Artificial Prediction Markets Present a Novel Opportunity for Human-AI Collaboration

Extended Abstract

Tatiana Chakravorti
Pennsylvania State University
State College, USA
tfc5416@psu.edu

Vaibhav Singh
Pennsylvania State University
State College, USA
vxs5308@psu.edu

Sarah Rajtmajer
Pennsylvania State University
State College, USA
smr48@psu.edu

Michael McLaughlin
Pennsylvania State University
State College, USA
mvm7085@psu.edu

Robert Fraleigh
Pennsylvania State University
State College, USA
rdf5090@psu.edu

Christopher Griffin
Pennsylvania State University
State College, USA
cxg286@psu.edu

Anthony Kwasnica
Pennsylvania State University
State College, USA
amk17@psu.edu

David Pennock
Rutgers University
New Jersey, USA
david.pennock@rutgers.edu

C. Lee Giles
Pennsylvania State University
State College, USA
clg20@psu.edu

ABSTRACT

Despite high-profile successes in the field of Artificial Intelligence, machine-driven technologies still suffer important limitations, particularly for complex tasks where creativity, planning, common sense, intuition, or learning from limited data is required. These limitations motivate effective methods for human-machine collaboration. Our work makes two primary contributions. We thoroughly experiment with an artificial prediction market model to understand the effects of market parameters on model performance for benchmark classification tasks. We then demonstrate, through simulation, the impact of exogenous agents in the market, where these exogenous agents represent primitive human behaviors.

KEYWORDS

prediction markets; machine learning; human-AI collaboration

ACM Reference Format:
Tatiana Chakravorti, Vaibhav Singh, Sarah Rajtmajer, Michael McLaughlin, Robert Fraleigh, Christopher Griffin, Anthony Kwasnica, David Pennock, and C. Lee Giles. 2023. Artificial Prediction Markets Present a Novel Opportunity for Human-AI Collaboration: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

A body of work on artificial prediction markets is emerging. These are numerically simulated markets, populated by artificial agents for the purpose of supervised learning of probability estimators [4]. While nascent, this literature has demonstrated the plausibility of using a trained market as a supervised learning algorithm, achieving comparable performance to standard approaches on simple classification tasks [3, 4, 12, 14].

Like other machine learning algorithms, functioning of an artificial prediction market depends on several researcher-determined parameters: number of agents; liquidity; initial cash; alongside parameters related to training processes. Scenarios in which performance is robust or brittle to these settings are yet unclear. Prior work has observed that artificial markets may suffer from a lack of participation [16]. That is, like their human counterparts in traditional prediction markets, agents may not invest in the market if they do not have sufficient information [2, 17, 18].

We suggest that a promising opportunity afforded by artificial prediction markets is eventual human-AI collaboration – a market framework should support human traders participating alongside agents to evaluate outcomes. That this approach may be particularly valuable in contexts where machine learning falls short (e.g., lack of training data, complex tasks) and the potential for human-only approaches is either undesirable or infeasible.

Our work is framed by two primary research questions.

RQ1: How does the performance of a simple artificial prediction market depend on hyper-parameter selection?

RQ2: What impact does the inclusion of exogenous agents representing simple, human-like behaviors have on market performance?

2 DATA

We consider three classification tasks. The first two are benchmark ML tasks [8, 13] used broadly to compare the performance of machine learning algorithms. The third is the task of classifying scientific research outcomes as replicable or not replicable – a challenging, complex task on which both machine learning algorithms [1, 15, 19, 20] and human assessment [5–7, 9–11] have achieved respectable but not excellent performance. Specifically, we use the dataset and extracted features considered by [16] for ease of comparison. The dataset contains 192 findings in the social

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

and behavioral sciences, each labeled either Replicable or Not Replicable, and a set of 41 features extracted from each associated paper representing bibliometric, venue-related, author-related, statistical, and semantic information.

3 EXPERIMENTAL DESIGN

RQ1. We use as a base model the artificial binary prediction market described in [14] to study the effects of inter-arrival rate λ, agent initial bank value $B_i(0)$ (or, "cash"), and market liquidity factor $1/\beta$ on artificial market performance. Number of generations is fixed at five during training; while, market duration is fixed at 20. These parameters were fixed (vs. manipulated) to avoid combinatorial complexity during this initial study; however, they should be studied in future work.

RQ2. We introduce three classes of exogenous agents representing primitive behaviors that operate fully separate from the agent logic and feature-based training protocol used for the other agents in the market. The first, *ground truth* agents GT have perfect knowledge of the outcome and always buy contracts corresponding to the correct outcome whenever they have the opportunity to participate (moderated by arrival rate, λ). The second is *ground truth inverse* agents GT_{inv}. These agents always buy contracts corresponding to the incorrect outcome whenever they have an opportunity to participate. The third class is *random* agents *rand* which purchase contracts corresponding to one or the other outcome randomly.

4 RESULTS

RQ1. *Task 1:* Best **F1** of **0.91** is achieved for the first benchmark ML task, Iris image classification [8]. for {liquidity factor = 300, λ = 1.0, initial cash = 1}. In this setting, accuracy is 0.94 and 100% of the data is scored. Generally, better performance is obtained when initial cash ranges between 1 and 4 and when liquidity is greater than 100. Choice of λ does not appear to significantly impact performance.

Task 2: Performance is generally poorer on the benchmark heart disease classification task [13] than for the Iris image classification task, and there is also less clear region of best performance in hyper-parameter space. Highest **F1** of **0.71** is achieved for {liquidity factor = 50, λ = 0.05, initial cash = 20}. In this setting, accuracy is 0.66 and 99.67% of the data is scored.

Task 3: In the context of replication outcomes prediction, best **F1** of **0.84** is achieved for {liquidity factor = 5, λ = 0.05, initial cash = 1}. Accuracy is 0.79 and 36% of the test data is scored. The market algorithm struggles with agent participation on this task; all but two hyper-parameter combinations leave at least 40% of the test data unscored. Performance increases with liquidity and decreases with initial cash, while the effect of λ reveals no clear pattern.

RQ2. *Task 1:* We introduce GT, GT_{inv} and *rand* agents into the market. These agents operate outside of the training process and, as such, represent primitives that may underlie simple human participant inputs. Exogenous agents are introduced into the general agent pool and are subject to the same arrival rate, λ, as trained agents. We find that the inclusion of even a very small population of GT agents improves market performance substantially. The impact of random agents is relatively lesser (Table 1).

Table 1: Average F1 on 10 best and worst-scoring replication markets, for different exogenous agent populations.

Baseline	GT 0.1%	GT 1%	GT_{inv} 0.1%	GT_{inv} 1%	*rand* 1%	*rand* 10%
0.84	0.93	1	0.34	0.09	0.79	0.80
0.84	0.91	0.97	0.34	0.28	0.74	0.75
0.83	0.94	1	0.32	0.06	0.79	0.83
0.83	0.90	0.99	0.33	0.24	0.76	0.82
0.83	0.88	0.94	0.33	0.29	0.75	0.79
0.69	0.89	0.96	0.34	0.28	0.76	0.81
0.69	0.91	0.96	0.34	0.29	0.78	0.77
0.68	0.95	0.96	0.33	0.29	0.78	0.79
0.66	0.89	0.94	0.33	0.29	0.76	0.79
0.65	0.90	0.94	0.34	0.30	0.77	0.81

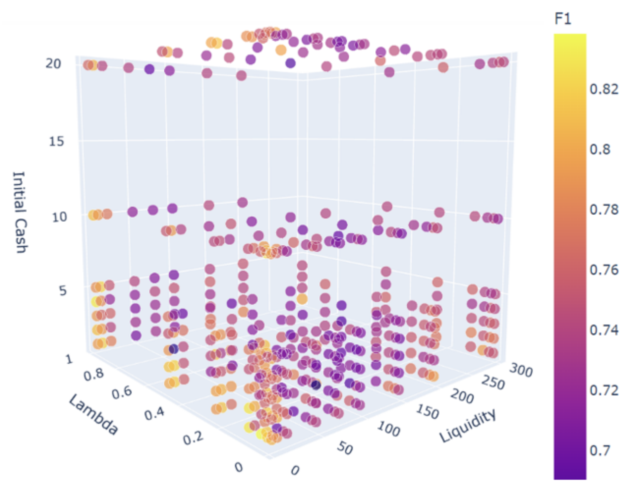

Figure 1: Average F1 score on the replication prediction task, plotted in hyper-parameter space.

5 CONCLUSIONS

The comprehensive study of a simple artificial prediction market we undertake here highlights a promising new machine learning algorithm, which achieves respectable performance on benchmark machine learning tasks but which, we argue, affords unique opportunities for human-AI collaboration.

ACKNOWLEDGMENTS

This work was sponsored by the Defense Advanced Research Projects Agency through Cooperative Agreement W911NF-19-2-0272.

REFERENCES

[1] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. 2019. Predicting the replicability of social science lab experiments. *PloS one* 14, 12 (2019), e0225826.

[2] Kenneth J. Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E.

Tetlock, Hal R. Varian, Justin Wolfers, and Eric Zitzewitz. 2008. The Promise of Prediction Markets. *Science* 320, 5878 (May 2008), 877–878.

[3] Adrian Barbu and Nathan Lay. 2013. Artificial prediction markets for lymph node detection. In *2013 E-Health and Bioengineering Conference (EHB)*. IEEE, 1–7.

[4] Adrian Barbu, Nathan Lay, and Shie Mannor. 2012. An Introduction to Artificial Prediction Markets for Classification. *Journal of Machine Learning Research* 13, 7 (2012).

[5] Colin F Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 6280 (2016), 1433–1436.

[6] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* 2, 9 (2018), 637–644. https://doi.org/10.1038/s41562-018-0399-z

[7] Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A Nosek, and Magnus Johannesson. 2015. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* 112, 50 (2015), 15343–15347.

[8] R.A. Fisher. 1988. Iris. UCI Machine Learning Repository.

[9] Eskil Forsell, Domenico Viganola, Thomas Pfeiffer, Johan Almenberg, Brad Wilson, Yiling Chen, Brian A Nosek, Magnus Johannesson, and Anna Dreber. 2019. Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology* 75 (2019), 102117.

[10] Michael Gordon, Domenico Viganola, Michael Bishop, Yiling Chen, Anna Dreber, Brandon Goldfedder, Felix Holzmeister, Magnus Johannesson, Yang Liu, Charles Twardy, et al. 2020. Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society open science* (2020).

[11] Michael Gordon, Domenico Viganola, Anna Dreber, Magnus Johannesson, and Thomas Pfeiffer. 2021. Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *Plos one* 16, 4 (2021), e0248780.

[12] Fatemeh Jahedpari, Julian Padget, Marina De Vos, and Benjamin Hirsch. 2014. Artificial prediction markets as a tool for syndromic surveillance. *Crowd Intelligence: Foundations, Methods and Practices* (2014).

[13] Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. 1988. Heart disease data set. *The UCI KDD Archive* (1988).

[14] Nishanth Nakshatri, Arjun Menon, C Lee Giles, Sarah Rajtmajer, and Christopher Griffin. 2021. Design and Analysis of a Synthetic Prediction Market using Dynamic Convex Sets. *arXiv preprint arXiv:2101.01787* (2021).

[15] Samuel Pawel and Leonhard Held. 2020. Probabilistic forecasting of replication studies. *PloS one* 15, 4 (2020), e0231416.

[16] Sarah Rajtmajer, Christopher Griffin, Jian Wu, Robert Fraleigh, Laxmaan Balaji, Anna Squicciarini, Anthony Kwasnica, David Pennock, Michael McLaughlin, Timothy Fritton, et al. 2022. A synthetic prediction market for estimating confidence in published work. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 13218–13220.

[17] David Rothschild and David M Pennock. 2014. The extent of price misalignment in prediction markets. *Algorithmic Finance* 3, 1-2 (2014), 3–20.

[18] Paul C Tetlock. 2008. Liquidity and prediction market efficiency. *Available at SSRN 929916* (2008).

[19] Jian Wu, Rajal Nivargi, Sree Sai Teja Lanka, Arjun Manoj Menon, Sai Ajay Modukuri, Nishanth Nakshatri, Xin Wei, Zhuoer Wang, James Caverlee, Sarah M Rajtmajer, et al. 2021. Predicting the Reproducibility of Social and Behavioral Science Papers Using Supervised Learning Models. *arXiv preprint arXiv:2104.04580* (2021).

[20] Yang Yang, Wu Youyou, and Brian Uzzi. 2020. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* 117, 20 (2020), 10762–10768.