Efficiency in prediction markets, evidence from SciCast.org

Nicolás Della Penna, The Australian National University & NICTA
Dhaval Adjodah, Massachusetts Institute of Technology, Media Lab
Alex Pentland, Massachusetts Institute of Technology, Media Lab

1. INTRODUCTION

Prediction markets have been proposed as a mechanism for collective information elicitation. We analyze data from SciCast, an online prediction market focused on events of scientific interest, to understand the predictive power of the prices to forecast events. Users receive a fixed number of points (5,000) when they join the site, and can trade contingent claims on the different outcomes of the questions of interest using an automated market maker. Prizes - amazon gift vouchers valued in total at 80,000 USD - are given to members who make the most profit for the subset of questions that are prize-eligible.

As can be seen in the plots below, prediction error decreases over time down to less than 10% on average for all questions on the system, showing that the SciCast prediction market is relatively accurate for predicting outcomes.

Our main finding is that the market aggregates all information available to it into current prices. We tested a number of models and combinations of variables and found that including any variables other than price does not lead to any improved prediction. The price current price at any moment appears to be the the best predictor of future outcome.

2. LITERATURE REVIEW

Previous literature has highlighted the efficiency of prediction markets at aggregating information into the price - as we found in our work - but there are still many open questions about how best optimize their forecasting potential, and research on prediction markets is growing as highlighted by Tziralis and Tatsiopoulos [2012] in their extensive literature review.

In Gillen et al. [2013], a prediction market was rolled out inside of Intel and is a very powerful example of prediction markets being used for making real-world decision. The experiment was very successful and led to better predictions than the official Intel forecast 75% of the time in the case of Collective Intelligence 2015.
short-term predictions involving direct distribution channels. One very powerful feature of prediction markets in this case was the fact that for each forecast, the distribution of beliefs was available instead of just a point-probability forecast, which helped with “asymmetric forecast loss and better control risks”. Cowgill and Zitzewitz [2013] similarly shows that prediction markets leads to "as much as a 25% reduction in mean squared error" over the prediction of official experts in the companies Google, Ford and Koch industries and that errors in prediction decrease over time normally because good traders learn to deal with inefficiencies, or leave the system. Their research additionally highlight the potential conflict between existing organizational practices and incentives, and the information sharing required for prediction markets to work well in a real organization that standalone general prediction market systems are not plagued with.

Servan-Schreiber et al. [2004] investigate whether real monetary incentives - as opposed to online points, play money - lead to better outcome in prediction markets. This is an important question because most questions on SciCast use only online points as incentives, although some questions are eligible for prizes. They find that, interestingly, real money does not lead to better outcomes and their hypothesis is that real money leads to more information discovery whereas online points lead to better information aggregation, two competing forces that they find to lead to the same results in 208 experiments, lead to no difference in outcomes.

Prediction markets perform very well in the long term, as shown by the work of Berg et al. [2008] who show that prediction markets do better than polls in 74% of the time, and that prediction markets are always better for outcomes that are more than 100 days away. This finding is very relevant because most questions are SciCast tend to be more long term questions. On the other hand, Page and Clemen [2013] shows that prediction markets are susceptible to biases in the long term: low-likelihood events are over-predicted and high-likelihood events are under-predicted, and long-term events are mis-calibrated while short-term events tend to be correct, and that these asymmetries can be exploited by betters on the system.

Goel et al. [2010] investigates the exact magnitude of how better prediction markets are compared to polls and statistical models and find that their predictive power is indeed bigger than alternative methods, but not by enough to justify their costs of implementation compared with using simpler forecasting methods.

Jian and Sami [2012] show through experiments that the trading sequence structure has an effect on the speed of information aggregation and accuracy: when the order of trading is per-determined, information aggregation is more accurate but also slower. In the structured case only, when trades are complementary instead of when trades are substitutes, there is more “bluffing and delaying strategies” by agents, who confirm in surveys that they can tell the difference when using the prediction market to trade. The paper also confirms that direct (where traders report probabilities directly) and indirect (where trader’s reveal their predictions by trading securities) Market Scoring Rules (MSR’s)

3. RESEARCH DESIGN

We selected only binary questions (True/False) and unordered multinomial questions (multiple choice questions where the distance to the right answer has no weight), and for each possible outcome of each question we create a datapoint. We also chose to only analyze questions that have already closed using the training set/prediction set framework, as opposed to predicting the future outcome of open questions, so that the outcomes are known and we can thus evaluate their objective accuracy.

We used several versions of outcome measures: 1) The marker of which of the choices of each question was actually chosen by the SciCast curators to be the winning option, called the resolved option 2) The probability of the each choice at the last trade (which is in almost all cases practically the same as the resolved option. We can see in the figure that transition cells cluster 3) The probability of each choice
50% into the lifetime of the question on the platform 4) The probability of each choice at 90% of the trade volume for each question

As features, we selected the following: 1) Whether a question is prize-eligible: SC curates certain questions that are eligible for real prizes (instead of platform points) and we believe that the market will be more efficient due to higher competition because real incentives are present 2) Whether a trade was performed by a high-point user: we believe that such users are more skilled because they might be domain experts, be better at trading, etc. such that their trades might be more predictive of outcomes 3) Whether a trade was performed by a trader who was participated in writing the question because question collaborators might know more about the outcome of a question, so their trades will be more informative 4) We compute the probability change for each choice during each trade and get a history of probability change over time 5) We compute the average of the last five trades for every 5 trades as a smoothed measure of probability 6) We compute how many points were bet on each choice during each trade and get a history of probability change over time 7) We flag if each trade was made using the power mode interface which is an advanced way for users to decide how many points exactly they want to bet during a trade. We believe that trades made using the power mode are better indicators of an outcome because the users have cared more about the bet 8) We count how many trades were made for each question so that we can discard questions that had too few trades 9) We also parse the comment history of each question and count how many comments there were in total for each question, and how many upvotes and downvotes all comments for a question got. The intuition is that questions that got a lot of discussion led to better information and hence prediction.

4. RESULTS & CONCLUSION

Relative to the starting uniform distribution, about half of the uncertainty is removed in the first few trades.

To test if the forecasts contain information that is masked by some form of bias that can be learned and reverted, we deploy machine learning methods to map the features and prices of the market. All classifiers were evaluated using stratified 5-fold cross validation, for those where parameters needed to be tuned, an internal cross validation step was carried out. The classifiers used were AdaBoost, Random Forest, Gradient Boosting, logistic, Ridge, Passive Aggressive perceptron, and SVM. For all methods we use the implementations in Scikit Learn.

None of the combinations of features with any classifier was able to beat the baseline of the price at the time the last feature was determined.

Collective Intelligence 2015.
REFERENCES


