Figure 1 we can see how the aggregate prediction changed for round 3; in Figure 2 we can see how much each individual prediction varied for all rounds (error decreases over time as shown by the convergence of values over time) .

2 Weights

Overall, we also find that weighing predictions made by students who have done historically better decreases error significantly: the weighted cumulative mean prediction of the collective, as shown in Figure 3 where historical weights are being used, is better than the simple unweighted mean prediction.

Different weights can be used, such as past historical accuracy as well as social weights: how much students improved their estimates after seeing the collective estimate of their peers (i.e. how much their pre-social prediction varied from their post-social prediction where 'social' here refers to viewing the histogram of their peers' values).

These social weights were calculated as values between 0 and 1 to indicate how much users were influenced by the histogram that they were shown, 0 meaning the users were not influenced and 1 meaning that the users were the most influenced by the histogram. Pre-histogram predictions are generally better for users with low social weights (they are influenced less) , but they also have the least improvement in their post-histogram predictions. Users with high social weight, although having poor performance initially, tend to perform better than their low social weight counterparts after being exposed to the histogram as shown in Figure 4 below:

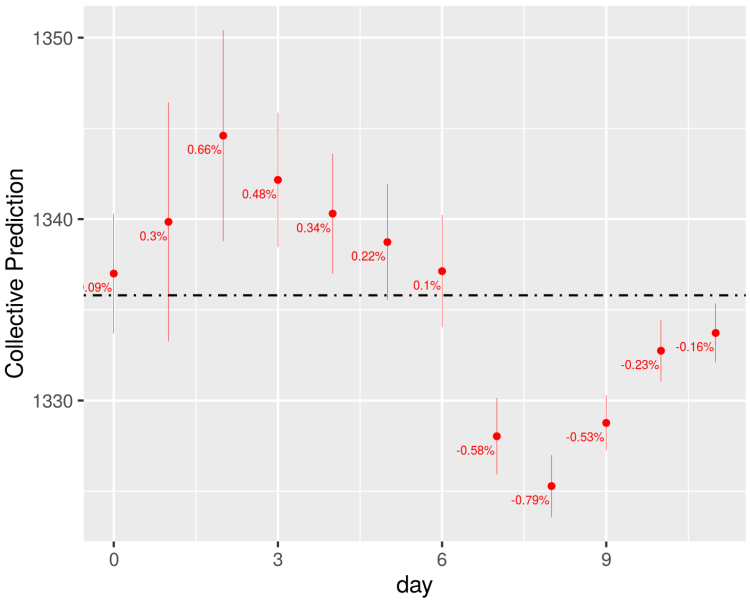

Figure 1: Error (as a percentage) progression over time (last day is the prediction deadline) and dotted line is the true price.

3 Past Round Performance and Learning

The graph in Figure 5 below plots the change in a user's prediction deviation from the mean of the histogram versus their average percentage error from the previous round. Change on the y-axis is defined as such:

$$change = \frac{|Postsocial\ Prediction - mean| - |Presocial\ prediction - mean|}{mean}$$

It shows that for users who performed poorly in the previous round, they learn to deviate less from the mean of their peer's predictions.

4 Feedback Improves Accuracy

One of the novel introductions we made in this prediction exercise was to show students their peers' predictions. The hypothesis was that if students are shown their peers' predictions, they might learn to predict better. This goes counter to the general theory of the 'Wisdom of the Crowd' whereby many people are asked to give an answer to a question, and the average of everybody's answer is close to the truth. In this theory, people overestimate and underestimate equally, and therefore the overall error is statistically canceled. By this logic, the wisdom of the crowd might break down if people are exposed to each other's values because answers – and therefore errors – become correlated instead of canceling

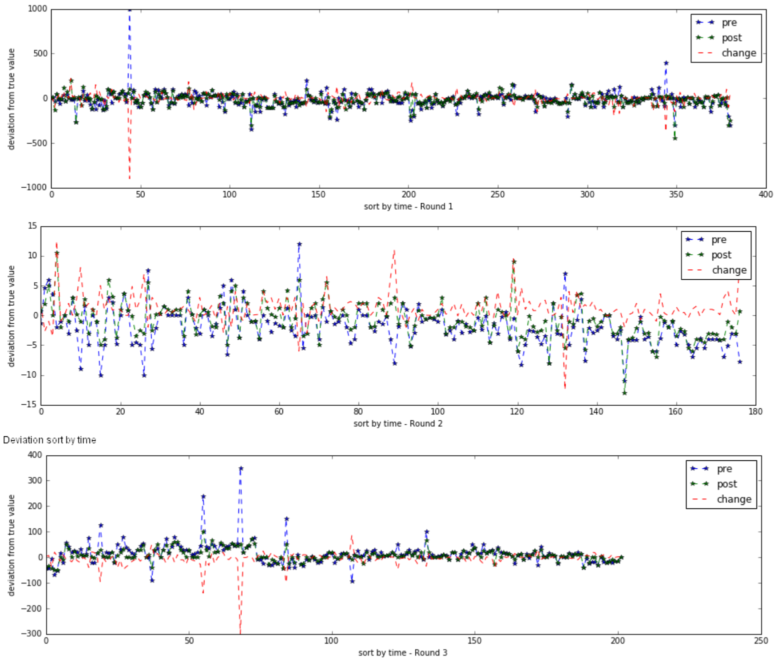

Figure 2: Variation of each individual predictions over time for all rounds

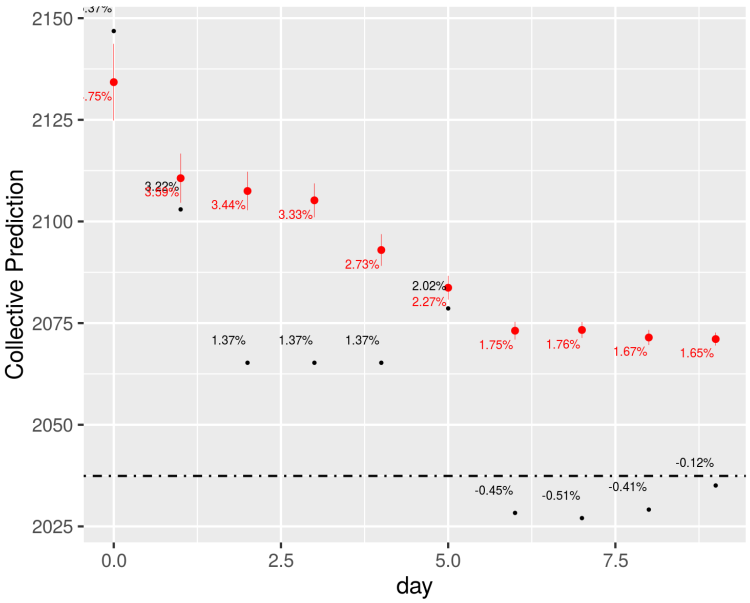

Figure 3: Weighted (in black) and unweighted (in red) error over time

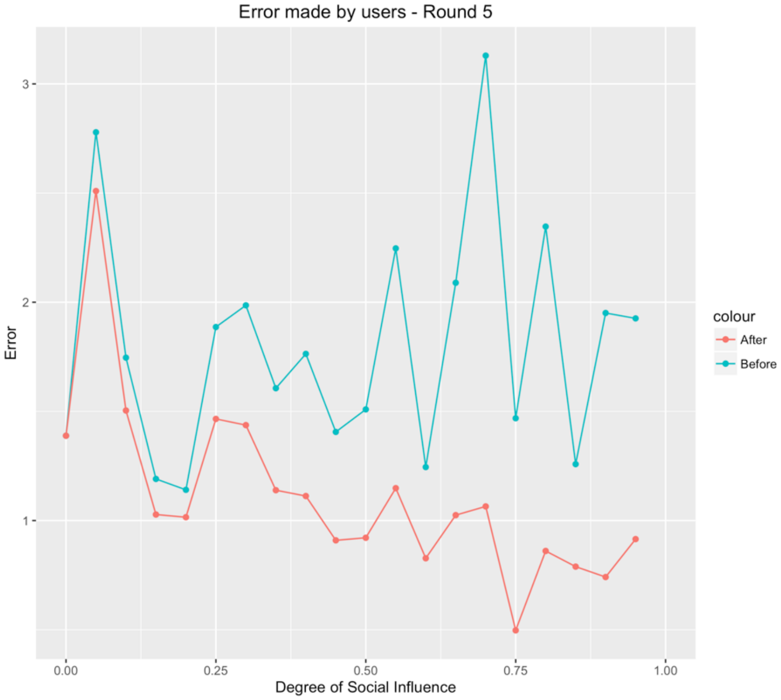

Figure 4: Error and Social Weight

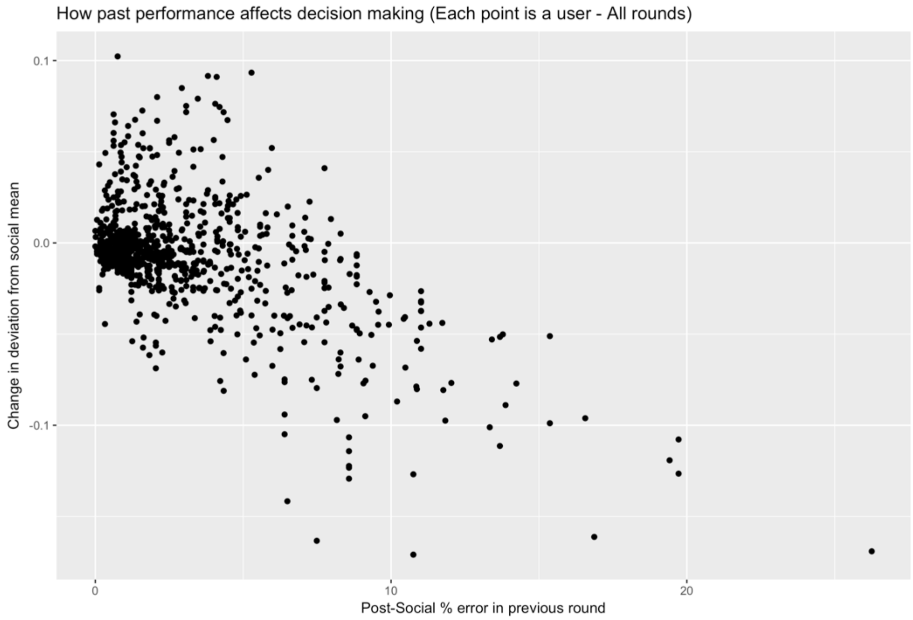

Figure 5: Change in a user's prediction deviation from the mean of the histogram versus their average percentage error from the previous round

each other out. Because we recorded both the pre and post-social predictions, we were able to test this. As can be seen in Figure 6, we found that showing peer information does not make prediction worse: in round 1 and 2, the median pre and post predictions are virtually indistinguishable; in round 3, the post value is actually better than the pre-predictions, meaning that seeing peers' predictions made the crowd more accurate.

Additionally, we gave feedback to each student as they were making their predictions, letting them know whether they had been more or less accurate after seeing their peers' predictions in the past. The idea was that if somebody was on average worse after seeing peer information, we would like them to 'trust' their peers' predictions less and revise their social prediction. As can be seen in round 2 and 3, the revised predictions are even more accurate than the post values. We do not have these revised predictions for round 1 because we needed prior history to show these values. Figure 7 shows the full histograms of pre, post, and revised values for each round.

In general, the number of users who benefited from exposure (meaning that their average performance improved) is about the same as the number of users who did not benefit from social exposure (where their average performance worsened). In rounds 4 and 5, there are a large proportion of user who did not adjust their predictions after viewing the histogram of other user's predictions as be seen in Figure 8.

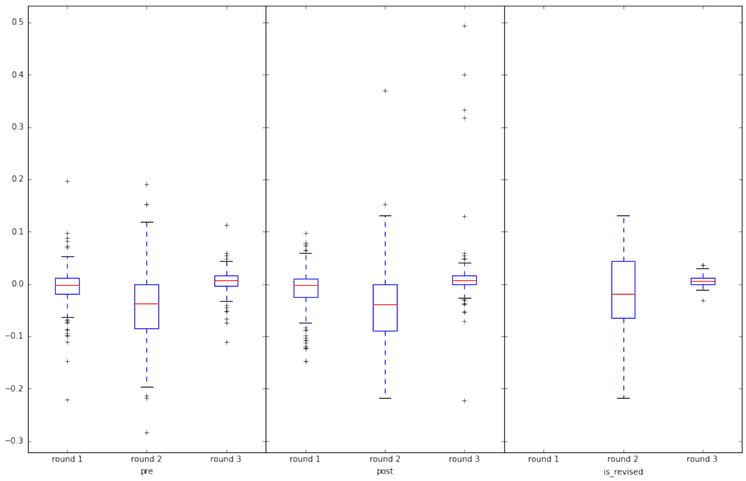

Figure 6: Wisdom of the Crowd, pre and post-social information, and revised values after feedback

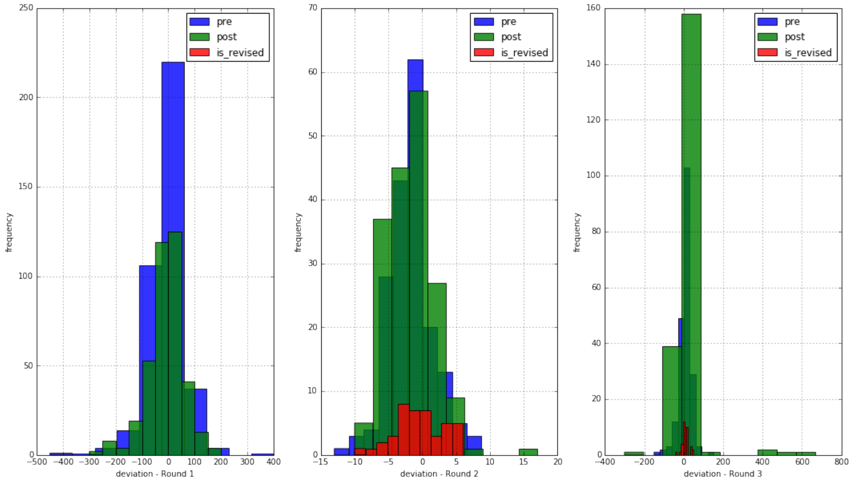

Figure 7: Full histograms for pre and post-social information, and revised values after feedback

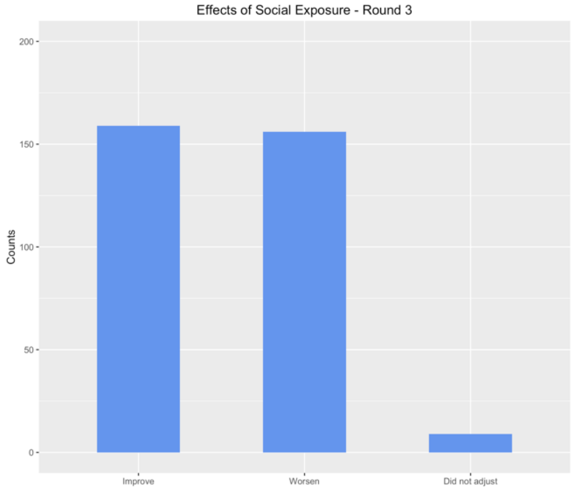

Figure 8: the number of users who benefited (or not) from exposure

Even though the number of users who benefit is roughly the number of users who did not improve after looking at their peer's predictions, those who benefited improved by a larger margin as compared to those whose predictions worsened. As a result, there is an average improvement of predictions across the board for all rounds. The graph in Figure 9 below shows the improvement of each of the users.

As observed by previous research, predictions made after histogram exposure have a lower standard deviation. Predictions for all rounds generally follow a log-normal distribution as can be seen in Figure 10, with some rounds having lighter tails.

5 Correlations

We also gave each student a brief survey after they made a prediction. For instance, we asked them how much experience they have with regards to this particular financial instrument, or how long they have worked in finance. The most variation we found with respect to how much people change their pre and post prediction was with years of experience as shown in Figure 11.

6 Models of Learning

In order to model the actual decision process of each student as they decide what their predictions are after seeing their peers' predictions, we tested different models of learning. The most straightforward one is known as DeGroot learning

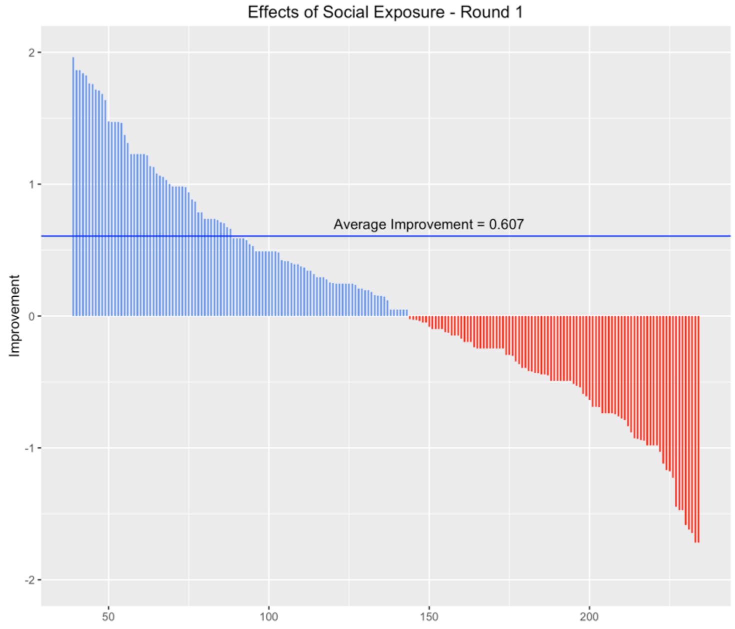

Figure 9: Those who benefited improved by a larger margin as compared to those whose predictions worsened

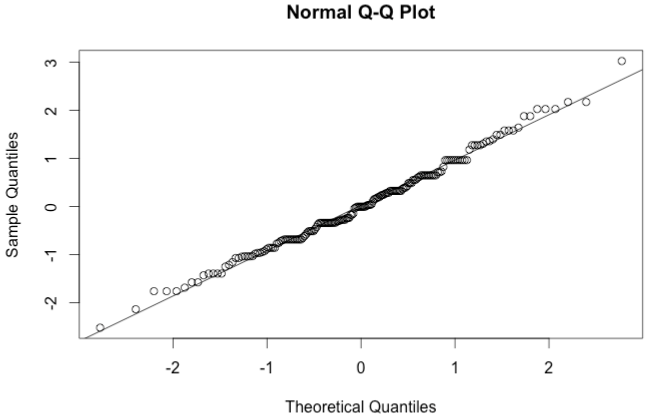

Figure 10: Predictions for all round generally follow a log-normal distribution

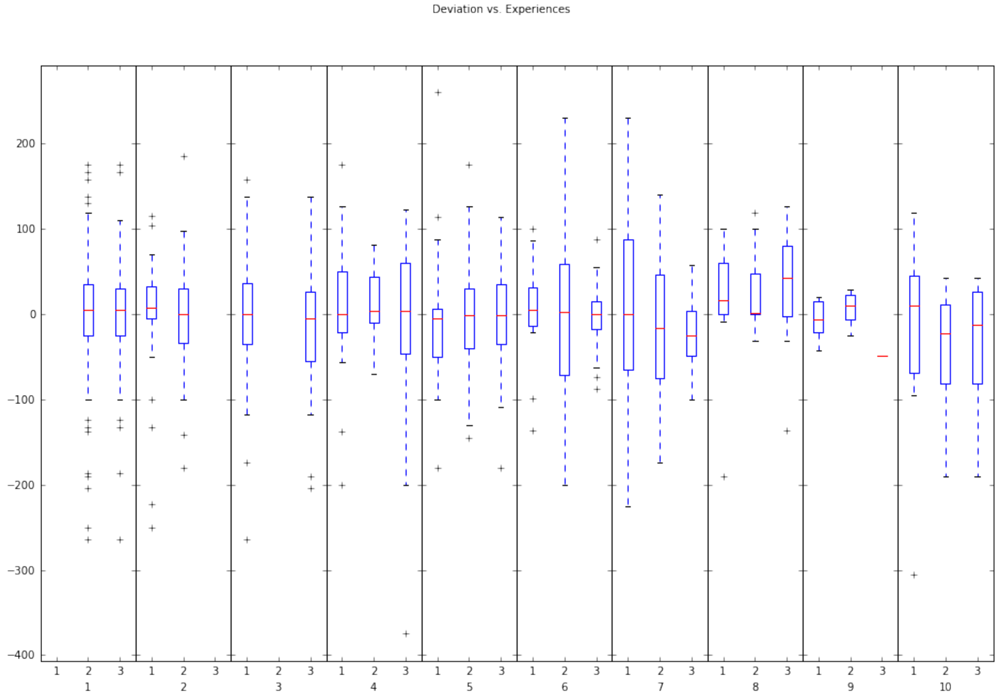

Figure 11: Variation of changes between pre and post-social information, and experience

whereby the post-social prediction of each individual is the average of all the information she has access to, including her own.

We also tested various Bayesian learning models whereby the pre-social value is somebody's prior. We tested whether this prior is normally distributed (using the stock's volatility as the standard deviation) or uniformly distributed. The histogram of their peers' values they see is the likelihood (we test using both a normal approximation of the likelihood histogram, and using the full empirical values of the histogram). We use the posterior of this social learning model to predict the crowd prediction by testing how the mode and mean of this posterior distribution performs. As can be seen in Figure 12, empirical predictions do better than normal approximations, and do almost as well as the DeGroot learning model as expected from past research.

People demonstrate Bayesian Learning in S&P 500 market: Table 1 combines the three rounds related to S&P 500. This includes 9334 predictions from 2196 individuals. The p-value of **normEmpiricalPostMode** (the case where the likelihood was modeled empirically as opposed to being approximated, and the mode of this distribution was used, and the prior is normally distributed) is larger than 0.05, which indicates that crowd's learning in S&P 500 predictions follows Bayesian Learning - with normal distribution to proxy the likelihood and mode from the posterior as the final prediction. Using a statistical explanation, the distribution of actual post-histogram prediction and **normEmpiricalPostMode** are not significantly different. We use the normal distribution to proxy the likelihood and the mode of the posterior distribution as the predicted value.

Table 1: KS test of social learning on S&P 500 [1]

	KS Stats	p-value
bayesian	0.181	0.000
deGroot	0.181	0.000
normEmpiricalPostMean	0.041	0.000
uniEmpiricalPostMean	0.414	0.000
normEmpiricalPostMode	**0.016**	**0.202**
uniEmpiricalPostMode	0.337	0.000

.

People demonstrate Bayesian Learning in Gold market: Table 2 summarizes results related to Gold prediction round which includes 2751 predictions from 552 individuals. The p-value of normEmpiricalPostMode is larger than 0.05, which indicates that crowd's learning in Gold predictions follows Bayesian Learning - with normal distribution to proxy the likelihood and mode from the posterior as the final prediction.

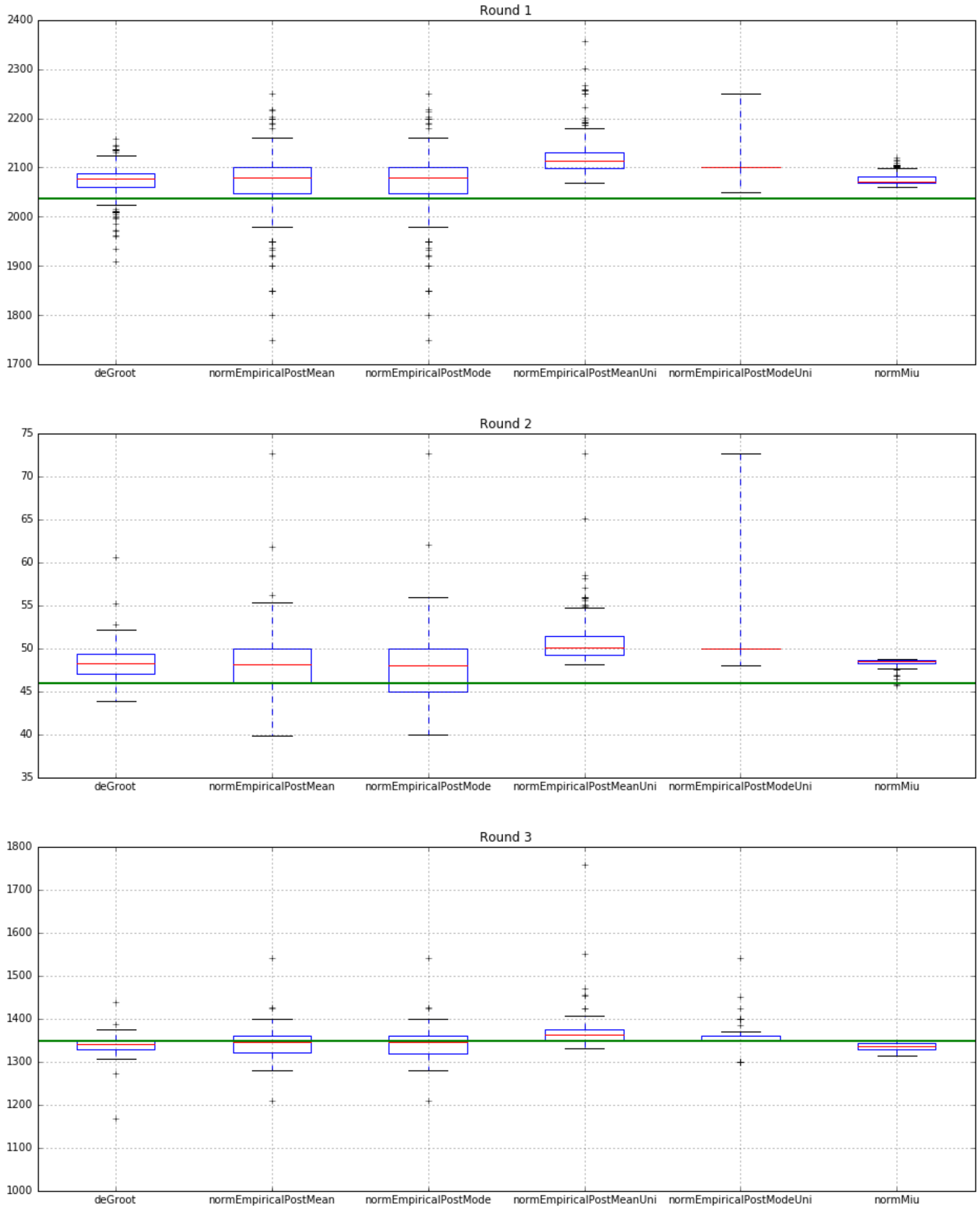

Figure 12: Predictions of different decision models. normEmpiricalPost refers to using empirical values of the likelihood/social histogram, mean and mode refers to using the mean and mode of the posterior distribution. normMiu refers to using normal approximation to the prior, likelihood and posterior.

Table 2: KS test of social learning on gold market

	KS Stats	p-value
bayesian	0.304	0.000
deGroot	0.301	0.000
normEmpiricalPostMean	0.054	0.001
UniEmpiricalPostMean	0.339	0.000
normEmpiricalPostMode	**0.035**	**0.070**
uniEmpiricalPostMode	0.428	0.000

People do not demonstrate Bayesian or Degroot learning in Oil market: As shown in Table 3, people do not follow Bayesian or Degroot learning in oil market. We discovers several unusual behaviors in the predictions in this market. More founds in this market will be helpful in better analysis.

Table 3: KS test of social learning on gold market

	KS Stats	p-value
bayesian	0.215	0.0
deGroot	0.217	0.0
normEmpiricalPostMean	0.086	0.0
normEmpiricalPostMeanUni	0.514	0.0
normEmpiricalPostMode	0.116	0.0
normEmpiricalPostModeUni	0.682	0.0

7 Machine Learning

Finally, we also ran many machine learning models to see if we could do better than the crowd using the predictions, survey questions, and discussion forum interaction (net). As can be seen from Figure 13, they do about as well as each other. The models tested were simple linear regression, lasso regression, Support Vector Regression (svr), Elastic net regularization, Random Forests, and the same models including interaction data (net).

8 Score Calculations

The most common question we had from students was how their score was calculated. First, we calculated the error of each individual prediction: how close each prediction by each student at each point in time was to the real value. This error was then scaled (using exponential smoothing) in terms of how early the prediction was made – e.g. if a prediction was made really early, it earned more points than if it was made the day before the deadline. The score of each student was then the sum of 1) the smallest time-scaled error, 2) the last time-scaled error, 3) the number of predictions and 4) the average time-scaled error of all predictions made by this student. We gave the most importance to the last prediction's error and the mean prediction error. This score was then used as a measure of historical performance, which was in turn used as weights

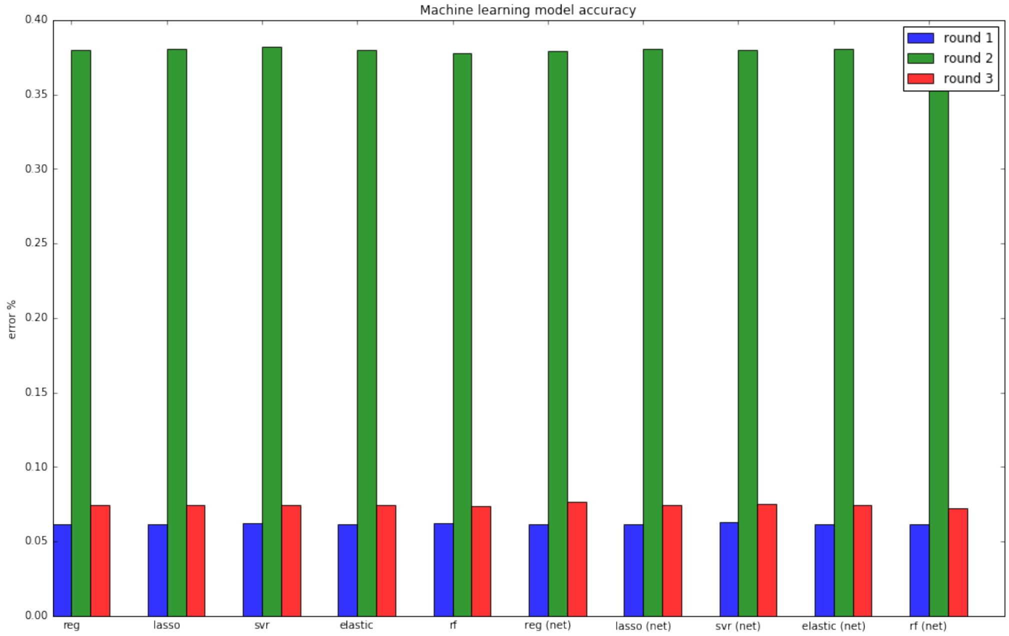

Figure 13: Predictions of different machine learning models.

to improve predictions.

9 Conclusion

We hope that this document answers the questions that many asked us throughout the course. We welcome any more questions you might have – please contact researchers at fc-fall-16-research@mit.edu. If you would like to see the code we use, we will be happy to share as promised.

We would like to end by thanking all the students who participated in this exercise. We really appreciated all of your participation, feedback, questions, and suggestions for improvement.