https://pubsonline.informs.org/journal/ijds

INFORMS JOURNAL ON DATA SCIENCE

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000

Submitted to *INFORMS Journal on Data Science*

ISSN 2694-4022, EISSN 2694-4030

Towards Replication-Robust Analytics Markets

(Authors' names are not included for peer review)

Abstract. Despite widespread adoption of machine learning, many firms face the common challenge of relevant datasets being distributed amongst market competitors whom are reluctant to share information. Accordingly, recent works propose analytics markets as a way to provide monetary incentives for collaboration, where agents share features and are rewarded based on their contribution to improving the predictions of others. These contributions are determined by their relative Shapley value, computed by treating features as players and their interactions as a cooperative game. However, this setup is known to incite agents to strategically replicate their data and act under multiple false identities to increase their own revenue whilst diminishing that of others, which limits the viability of these markets in practice. In this work, we develop an analytics market robust to such strategic replication for supervised learning problems. We adopt Pearl's do-calculus from causal inference to refine the cooperative game by differentiating between observational and interventional conditional probabilities. As a result, we derive Shapley value-based rewards that deter replication by design.

Key words: Regression, Collaborative Analytics, Market Design, Causality

1. Introduction

Machine learning relies heavily on both the quality and quantity of input data, however, firms often find it difficult, if not impossible, to acquire rich datasets themselves. This is often due to privacy constraints. For instance, in the medical domain, data is highly sensitive and subject to strict regulations (Rieke et al. 2020), yet hospitals could benefit from sharing patient information to mitigate social biases in diagnostic support systems. Similar examples include rival distributors sharing sales data to improve supply forecasts, or hotel operators using airline data to better anticipate demand. One promising solution to this is federated learning, where multiple agents collaborative by pooling their local resources for a central learning task without directly exchanging raw data (Yang et al. 2019). However, this approach presupposes that agents are willing to share information altruistically—an assumption that may not hold if these agents also compete in downstream markets (Gal-Or 1985). To incentivize data sharing, one can instead frame data as a commodity within a market-based framework (Bergemann and Bonatti 2019). This is by no means a new concept, as

many platforms already exist to purchase of raw datasets directly from their owner via bilateral transactions (Rasouli and Jordan 2021). That said, pricing these datasets is not easy as their value ultimately depends on when, how, and by whom they are eventually used (Mussell 2014). Further, since datasets often contain overlapping information, their value is inhernetly combinatorial, so these seemingly straightforward transactions can easily become intractable to price.

Data can also be monetized via *prediction markets*—exchanges where participants buy and sell securities whose payoffs depend on the outcome of future events (Wolfers and Zitzewitz 2004, Storkey 2011). The prices of these securities reflect the collective belief about the probability of the event occurring, effectively crowdsourcing diverse information from the market (Abernethy et al. 2015). These markets were some of the first to question the notion that raw data possesses a singular intrinsic value, arguing instead that its usefulness depends on the specific task at hand. However, a key limitation is that sellers must choose which tasks to predict without prior knowledge of how relevant their data might be. To this end, recent works instead advocate for *analytics markets*—real-time mechanisms that match datasets to machine learning tasks based on predictive performance, which build on federated learning by retaining the possibility to distribute compute and preserve privacy (Agarwal et al. 2019). In these markets, revenue is generated based on the value task owner's value for accuracy. Such markets have been proposed for both classification (Koutsopoulos et al. 2015) and regression (Pinson et al. 2022, Falconer et al. 2024) tasks.

In analytics markets, each feature is allocated a portion of the revenue by treating each as a player in a cooperative game and using well-established solution concepts from game theory—specifically, semivalues (Dubey et al. 1981)—a framework also used in machine learning for feature importance (Ghorbani and Zou 2019). The key advantage of using semivalues is that they are characterized by a set of axioms, namely symmetry, efficiency, null-player, and additivity, that lead to desirable market properties by design (for precise definitions, see Chalkiadakis et al. 2011). A feature's semivalue represents its expected marginal contribution to predictive performance across all subsets of other features. In many applications, the Shapley value (Shapley 1997) is favored as it is the unique semivalue that satisfies all four axioms.

1.1. Challenges

For an arbitrary feature vector $\boldsymbol{x} \in \mathbb{R}^n$, a revenue allocation function should ideally take the form $\phi : \mathcal{H} \times \mathbb{R}^n \mapsto \mathbb{R}^n$, where $\mathcal{H}$ is the set of all possible scoring rules $h : \mathbb{R}^n \mapsto \mathbb{R}$. In other words, the output of $h(\boldsymbol{x})$ is directly decomposed into contributions $(\phi(h,\boldsymbol{x})_1, \phi(h,\boldsymbol{x})_2, \ldots, \phi(h,\boldsymbol{x})_n)$ for each feature, such that h need only be evaluated for the complete input vector $\boldsymbol{x}$. However, to compute

the Shapley values, the scoring rule needs to be evaluated for each subset of features. The problem is, standard machine learning models are typically defined only for complete input vectors (to avoid issues such as matrix dimension mismatches), so they do not naturally produce outputs for partial inputs. To address this, one must define a so-called *lift* function $\zeta : \mathcal{H} \times \mathbb{R}^n \times 2^n \mapsto \mathbb{R}$, also referred to as a characteristic function, which extends h to operate on subsets $C \subseteq \{1, \ldots, n\}$ of features (Merrill et al. 2019). That is, the lift $\zeta(h, \boldsymbol{x}, C)$ assigns a value for each C, thereby *lifting* h from $\mathbb{R}^n$ to the $\mathbb{R}^n \times 2^n$.

The computed Shapley values are therefore contingent upon the particular lift used to map the dense and uncountable input space into the required discrete domain. As there are several ways to formulate such a lift, it is not immediately clear which one to use (Sundararajan and Najmi 2020). If we refer to each subset C of features as a *coalition*, these lifts simulate the inclusion or exclusion of features, differing in how they model the distributions of *in-coalition* features conditioned on *out-of-coalition* features, most of which can be categorized as either *observational* or *interventional*. From the lens of causal inference, an observational conditional probability describes the relationship between two or more variables as they occur naturally, whereas an interventional conditional probability is the result when one "intervenes" by fixing a particular variable's value (Pearl 2010).

In existing works on analytics markets (e.g., Agarwal et al. 2019, Pinson et al. 2022) the choice of lift uses observational conditional probabilities. These works highlight that if an agent's feature is highly correlated with that of another agent, they are able to strategically submit many replicates of their feature under different identities to increase their revenue and dimish that of others. This can be done freely since, unlike material commodities, data can be replicated at no additional cost. Whilst many attempts have been made to remedy this problem, doing so typically requires a trade-off. For instance, Ohrimenko et al. (2019) propose a more elaborate mechanism design, requiring each seller to also have their own machine learning task for which they want to procure data, which has practical limitations. In Agarwal et al. (2019), a modification to the Shapley value is proposed which penalizes similar features, thereby deterring replication. However, budget balance is sacrificed to achieve this, meaning that that some of the market revenue, that perhaps should have been portioned to other agents, remains unallocated. Their setup is also vulnerable to spiteful agents—those who seek to minimize the revenue of others as well as maximize their own. A similar shortcoming is observed in the proposal of Han et al. (2023), as both natural correlations and deliberate replications are penalized. In this work, we show that the choice of lift is responsible for these grossly undesirable allocations.

1.2. Contributions

The key contributions of our work are as follows: (i) we propose a general analytics market design for supervised learning problems that subsumes recent proposals in literature; (ii) we show that there are many ways in which Shapley values can be used to allocate revenue and that the differences between them can be explained from a causal perspective; (iii) by applying its recent links to feature importance, we show that the replication incentives in existing works can be explained using Pearl (2012)'s seminal work on causality; (iv) by replacing the conventional approach of conditioning by *observation* with conditioning by *intervention*, we design a market that is robust to replication whilst also accounting for spiteful agents, thereby taking a step toward the practical application of these markets; and finally (v) we demonstrate our findings on a real-world case study—out of many potential applications, we choose to study wind power forecasting due to data availability, the known value of sharing distributed data, and the fact it is a sandbox that can be easily shared and used by others.

The remainder of this paper is structured as follows: Section 2 presents our general market design framework. In Section 3 we derive variants of the characteristic function and analyze each from a causal perspective. In Section 4 we discuss the impact of each on the robustness of the market to replication. Section 5 then illustrates our findings on a real-world case study. Finally, Section 6 gathers a set of conclusions and perspectives for future work.

2. Preliminaries

As many machine learning applications involve forecasting, we focus on regression analysis in the context of analytics markets, yet our setup can be used for any supervised learning problem. This builds upon prior work on data acquisition for machine learning tasks from both strategic (Dekel et al. 2010) and privacy-conscious (Cummings et al. 2015) agents. We assume an owner of a regression task has a valuation for a marginal improvement in predictive performance, which sets the price for the distributed agents, whom in turn propose their own data as features and are rewarded based on their marginal contributions to this improvement. We denote this valuation $\lambda \in \mathbb{R}_{\geq 0}$, the value of which we assume to be known and reported truthfully. We refer the reader to Ravindranath et al. (2024) for a recent proposal of how λ may be elicited in practice.

2.1. Market Agents

Let $\mathcal{A}$ be the set of market agents, one of which $c \in \mathcal{A}$ is a *central agent* seeking to improve their predictions, whilst the remaining agents $a \in \mathcal{A}_{-c}$ are *support agents*, whom propose their own data

as features, whereby $\mathcal{A}_{-c} = \mathcal{A} \setminus \{c\}$. Let $\{y^{(t)}\}$ be the target signal recorded by the central agent, with $y^{(t)} \in \mathbb{R}$ a sample from the stochastic process $\{Y^{(t)}\}$ at time t. Assuming there are M available features in the market, we let $\boldsymbol{x}^{(t)} = [x_1^{(t)}, \ldots, x_M^{(t)}]^\top$ be the vector of values at time t, indexed by the ordered set $\mathcal{I} = \{1, \ldots M\}$. If only a particular subset of features $C \subseteq \mathcal{I}$ is used, we add an index for the set itself, such that the vector of values for features in C at time t is denoted by $\boldsymbol{x}_C^{(t)}$. Each agent $a \in \mathcal{A}$ owns a subset $\mathcal{I}_a \subseteq \mathcal{I}$ of indices. We write $\mathcal{I}_{-c}$ as the set of indices for features owned only by the support agents. For each subset of features $C \subseteq \mathcal{I}$ we write $\mathcal{D}_C^{(t)} = \{\boldsymbol{x}_C^{(t)}, y^{(t)}\}$ to be the input-output pair observed at time t.

2.2. Regression Framework

To model the target signal, $Y^{(t)}$, we use a parametric Bayesian regression framework, formulating the likelihood as a deviation from a deterministic mapping under an independent Gaussian noise process, the variance of which is treated as a hyperparameter. We use a linear interpolant parameterized by a vector of coefficients which represents the conditional expectation of the target signal, such that the interpolant using all available features at time t can be decomposed as:

$$f(\boldsymbol{x}^{(t)}, \boldsymbol{w}) = \underbrace{w_0 + \sum_{i \in \mathcal{I}_c} w_i x_i^{(t)}}_{\substack{\text{Terms belonging} \\ \text{to the central agent.}}} + \underbrace{\sum_{a \in \mathcal{A}_{-c}} \sum_{j \in \mathcal{I}_a} w_j x_j^{(t)}}_{\substack{\text{Terms belonging} \\ \text{to the support agents.}}}.$$

REMARK 1. *We focus on parametric regression with functions that are linear in their coefficients to guarantee certain market properties. One can of course obtain a rich class of models with linear combinations of nonlinear basis functions or splines, however we adopt only a linear basis in this work for ease of exposition. For an application of nonlinear basis functions to analytics markets, the reader is referred to Falconer et al. (2024).*

2.3. Market Clearing

As in Pinson et al. (2022), we adopt a two-stage (i.e., in-sample and out-of-sample) market. In the first stage, parameters are inferred using observed input–output pairs. In the second stage, the trained model is deployed to forecast on previously unseen data, thereby testing its ability to generalize beyond the training set. Both stages require performance evaluation, as well as processes for payment collection and revenue allocation. We relax their assumption that features are independent, yet still remove redundant features owned by the support agents (i.e., those that are highly correlated with the central agent's features) via the detailed feature selection process.

Parameter Inference We opt for a centered isotropic Gaussian prior, which is conjugate for our likelihood, resulting in a tractable Gaussian posterior that summarizes the updated beliefs, which, for a particular subset of features at time t is given by

$$\begin{aligned} & p(\boldsymbol{w}_C|\mathcal{D}_C^{(t)}) \\ & \quad \propto p(\mathcal{D}_C^{(t)}|\boldsymbol{w}_C)p(\boldsymbol{w}_C|\mathcal{D}_C^{(t-1)}), \\ & \quad = p(\mathcal{D}_C^{(t)}|\boldsymbol{w}_C)\left(p(\boldsymbol{w}_C)\prod_{t'<t}p(\mathcal{D}_C^{(t')}|\boldsymbol{w}_C)\right), \end{aligned}$$

where recall $\mathcal{D}_C^{(t)}$ is the input-output pair observed at time t. We note the use of Gaussians is only for mathematical convenience, and our framework can be readily extended to more general hypotheses (e.g., without conjugate priors). Our Bayesian approach also subsumes many frequentist methods, making it easy to apply, for instance, ordinary least-squares or maximum likelihood estimation.

Performance Evaluation At time $t+1$, the predictive density is equal to the convolution of the likelihood with the posterior at time t such that

$$\hat{y}_C^{(t+1)} = \int_\Theta p(y^{(t+1)}|\boldsymbol{x}_C^{(t+1)};\boldsymbol{w})p(\boldsymbol{w}_C|\mathcal{D}_C^{(t)})d\boldsymbol{w},$$

where $\hat{y}^{(t+1)} = p(y^{(t+1)}|\boldsymbol{x}_C^{(t+1)})$ is the prediction for the features in C. We measure performance using the negative log likelihood, $h(\boldsymbol{x}_C^{(t+1)}) = -\log \hat{y}_C^{(t+1)}$, which can be described as a negatively oriented strictly proper scoring rule. Ergo, the following properties hold: (i) between any two models, the one with a more accurate description of the data produces a lower score; and (ii) the score is uniquely minimized when the predicted distribution matches the true distribution. We retain a recursive estimate of its expected value as observations arrive for each subset of features, $\mathbb{E}[h(\boldsymbol{x}_C)]^{(t)}$.

Payment Collection Market revenue is a function of the exogenous valuation, $\lambda \geq 0$, and the extent to which model-fitting is improved. This is measured using the current estimate of the expected value of the scoring rule, such that the market revenue at time t is given by

$$\pi^{(t)} = \lambda\left(\mathbb{E}[h(\boldsymbol{x}_{\mathcal{I}_c})]^{(t)} - \mathbb{E}[h(\boldsymbol{x}_{\mathcal{I}})]^{(t)}\right)$$

which is the payment collected from the central agent.

Revenue Allocation We use the Shapley value to reward each feature for their contribution to the improved predictive performance.

DEFINITION 1 (**CHARACTERISTIC FUNCTION**). For a given scoring rule h and feature vector $\boldsymbol{x}^{(t)}$, a characteristic function $\zeta : \mathbb{R}^{|\mathcal{I}|} \times \mathcal{P}(\mathcal{I}_{-c}) \mapsto \mathbb{R}$ assigns a real number $\zeta(\boldsymbol{x}^{(t)}, C)$ to each subset $C \subseteq \mathcal{I}_{-c}$.

For brevity, we write $\zeta_C^{(t)} = \zeta(\boldsymbol{x}^{(t)}, C)$. Each subset $C \subseteq \mathcal{I}_{-c}$ is a coalition in the cooperative game, with $\mathcal{I}_{-c}$ the so-called grand coalition. Let $m = |\mathcal{I}_{-c}|$ be the number of support agents, such that the Shapely value for feature i at time t is

$$\phi_i^{(t)} = \frac{1}{m} \sum_{C \in \mathcal{P}(\mathcal{I}_{-c} \setminus \{i\})} \binom{m-1}{|C|}^{-1} \delta_i^{(t)}(C), \tag{1}$$

where $\delta_i^{(t)}(C) = \zeta^{(t)}_{\mathcal{I}_c \cup C} - \zeta^{(t)}_{\mathcal{I}_c \cup C \cup i}$ is the marginal contribution of feature i to coalition C.

We acknowledge that evaluating $\phi_i^{(t)}$ is NP-hard in general (Deng and Papadimitriou 1994), with a time complexity of $O(2^m)$, hence in practice one must rely on approximation methods (Castro et al. 2009, Mitchell et al. 2022). An obvious method is to obtain a Monte-Carlo estimate by sampling $d < m$ terms from the sum in (1) with probability $p(C) = 1/\binom{m-1}{|C|}$ such that an approximate Shapley value is given by

$$\hat{\phi}_i^{(t)} = \frac{1}{d} \sum_{j=1}^{d} \delta_j^{(t)}(C_j),$$

which is an unbiased estimator that converges asymptotically at a rate of $O(1/\sqrt{d})$, according to the Central Limit Theorem. However, in this work we are solely focused on the functional form of ζ, which is agnostic to the choice of sampling method, so exploring state-of-the-art approximations is outwith the scope of this work so we revert to (1) to compute the Shapley values.

The reward for each support agent can then be written as

$$\pi_a = \sum_{i \in \mathcal{I}_a} \lambda \mathbb{E}[\phi_i]^{(t)}, \quad \forall a \in \mathcal{A}_{-c}.$$

DEFINITION 2 (**MARKET PROPERTIES**). With the proposed regression framework and Shapley value-based revenue allocation, regression markets have the following properties:

1. *Symmetry*—Any two features with the same marginal contribution to all coalitions obtain equal reward, that is, $\forall C \in \mathcal{I}_{-c} \setminus \{i, j\} : \zeta^{(t)}_{\mathcal{I}_c \cup C \cup i} \equiv \zeta^{(t)}_{\mathcal{I}_c \cup C \cup j} \mapsto \phi_i^{(t)} \equiv \phi_j^{(t)}, \forall (i, j) \in \mathcal{I}_{-c}, i \neq j, \forall t$.

2. *Linearity*—For any two features, their joint contribution to coalition is equal to the sum of their marginal contributions, that is, $\zeta^{(t)}_{\mathcal{I}_c \cup C \cup i} + \zeta^{(t)}_{\mathcal{I}_c \cup C \cup j} = \zeta^{(t)}_{\mathcal{I}_c \cup C \cup i,j}, \forall (i, j) \in \mathcal{I}_{-c}, \forall t$.

3. *Budget balance*—The payment of the central agent is equal to the sum of rewards received by all the support agents, that is, $\pi = \sum_{a \in \mathcal{A}_{-c}} \pi_a$.

4. *Individual rationality*—Support agents have a weak preference to participate in the market rather than the outside option, that is, $\pi_a \geq 0, \forall a \in \mathcal{A}_{-c}$.

5. *Zero-element*—If a support agent provides no feature, or provide features with zero marginal contribution to all coalitions, they earn no reward, that is, $\forall C \in \mathcal{I}_{-c} : \zeta^{(t)}_{\mathcal{I}_c \cup C \cup i} \equiv \zeta^{(t)}_{\mathcal{I}_c \cup C}, \forall i \in \mathcal{I}_a \mapsto \pi_a = 0$.

6. *Truthfulness*—Support agents maximize their reward by reporting their true data.

These desirable market properties stem from the axioms of the Shapley value, a detailed proof of which is provided in Falconer et al. (2024). Recall that our scoring rule h relates to the linear interpolant $f : \mathbb{R}^{|\mathcal{I}|} \mapsto \mathbb{R}$ and is therefore itself only defined on $\mathbb{R}^{|\mathcal{I}|}$. To compute $\phi^{(t)}_i$, an evaluation of h for each coalition $C \in \mathcal{P}(\mathcal{I}_{-c})$ of features is needed, where $|\mathcal{P}(\mathcal{I}_{-c})| = 2^m$. Accordingly, we lift the scoring rule to a higher dimensional space with the characteristic function. For machine learning problems, one could argue that the Shapley value is not well-defined in general, as there exists many methods to formulate this lift (Sundararajan and Najmi 2020). In the following section, we explore these methods and their differences from a causal perspective.

3. Characteristic Function

Methods to compute Shapley values for machine learning problems can broadly be categorized as either *observational* or *interventional*, relating to the formulation of the characteristic function that underpins the cooperative game. The former is typically found in work related to analytics markets (e.g, Agarwal et al. 2019, Pinson et al. 2022) and the latter used for interoperability in machine learning (Lundberg and Lee 2017). Recall that the purpose of the lift is to simulate the removal of features to obtain partial evaluations of h. These two formulations differ in how they model the distribution of features, in particular, the distribution of features within a coalition C conditioned on those not in C.

The observational lift uses the *observational conditional expectation*, the expectation of the scoring rule at time t, where the integral is taken with respect to the out-of-coalition features given the in-coalition features take on their observed values, such that

$$\zeta^{(t),\text{obs}}_C = \int h(\boldsymbol{x}^{(t)}_C, \boldsymbol{x}^{(t)}_{C'}) p(\boldsymbol{x}^{(t)}_{C'} | \boldsymbol{x}^{(t)}_C) d\boldsymbol{x}^{(t)}_{C'}, \tag{2}$$

where $C' = \mathcal{I} \setminus C$ denotes the out-of-coalition features.

The interventional lift uses the *interventional conditional expectation*, which is given by

$$\zeta_C^{(t),\text{int}} = \int h(\boldsymbol{x}_C^{(t)}, \boldsymbol{x}_{C'}^{(t)}) p(\boldsymbol{x}_{C'}^{(t)} | \operatorname{do}(\boldsymbol{x}_C^{(t)})) d\boldsymbol{x}_{C'}^{(t)}, \tag{3}$$

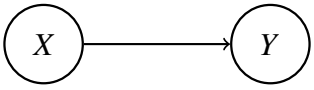
where $\operatorname{do}(\cdot)$ is an operator from Pearl's *do*-calculus (Pearl 2012) that represents an "intervention" where the data generating process is manipulated by manually fixing the features in the coalition to their observed values. The key difference between (2) and (3) is that in the former, conditioning on the observed values of the features in the coalition can alter the distribution of the out-of-coalition features if any latent dependencies exist, an effect which is ignored in the latter by fixing the in-coalition features via the *do*-intervention. For further intuition we provide the following example.

Illustrative Example Consider the causal graph in Figure 1. In this setup, two variables, X and Y, are connected by a single directed edge. If we observe some value $X = x$, the observational conditional distribution of Y describes: *the distribution of Y given that X is observed to take on the value x*, written as $p(y|x) = p(x, y)/p(x)$. By contrast, the interventional conditional distribution $p(y|\operatorname{do}(x))$ describes instead: *the distribution of Y given that we artificially set the value of X to x*, denoted $p(y|\operatorname{do}(x))$. Graphically, an intervention removes all edges going into the variable. As there are no parents of X, intervening on X does not alter any other part of the system, so $p(y|\operatorname{do}(x)) = p(y|x)$. Yet, if we intervene on Y by setting $Y = y$, we remove all edges into Y, so $p(x|\operatorname{do}(y)) = p(x)$, hence x simply governed by its marginal distribution.

Computation These two lifts also differ significantly in their computational expense (Lundberg and Lee 2017). In particular, computing the observational conditional expectation of h is generally intractable, requiring complex and expensive approximations (Covert et al. 2021). By contrast, intervening on features can be done via comparatively simple and efficient methods (Sundararajan and Najmi 2020). Although there is ongoing debate regarding the most suitable way to evaluate the conditional expectation (Chen et al. 2022), one common approach is to train a separate model for each subset of features; if each model is optimal with respect to the scoring rule, then marginalizing out features via their conditional distribution is effectively achieved.

In the context of linear regression over τ time steps, fitting a model for a coalition and evaluating h incur complexities of $O(\tau|C|^2 + |C|^3)$ and $O(\tau|C|)$, respectively, which are calculated for all 2^m coalitions, scaling poorly to high dimensions. In contrast, the interventional lift can be computed

Figure 1 Causal graph indicating a direct effect between two random variables, X and Y.



much faster by simply imputing out-of-coalition features, requiring only a single model (i.e., the grand coalition) with just the scoring rule evaluated for each coalition which is computed in linear time. Note that, both lifts preserve the axioms of the original Shapley value, and subsequently the desirable market properties.

Causal Perspectives When features are mutually independent, the two lifts coincide. To see this, we can think of $\text{do}(\boldsymbol{x}_C^{(t)})$ in (3) as breaking the dependence to $\boldsymbol{x}_C^{(t)}$, without affecting the distribution of $\boldsymbol{x}_C^{(t)}$, thus we can re-write this operation as $p(\boldsymbol{x}_C^{(t)} | \,\text{do}(\boldsymbol{x}_C^{(t)})) = p(\boldsymbol{x}_C^{(t)})$ so the interventional expectation coincides with the marginal expectation (Janzing et al. 2020). If features are independent, we can then calculate (3) from (2) by simply replacing $p(\boldsymbol{x}_C^{(t)}|\boldsymbol{x}_C^{(t)})$ with the marginal distribution, which would be equivalent in this case. With this in mind, we use the following theorem to further analyze these lifts from a causal perspective.

THEOREM 1. *Marginal contributions derived using the observational conditional expectation as defined in (2) can be decomposed into both indirect and direct causal effects.*

Proof First, if we let Θ be the set of all possible permutation of indices in $\mathcal{I}_{-c}$, we can reformulate the Shapley value in (1) for feature i at time t as follows:

$$\phi_i^{(t)} = \frac{1}{m!} \sum_{\theta \in \Theta} \delta_i^{(t)}(\theta),$$

where now $\delta_i^{(t)}(\theta) = \zeta^{(t)}_{\mathcal{I}_c \cup \{j : j \prec_\theta i\}} - \zeta^{(t)}_{\mathcal{I}_c \cup \{j : j \preceq_\theta i\}}$, with $j \prec_\theta i$ meaning j precedes i in permutation θ. Then, using the formulation in (2), the marginal contribution of feature i for a single permutation

$\theta \in \Theta$ derived using the observational lift can be written as

$$\begin{aligned}
\delta^{(t),\text{obs}}(\theta) &= \zeta^{(t),\text{obs}}_{\underline{C}} - \zeta^{(t),\text{obs}}_{\underline{C}\cup i}, \\
&= \int h(\boldsymbol{x}^{(t)}_{\underline{C}}, \boldsymbol{x}^{(t)}_{\bar{C}\cup i}) p(\boldsymbol{x}^{(t)}_{\bar{C}\cup i} | \boldsymbol{x}^{(t)}_{\underline{C}}) d\boldsymbol{x}^{(t)}_{\bar{C}\cup i} \\
&\quad \underbrace{- \int h(\boldsymbol{x}^{(t)}_{\underline{C}\cup i}, \boldsymbol{x}^{(t)}_{\bar{C}}) p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}\cup i}) d\boldsymbol{x}^{(t)}_{\bar{C}}}_{\text{Total effect}}, \\
&= \int h(\boldsymbol{x}^{(t)}_{\underline{C}}, \boldsymbol{x}^{(t)}_{\bar{C}\cup i}) p(\boldsymbol{x}^{(t)}_{\bar{C}\cup i} | \boldsymbol{x}^{(t)}_{\underline{C}}) d\boldsymbol{x}^{(t)}_{\bar{C}\cup i} \\
&\quad \underbrace{- \int h(\boldsymbol{x}^{(t)}_{\underline{C}\cup i}, \boldsymbol{x}^{(t)}_{\bar{C}}) p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}}) d\boldsymbol{x}^{(t)}_{\bar{C}}}_{\text{Direct effect}} \\
&\quad + \int h(\boldsymbol{x}^{(t)}_{\underline{C}\cup i}, \boldsymbol{x}^{(t)}_{\bar{C}}) p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}}) d\boldsymbol{x}^{(t)}_{\bar{C}} \\
&\quad \underbrace{- \int h(\boldsymbol{x}^{(t)}_{\underline{C}\cup i}, \boldsymbol{x}^{(t)}_{\bar{C}}) p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}\cup i}) d\boldsymbol{x}^{(t)}_{\bar{C}}}_{\text{Indirect effect}},
\end{aligned}$$

where $\underline{C} = \{j : j \prec_\theta i\}$ and $\bar{C} = \{j : j \succ_\theta i\}$. Thus, the marginal contribution captures two distinct effects. The first is the direct effect on the expected score when feature i is observed and added to the coalition, keeping the distribution of the out-of-coalition features unchanged, in other words, using $p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}})$ instead of $p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}\cup i})$. The other is the indirect effect on the expected score when the distribution of the out-of-coalition features does change as a result of observing feature i, that is, when $p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}})$ changes to $p(\boldsymbol{x}^{(t)}_{\bar{C}} | \boldsymbol{x}^{(t)}_{\underline{C}\cup i})$. □

Following Theorem 1, we can see that by replacing conditioning by observation with the marginal distribution as in (2), the indirect effect disappears entirely. Hence, the interventional lift disregards causal effects *between features*, and subsequently any latent confounders or root causes with *indirect effects* (Heskes et al. 2020). As a result, the interventional lift is more effective at crediting features upon which the regression model has an explicit algebraic dependence. In contrast, the observational lift attributes features in proportion to indirect effects (Frye et al. 2020), which some argue is illogical as features not explicitly used by the model can receive non-zero allocation.

Whilst this dispute has been used to reject the general use of Shapley values for interoperability in machine learning (Kumar et al. 2020) and argue that Lundberg and Lee (2017) were mistaken to simply convey (3) as a cheap approximation of (2), the choice between observational and

interventional lifts can be viewed as whether one intends to be *true to the data* or *true to the model*, respectively, meaning the trade-offs of each approach can be seen as context-specific (Chen et al. 2020). We argue that the former is best suited for analytics markets.

Interpreting Rewards We can explore this last conjecture by considering how the rewards of the support agents may differ depending on the choice of lift. We know that the predictive performance of the regression model out-of-sample is contingent upon the availability of features that were used during training, which, in practice, requires data of the support agents to be streamed continuously in a timely fashion, particularly for an online setup. If a feature was missing, the efficacy of the forecast may drop, the extent to which would relate not to any root causes or indirect effects regarding the data generating process, but rather the magnitude of direct effects.

Specifically, larger rewards would be made to support agents with features to which the predictive performance of the model is most sensitive, providing incentives to reduce data being unavailable, somewhat resembling reserve payments in energy markets, where assets are remunerated for being available in times of need. With the observational lift, it would instead be unclear as to whether comparatively larger rewards in the regression market are consequential of features having a sizeable impact on predictive performance, or merely a result of indirect effects through those that do. The interventional lift therefore better aligns with desirable intentions of the market.

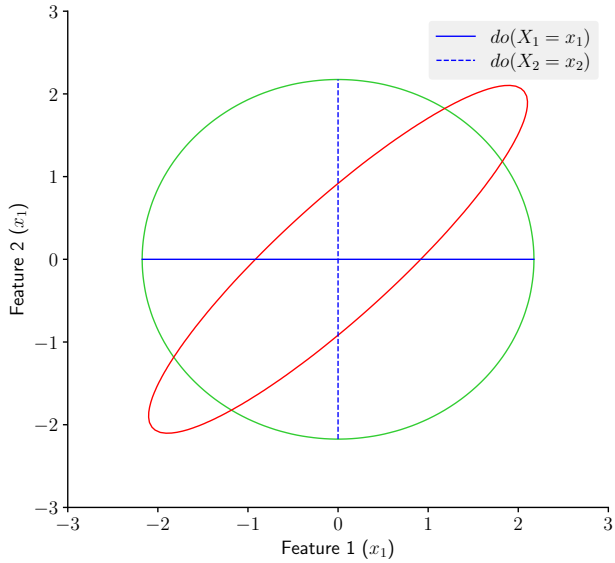
Limitations There is, of course, no free lunch, as if features are strongly correlated, conditioning by intervention can lead to model evaluation on points outwith the true data manifold. This can visualized with the simple illustration in Figure 2. Whilst intervening on independent features always yields samples within the original manifold, if features are very correlated, there is a possibility of extrapolating beyond the training distribution, where model behavior is unknown. In the remainder of this section we consider what impact this may have on the market outcomes. Multicollinearity inflates the variance of the coefficients, which can distort the estimated mean when the number of in-sample observations is limited.

The posterior variance of the i-th coefficient can be written as $\sigma^2(w_i) = \kappa_i / \xi |\mathcal{D}_t|$, where ξ is the intrinsic noise precision of the target and κ_i is the variance inflation factor, given by

$$\kappa_i = \mathbf{e}_i^\top \Big(\sum_{t' \leq t} (\boldsymbol{x}^{(t)})^\top \boldsymbol{x}^{(t)} \Big)^{-1} \mathbf{e}_i, \quad \forall i \in \mathcal{I},$$

where $\mathbf{e}_i$ is the i-th basis vector. Whilst $\kappa_i \geq 1$, it has no upper bound, meaning $\kappa_i \mapsto \infty$, $\forall i$, with increasing extent of collinearity.

Figure 2 Interventions producing points outwith the data manifold.



Note. Green and red lines are level sets within which 0.99 quantile of the training data when features are independent and correlated, respectively. The blue lines represent the data extrapolated as a result of intervening on X_1 and X_2.

From a variance decomposition perspective, the Shapley value of feature i equals the variance in the target signal that it explains, such that, $\mathbb{E}[\phi_i]^{(t)} = (\mathbb{E}[w_i]^{(t)})^2\ var(X_i^{(t)})$, approximating the behaviour of the interventional Shapley value when features are correlated (Owen and Prieur 2017). With a Gaussian posterior, the Shapley values follow a noncentral Chi-squared distribution with one degree of freedom. We can write the probability density function for the distribution of the Shapley value for feature i in closed-form as

$$\begin{aligned} &p(\phi_i^{(t)}) \\ &= var(X_i^{(t)}) var^{(t)}(w_i) \sum_{n=0}^{\infty} \frac{e^{\eta/2}}{n!} \left(\frac{\eta}{2}\right)^n \chi^2(1+2n), \end{aligned}$$

where $var^{(t)}(\cdot)$ is the estimated variance at time t and the noncentral Chi-squared distribution is seen to simply be given by a Poisson-weighted mixture of central Chi-squared distributions, $\chi^2(\cdot)$, with noncentrality $\eta = (\mathbb{E}[w_i]^{(t)})^2 / var^{(t)}(w_i)$, for which the moment generating function is known

in closed form. For feature i, the centered second moment is

$$var^{(t)}(\phi_i) = 2var^{(t)}(w_i) \times \left(2\mathbb{E}[w_i]_t^2 + var^{(t)}(w_i)\right)\left(var(X_i^{(t)})\right)^2$$

so the variance of the allocation for any feature is a quadratic function of the variance of the corresponding coefficient, thus the variance inflation induced by multicollinearity. That being said, this is only a problem for small sample sizes and vanishes with increasing t, as $var^{(t)}(w_i) \mapsto 0$, $\forall i$ (Qazaz et al. 1997). If only a limited number of observations are available, distorted revenues could be remedied using *zero-Shapley* or *absolute-Shapley* proposed in Liu (2020), or restricting evaluations to the data manifold (Taufiq et al. 2023). We leave an investigation into these remedies in relation to analytics markets to future work.

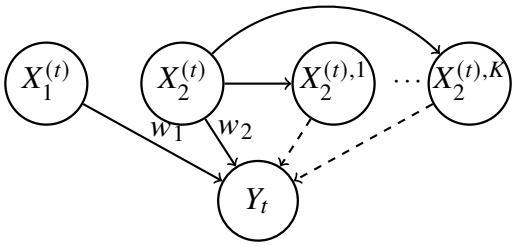
4. Robustness To Replication

Although it is natural for datasets to contain some overlapping information, Agarwal et al. (2019) show that in analytics markets such redundancy may also arise as a result of *malicious* behavior, in the sense that it is done to increase one's own reward at the expense of others'. This problem arises from the fact that data can be replicated freely, which differentiates it from material commodities, a trait which has motivated reassessments into fundamental mechanism design concepts for selling it (Aiello et al. 2001). In this section, we show that the use of observational conditional expectations in existing works explains the existence of replication incentives, the downsides of this, and how it can be remedied with the interventional lift.

DEFINITION 3 (**REPLICATE**). A replicate feature i is the original data obfuscated with noise, $x_i^{(t)} + \eta_i^{(t)}$, where $\eta_i^{(t)}$ is drawn from a centered distribution with finite variance, conditionally independent of the target given the feature.

Obfuscating a feature in this manner is equivalent to regularizing it's coefficients during training (Bishop 1995), inducing an endogeneity bias that diminishes the feature's contribution and, consequently, the revenue generated by the support agent. This idea underpins the proof of the truthfulness property described in Definition 2 as provided in Falconer et al. (2024). However, this property does not account for the fact that agents could, in theory, submit multiple replicates along with their original feature, each under a false identity. Whilst this would not impact predictive performance, it allows agents to increase their own revenue and diminish that of others whilst providing no additional improvements if the observational lift is used to calculate the Shapley values.

Figure 3 Direct effects (solid) and indirect effects (dashed) induced by replicating $X_2^{(t)}$.



Note. The k-th replicate of $X_2^{(t)}$ is denoted by $X_2^{(t),k}$.

To illustrate this, consider the causal graph in Figure 3 and let $x_i^{(t),k} = x_i^{(t)} + \eta_i^{(t),k}$ denote the k-th replicate of feature i. Suppose that $x_1^{(t)}$ and $x_2^{(t)}$ are identical features, such that $w_1 = w_2$, and that each is owned by a unique support agent, a_1 and a_2, respectively. With Theorem 1, the reward to each support agent without any replication will be $\pi^{(t)}/2$, where recall $\pi^{(t)}$ is the market revenue. Now suppose that a_2 replicates their feature K times and for simplicity assume $var(\eta_i^{(t),k}) = 0$ for every k. Using the same logic, the reward of a_1 is

$$\pi_{a_1}^{(t)} = \frac{\pi^{(t)}}{2+K},$$

and for agent a_2 the reward will be

$$\pi_{a_2}^{(t)} = \sum_{k=1}^{1+K} \frac{\pi^{(t)}}{2+K} = \frac{\pi^{(t)}(1+K)}{2+K},$$

hence a malicious agent can replicate their data many times so as to maximize their overall revenue, and diminish that of others, since $\pi_{a_1}^{(t)} \to 0$ as $K \to \infty$.

If support agent $a \in \mathcal{A}_{-c}$ replicates a feature K times, let the original feature vector augmented to include all of the additional replicates be $\boldsymbol{x}^{(t),+} \in \mathbb{R}^{|\mathcal{I}|\times K}$, with an analogous index set, $\mathcal{I}^+$.

DEFINITION 4 (**WEAKLY REPLICATION-ROBUST**). An analytics market is *weakly* robust to replication if $\pi_a^{(t),+} \leq \pi_a$, $\forall a \in \mathcal{A}_{-c}$, where $\pi_a^{(t),+}$ is the reward derived using $\boldsymbol{x}^{(t),+}$ instead.

REMARK 2. *Of course, a support agent may still choose to add noise to their feature for privacy reasons, wherein the loss of revenue can be perceived as the cost of privacy. So by replication-robust, we specifically refer to the malicious behavior of submitting multiple replicates of the same feature.*

Definition 4 is the definition of replication-robustness presented in Agarwal et al. (2019), stating that an agent who submits replicates of their feature in addition to the original should obtain weakly less reward than before. To achieve this, the authors propose *Robust-Shapley*, defined as follows:

$$\phi_i^{(t),\text{robust}} = \phi_i^{(t)} \exp\left(-\gamma \sum_{j \in \mathcal{I}_{-c}} \text{sim}\left(X_i^{(t)}, X_j^{(t)}\right)\right),$$

where $\text{sim}(\cdot,\cdot)$ is some measure of similarity (e.g., cosine similarity). This method penalizes similar features so as to remove the incentive for replication, thereby satisfying Definition 4. However, the issue with this approach is that not only replicated features are penalized, but also those with naturally occurring correlations between features. As a result, budget balance is lost, the extent to which depends on the chosen similarly metric and the value of γ. In addition, this leaves the market susceptible to spiteful agents—those willing to sacrifice their reward in order to minimize that of others. For this reason we refer to this definition as *weakly* robust.

A similar result is presented in Han et al. (2023) who consider the general set of semivalues, the class of solution concepts to submodular games to which the Shapley value belongs (Dubey et al. 1981). The authors show that the way in which a semivalue weights coalition sizes has an affect on the resultant properties, and that the Banzhaf value (Lehrer 1988) is in fact replication-robust by design (i.e., with respect to Definition 4), along with many other semivalues, albeit still penalizing naturally occurring correlations whilst being susceptible to spiteful agent.

DEFINITION 5 (**STRICTLY REPLICATION-ROBUST**). An analytics market is *strictly* robust to replication if $\pi_a^{(t),+} \equiv \pi_a$, $\forall a \in \mathcal{A}_{-c}$.

PROPOSITION 1. *With the proposed regression framework and Shapley value-based revenue allocation, regression markets using the interventional lift are strictly replication-robust.*

Proof With Definition 3, each replicate in $\boldsymbol{x}^{(t),+}$ only induces an indirect effect on the target. However, from Theorem 1, we know that the interventional lift only captures direct effects. Therefore, for each of the replicates, we write the marginal contribution for a single permutation $\theta \in \Theta$

as

$$\begin{aligned}\delta_i^{(t),\text{int}}(\theta) &= \zeta_{\underline{C}}^{(t),\text{int}} - \zeta_{\underline{C}\cup i}^{(t),\text{int}}, \\ &\int h(\boldsymbol{x}_{\underline{C}}^{(t)}, \boldsymbol{x}_{\underline{C}\cup i}^{(t)}) p(\boldsymbol{x}_{\underline{C}\cup i}^{(t)} | \boldsymbol{x}_{\underline{C}}^{(t)}) d\boldsymbol{x}_{\underline{C}\cup i}^{(t)} \\ &\quad - \int h(\boldsymbol{x}_{\underline{C}\cup i}^{(t)}, \boldsymbol{x}_{C}^{(t)}) p(\boldsymbol{x}_{C}^{(t)} | \boldsymbol{x}_{\underline{C}}^{(t)}) d\boldsymbol{x}_{C}^{(t)}, \\ &= 0, \quad \forall i \in \mathcal{I}_{-c}^{+} \setminus \mathcal{I}_{-c},\end{aligned}$$

and therefore $\phi_i \propto \sum_{\theta\in\Theta} \Delta_i(\theta) = 0$ for each of the replicates. For the original features, any direct effects will remain unchanged, as visualized in Figure 3. This leads to

$$\begin{aligned}\pi_a^{(t),+} &= \sum_{i\in\mathcal{I}_a} \lambda \mathbb{E}[\phi_i]^{(t)} + \sum_{i\in\mathcal{I}_a^+\setminus\mathcal{I}_a} \lambda \underbrace{\mathbb{E}[\phi_i]^{(t)}}_{=0} \\ &= \pi_a^{(t)}, \ \forall a \in \mathcal{A}_{-c},\end{aligned}$$

showing that by replacing the conventional observational lift with the interventional lift, Shapley value-based allocation is robust to replication *and* spitefulness by design.

5. Experimental Analysis

We now validate our findings on a real-world case study.We use an open source dataset to facilitate reproduction of our work, namely the Wind Integration National Dataset (WIND) Toolkit, detailed in Draxl et al. (2015). Our setup is a stylised continuous electricity market where agents—in our case, wind producers—need to notify the system operator of their expected electricity generation in a forward stage, one hour ahead of delivery, for which they receive a fixed price per unit. In real-time, they receive a penalty for deviations from the scheduled production, thus their downstream revenue is an explicit function of forecast accuracy.

Data Description This dataset contains wind power measurements simulated for 9 wind farms in South Carolina (USA), all located within 150 km of each other—see Table 1 for a characteristic overview. Although this data is not exactly *real*, it effectively captures the spatio-temporal aspects of wind power production, with the added benefit of remaining free from any spurious measurements, as can often be the case with real-world datasets. Measurements are available for a period of 7

Table 1 Agents and corresponding site characteristics considered in South Carolina (USA). C_f denotes the capacity factor and P the nominal capacity. The identify number is that from the WIND Toolkit database.

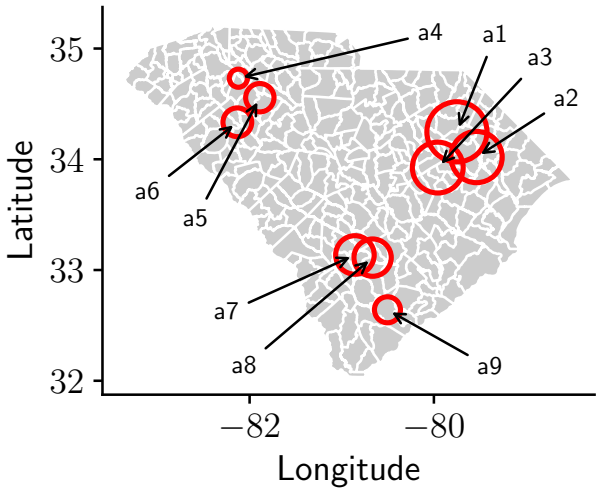
Agent	Id.	C_f (%)	P (MW)
a_1	4456	34.11	1.75
a_2	4754	35.75	2.96
a_3	4934	36.21	3.38
a_4	4090	26.60	16.11
a_5	4341	28.47	37.98
a_6	4715	27.37	30.06
a_7	5730	34.23	2.53
a_8	5733	34.41	2.60
a_9	5947	34.67	1.24

years, from 2007 to 2013, with an hourly granularity, which we normalize to take values in the range of $[0, 1]$.

Each wind farm is considered a market agent. For simplicity, we let a_1 be the central agent, however in practice each could assume this role in parallel. We assume each agent to have only 1 feature, namely the 1-hour lag of their power measurements—for wind power forecasting, the lag not only captures the temporal correlations of the production at a specific site, but also indirectly encompasses the spatial dependencies amongst neighboring sites due to the natural progression of wind. To illustrate this, we plot the location of each site in Figure 4. We see that the measurements at sites directly neighbouring a_1 have the largest dependency, which then decreases for the sites further away.

Methodology We use the regression framework described in Section 2, with an *Auto-Regressive with eXogenous input* model, such that each agent is assumed to own a single feature, namely a 1-hour lag of their power measurement. We are interested in assessing market outcomes rather than competing with state-of-the-art forecasting methods, so we use a very short-term lead time (i.e., 1-hour ahead), permitting fairly simple time-series analyses. We focus on assessing rewards rather than competing with state-of-the-art forecasting methods, so we use a very short-term lead time, permitting fairly simple time-series analyses. Nevertheless, our mechanism readily allows more complex models for those aiming to capture specific intricacies of wind power production, for instance the bounded extremities of the power curve (Pinson 2012).

Figure 4 Geographic location of each wind farm.



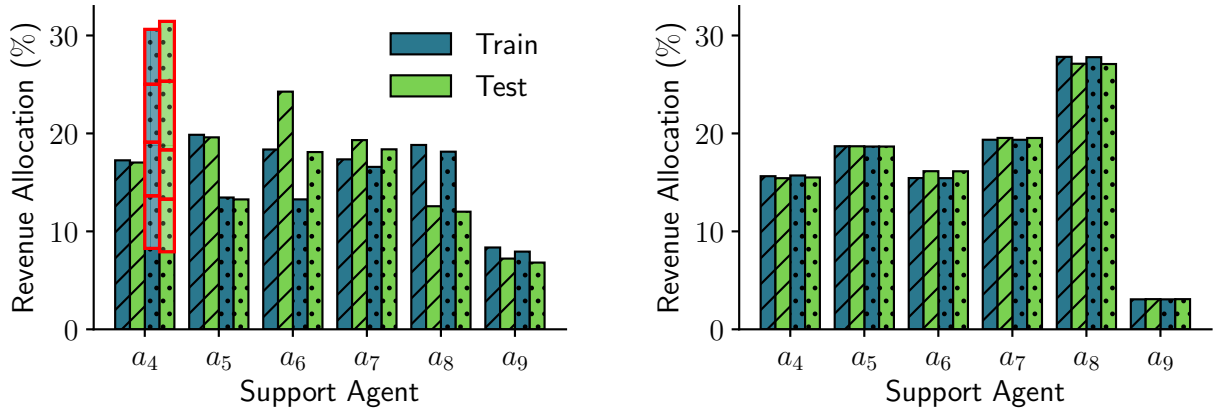
Note. The point sizes indicate the relative correlation between the measurements at each site and that of the central agent, a_1.

We perform a pre-screening, such that given the redundancy between the lagged measurements of a_2 and a_3 with that of a_1, we remove them from the market in line with our assumptions. At every time step, once a new observation of the target signal arrives, the previous time step's forecast is applied for out-of-sample market clearing. Simultaneously, the posterior is updated, the in-sample market is cleared, and a forecast for the next time step is generated. We clear both markets considering each agent is honest, that is, they each provide a single report of their true data. Next, we re-clear the markets, but this time assuming agent a_4 is malicious, replicating their data, thereby submitting multiple separate features to the market to increase their revenue. This problem size doesn't require approximate Shapley values, but recall findings hold either way, and generalize theoretically to arbitrary numbers of agents.

Results We set the central agent's to valuation to $\lambda = 0.5$ USD per time step and per unit improvement in h, for both in-sample and out-of-sample market stages. However, we are primarily interested in reward allocation rather than the magnitude—see Pinson et al. (2022) for a complete analysis of the monetary incentive to each agent participating in the market. Overall the expected in-sample and out-of-sample losses improved by 10.6% and 13.3% respectively with the help of the support agents. This improvement is unaffected bu the number of replicates, since they provide no additional information.

Setting $K = 4$, in Figure 5, we plot the expected allocation for each agent both with and without the malicious behavior of agent a_4, for each lift. When a_4 is honest, we observe that the observational lift spreads credit relatively evenly amongst features, suggesting that many of them have similar

Figure 5 Revenue allocations for each support agent.

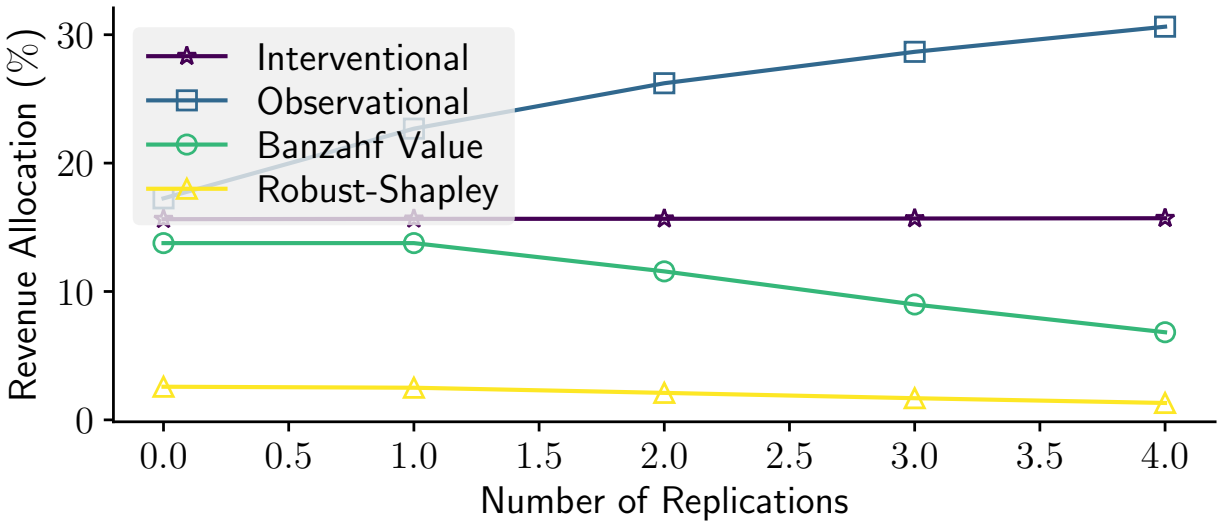


(a) *Observational*: Revenue of a_4 is increased due to indirect effects induced by the replicates.

(b) *Interventional*: Revenue of a_4 remains the same by accounting only for direct effects.

Note. Results for both (a) observational and (b) interventional lifts, when agent a_4 is honest (//) and malicious (∘) by replicating their feature. The blue and green bars correspond to in-sample and out-of-sample market stages, respectively. The revenue split amongst replicates is depicted by the stacked bars highlighted in red.

Figure 6 Revenue allocation of agent a_4 with increasing number of replicates



indirect effects on the target. The interventional lift favours agents a_7 and a_8, which, as one would expect, own the features with the most spatial correlation with the target. In this market, most of the additional revenue of agent a_8 appears to be lost from agent a_9 compared with the observational lift, suggesting that whilst these features are correlated, it is agent a_8 with the greatest direct effect, which is intuitive given their geographic location.

When agent a_4 replicates their data, with the observational lift, agents a_5 to a_8 earn less, whilst agent a_4 earns more. This shows that this lift indeed spreads rewards proportionally amongst indirect effects, of which there are four more due to the replicates, and so the malicious agent out-earns the others. Since the interventional lift only attributes direct effects, each replicate gets zero reward, so the malicious agent is no better off than before. Rewards were consistent between in-sample and out-of-sample, likely due to the large sample size and limited nonstationarities within the data.

To compare our work against current literature, in Figure 6 we plot the allocation of agent a_4 with increasing number of replicates. Here, *Robust-Shapley* and *Banzahf Value* refer to both the penalization approach of Agarwal et al. (2019) and the use of another semivalue in Han et al. (2023), respectively. With the observational lift, the proportion of revenue obtained increases with the number of replicates, as in the previous experiment. With *Robust-Shapley*, the allocation indeed decreases with the number of replicates, demonstrating this approach is *weakly* replication-robust, but is considerably less compared with the other approaches since natural similarities are also penalized. The authors argue this is an incentive for provision of unique information, but this allows agents to be spiteful. The *Banzahf Value* is strictly robust to replication for $K = 1$, but only weakly for $K \geq 2$. Lastly, unlike these methods, our proposed use of the interventional lift remains strictly replication-robust throughout as expected, with agent a_4 not able to benefit from replicating their feature, without penalizing the other agents.

6. Conclusions

Many machine learning tasks could benefit from using the data owned by others, however convincing firms to share information, even if privacy is assured, poses a considerable challenge. Rather than relying on data altruism, analytics markets are recognized as a promising way of providing incentives for data sharing, many of which use Shapley values to allocate revenue. Nevertheless, there are a number of open challenges that remain before such mechanisms can be used in practice, one of which is vulnerability to strategic replication, which we showed leads to undesirable reward allocation and restricts the practical viability of these markets.

We introduced a general framework for analytics markets for supervised learning problems that subsumes many of these existing proposals. We demonstrated that there are several different ways to formulate a machine learning task as cooperative game and analysed their differences from a causal perspectives. We showed that use of the observational lift to value a coalition is the source of these replication incentives, which many works have tried to remedy through penalization

methods, which facilitate only *weak* robustness. Our main contribution is an alternative algorithm for allocating rewards that instead uses interventional conditional probabilities. Our proposal is robust to replication without comprising market properties such as budget balance. This is a step towards making Shapley value-based analytics markets feasible in practice.

From a causal perspective, the interventional lift has additional potential benefits, including reward allocations that better represent the reliance of the model on each feature, providing an incentive for timely and reliable data streams for useful features, that is, those with greater influence on predictive performance. It is also favorable with respect to computational expenditure. That said, when it comes to data valuation, the Shapley value is not without its limitations—it is not generally well-defined in a machine learning context and requires strict assumptions, not to mention its computational complexity. This should incite future work into alternative mechanism design frameworks, for example those based on non-cooperative game theory instead.

References

Abernethy J, Chen Y, Ho CJ, Waggoner B (2015) Low-cost learning via active data procurement. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 619–636.

Agarwal A, Dahleh M, Sarkar T (2019) A marketplace for data: An algorithmic solution. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 701–726.

Aiello B, Ishai Y, Reingold O (2001) Priced oblivious transfer: How to sell digital goods. *International Conference on the Theory and Applications of Cryptographic Techniques*, 119–135 (Springer).

Bergemann D, Bonatti A (2019) Markets for information: An introduction. *Annual Review of Economics* 11(1):85–107.

Bishop CM (1995) Training with noise is equivalent to tikhonov regularization. *Neural computation* 7(1):108–116.

Castro J, Gómez D, Tejada J (2009) Polynomial calculation of the shapley value based on sampling. *Computers & operations research* 36(5):1726–1730.

Chalkiadakis G, Elkind E, Wooldridge M (2011) Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5(6):1–168.

Chen H, Janizek JD, Lundberg S, Lee SI (2020) True to the model or true to the data?

Chen H, Lundberg SM, Lee SI (2022) Explaining a series of models by propagating shapley values. *Nature Communications* 13(1):4512.

Covert IC, Lundberg S, Lee SI (2021) Explaining by removing: A unified framework for model explanation. *The Journal of Machine Learning Research* 22(1):9477–9566.

Cummings R, Ioannidis S, Ligett K (2015) Truthful linear regression. Grünwald P, Hazan E, Kale S, eds., *Proceedings of the 28th Conference on Learning Theory*, 448–483 (Paris, France).

Dekel O, Fischer F, Procaccia AD (2010) Incentive compatible regression learning. *Journal of Computer and System Sciences* 76(8):759–777.

Deng X, Papadimitriou CH (1994) On the complexity of cooperative solution concepts. *Mathematics of operations research* 19(2):257–266.

Draxl C, Clifton A, Hodge BM, McCaa J (2015) The wind integration national dataset (wind) toolkit. *Applied Energy* 151:355–366.

Dubey P, Neyman A, Weber RJ (1981) Value theory without efficiency. *Mathematics of Operations Research* 6(1):122–128.

Falconer T, Kazempour J, Pinson P (2024) Bayesian regression markets. *Journal of Machine Learning Research* 25(180):1–38, URL `http://jmlr.org/papers/v25/23-1385.html`.

Frye C, Rowat C, Feige I (2020) Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems* 33:1229–1239.

Gal-Or E (1985) Information sharing in oligopoly. *Econometrica: Journal of the Econometric Society* 329–343.

Ghorbani A, Zou J (2019) Data shapley: Equitable valuation of data for machine learning. *International conference on machine learning*, 2242–2251 (PMLR).

Han D, Wooldridge M, Rogers A, Ohrimenko O, Tschiatschek S (2023) Replication robust payoff allocation in submodular cooperative games. *IEEE Transactions on Artificial Intelligence* 4(5):1114–1128.

Heskes T, Sijben E, Bucur IG, Claassen T (2020) Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems* 33:4778–4789.

Janzing D, Minorics L, Blöbaum P (2020) Feature relevance quantification in explainable ai: A causal problem. *International Conference on Artificial Intelligence and Statistics*, 2907–2916 (PMLR).

Koutsopoulos I, Gionis A, Halkidi M (2015) Auctioning data for learning. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 706–713 (IEEE).

Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S (2020) Problems with shapley-value-based explanations as feature importance measures. *International Conference on Machine Learning*, 5491–5500.

Lehrer E (1988) An axiomatization of the banzhaf value. *International Journal of Game Theory* 17:89–99.

Liu J (2020) Absolute shapley value.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30.

Merrill J, Ward G, Kamkar S, Budzik J, Merrill D (2019) Generalized integrated gradients: A practical method for explaining diverse ensembles.

Mitchell R, Cooper J, Frank E, Holmes G (2022) Sampling permutations for shapley value estimation. *Journal of Machine Learning Research* 23(43):1–46.

Mussell J (2014) Raw data is an oxymoron.

Ohrimenko O, Tople S, Tschiatschek S (2019) Collaborative machine learning markets with data-replication-robust payments. URL `https://arxiv.org/abs/1911.09052`.

Owen AB, Prieur C (2017) On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification* 5(1):986–1002.

Pearl J (2010) An introduction to causal inference. *The international journal of biostatistics* 6(2).

Pearl J (2012) The do-calculus revisited. *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 3–11, UAI'12 (Arlington, Virginia, USA: AUAI Press), ISBN 9780974903989.

Pinson P (2012) Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61(4):555–576.

Pinson P, Han L, Kazempour J (2022) Regression markets and application to energy forecasting. *TOP* 30(3):533–573.

Qazaz CS, Williams CK, Bishop CM (1997) An upper bound on the bayesian error bars for generalized linear regression. *Mathematics of Neural Networks: Models, Algorithms and Applications*, 295–299 (Springer).

Rasouli M, Jordan MI (2021) Data sharing markets. *arXiv preprint arXiv:2107.08630* .

Ravindranath SS, Jiang Y, Parkes DC (2024) Data market design through deep learning. *Advances in Neural Information Processing Systems* 36.

Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, Bakas S, Galtier MN, Landman BA, Maier-Hein K, et al. (2020) The future of digital health with federated learning. *NPJ digital medicine* 3(1):1–7.

Shapley LS (1997) A value for n-person games. *Classics in Game Theory* 69.

Storkey A (2011) Machine learning markets. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 716–724 (JMLR Workshop and Conference Proceedings).

Sundararajan M, Najmi A (2020) The many shapley values for model explanation. *International Conference on Machine Learning*, 9269–9278.

Taufiq MF, Blöbaum P, Minorics L (2023) Manifold restricted interventional shapley values. *International Conference on Artificial Intelligence and Statistics*, 5079–5106 (PMLR).

Wolfers J, Zitzewitz E (2004) Prediction markets. *Journal of economic perspectives* 18(2):107–126.

Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10(2):1–19.