IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2022, accepted January 25, 2022, date of publication January 31, 2022, date of current version February 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3147799

Data Boundary and Data Pricing Based on the Shapley Value

YINGJIE TIAN[1,2,3], YURONG DING[1,2,3], SAIJI FU [1,2,3], AND DALIAN LIU[4,5]

[1]School of Economics and Management, University of Chinese Academy of Sciences, Haidian, Beijing 100190, China
[2]Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Haidian, Beijing 100190, China
[3]Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Haidian, Beijing 100190, China
[4]Institute of Mathematics and Physics, Beijing Union University, Beijing 100101, China
[5]Institute of Fundamental and Interdisciplinary Sciences, Beijing Union University, Beijing 100101, China

Corresponding author: Dalian Liu (ldtdalian@buu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 12071458, Grant 71731009, and Grant 61472390.

ABSTRACT Data pricing is to price data as asset to promote the healthy development of data share, exchange, and reuse. However, the uncertain value index and neglection of interactivity lead to information asymmetry in the transaction process. A perfect data pricing system and a well-designed data market can widely promote data transactions. We take a three-agent data market in this paper, the data owner who provides the data record, the model buyer who is interested in buying machine learning (ML) model instances, and the broker who interacts between the data owner and the model buyer. Two interaction scenarios are defined. In scenario 1, we introduce the Shapley value (SV) to measure values of data records fairly and construct a revenue optimization problem based on the sum of the SV for data boundary. The optimal solution is obtained by calculating their derivatives. In scenario 2, we utilize market research and construct a revenue maximization (RM) problem to price the ML models. Further, An Integer Linear Programming for the RM problem (RM-ILP) process is proposed to transform the RM problem into an equivalent integer linear programming (ILP) problem and solve it with the Gurobi solver. Finally, we conduct extensive experiments, which validate that the RM-ILP process can provide high revenue to the broker and high affordability to the model buyers compared to the benchmarks.

INDEX TERMS Data value, data pricing, the shapley value, model pricing.

I. INTRODUCTION

Data is the core asset in the era of data-driven economy, which significantly promotes the rapid development of the data transaction industry. Big data empowers business value, leads to significant changes, and even becomes a driving factor for business model innovation. According to the report from International Data Corporation, the worldwide revenue of big data analysis markets will grow up to $260 billion in 2022 [1]. However, using data to generate business value in many scenarios needs further improvement. For small organizations, insufficient data will hinder the establishment of better models and can not meet the needs of analysis, application and decision support [2]. Recent studies and practices have approached the commoditization of data in various ways, leading to the concepts of data pricing and data market. Data pricing is to price data as asset and ensure that data can be sold or purchased as a commodity in business and economic activities [3]. The data market [4], [5] is where we trade data for revenue. Perfect data pricing system and well-designed data market can promote data transactions broadly.

For data pricing systems, there are numerous ways to trade data or data products. A widely accepted approach for applications is to set prices to datasets directly based on their values. From an economic perspective, data value has been explored to study the impact on companies and industries [5]–[7]. From a computer science perspective, data value focuses on data compensation [8], [9] and data fusion [10], [11]. It is crucial to measure data value correctly to avoid the ''lemons'' market and promote the development of healthy share, exchange and reuse. However, the value of data is varied for different agents in trading process. So determining a unified value index and

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

getting a standard in the data market are worth further exploration.

Meanwhile, current data pricing system forces users to buy the entire dataset without considering the balance between utility and cost. When the cost is considered [10], ''the more, the better'' principle does not always hold for data trading. We assume that in the data market, the broker needs to make a trade-off between benefits and costs. They buy necessary data only if the marginal gain is higher than the marginal cost. The corresponding amount of purchased data is called the data boundary n. So, determining the data boundary and ensuring a well-designed data price mechanism to incentivize data flow are necessary.

For the data market, a three-agent data market based on model pricing has been proposed recently [9], [12], [13], shown in Fig.1, including data providers, the broker, and model buyers. The existing model-based pricing data market is a value stream of data supply, processing and application, dedicated to ML model instances and focusing on pricing models instead of the datasets. Data owners provide structured and labelled data records for sale, and the broker [14] combines data records into a dataset D to train models with various accuracy level options. The model buyers specify a model they are interested in and buy it directly. The main functions and assumptions of the three agents will be explained in detail in Section III-A.

Existing pricing schemes force the broker to buy the whole dataset when purchasing data from the data owner, without awareness of values of each data records because the data is usually packaged and sold as a dataset. This means that valuable datasets may not be defined clearly and broker operate in an inefficient market, where he do not maximize his revenue. The challenge is: How to formulate the broker's requisites on data records and data boundary based on the SV, and construct an efficient data trading process? As with the same practice of selling digital commodities in several versions, existing work [15] provides a set of models for sale with various accuracy levels. The challenge is: How to formulate the model prices directly align with the demands of model buyers on model utility? How to model the RM problem and solve it effectively with an arbitrage-free guarantee?

To solve the problems mentioned above, we define a three-agent data market and propose two revenue optimization frameworks for data pricing. The main contributions can be summarized as

- A three-agent data market consistent with the Dealer market [13] is introduced. Two interaction scenarios are clearly defined between data owners and a broker (scenario 1), a broker and model buyers (scenario 2).
- The significance of data boundary is proposed to build an efficient data trading process. To find the data boundary and records on the subscription service, we creatively use the sum of SV to fit the model buyers' WTP and assume that the broker determines the data records and data boundary based on the revenue optimization framework.
- To obtain the model prices align with the demands of model buyers on model utility, we design an RM-ILP process and transform the RM problem into an equivalent ILP problem, which can provide high revenue to the broker and high affordability to the model buyers.
- Extensive experiments are conducted on real datasets to find the data boundary and records in scenario 1. Results in scenario 2 demonstrate the effectiveness of the RM-ILP process compared with benchmarks.

The remainder of this article is organized as follows. Section II reviews the related work. Section III introduces the background and preliminaries, including data markets and interactions scenarios, the Shapley value, and the arbitrage-free properties. Section IV presents the revenue optimization problems. Section V provides the experimental results. Section VI summarizes the conclusions.

II. RELATED WORK

In this section, we make a detailed review of the pricing methods for data and SV for data valuation.

A. DATA PRICING

Pricing Methods for Data: There are numerous works directly trading and pricing data as an economical product in the data markets. According to the type of sold products and the corresponding pricing mechanism, the related studies can be divided into: **1. Data-based pricing.** They sell datasets and allow buyers to value the data directly based on the data features, including data information entropy levels [16], [17], data quality levels [18], [19], the fresh and real-time features [20], [21], and the amounts of data in the dataset [22]. Also, competitive data trading model [23] considered the willingness-to-sell of data providers and the willingness-to-pay of model buyers. They proved that the model has the unique Nash equilibrium and maximized the profits of business stakeholders. However, in the research of data-based pricing, data value only depends on the performance of a single dimension. With the increase of dimension, the complexity of calculation increases greatly. Meanwhile, data buyers have to purchase the whole dataset even if they are only interested in particular information extracted from the dataset. **2. Query-based pricing**. Query-based pricing was proposed as QueryMarket [24], [25]. They argued that the data owners upload the entire datasets with clear price points defined in advance to the market service platform. Then data buyers submit queries to the platform, and the platform generates the objective function and corresponding constraints, forming an ILP problem to get the prices. Researchers improved their studies and proposed a framework that allows the price of any query to be derived automatically [26]. QIRANA pricing system [27] allows flexible pricing by allowing the data owners to choose from a variety of pricing functions, specify relation and attribute-level parameters to control the prices of queries. However, most queries considered by these marketplaces are too simplistic to support sophisticated data analytics and decision making. Recently, Chawla *et al.* [28] investigated

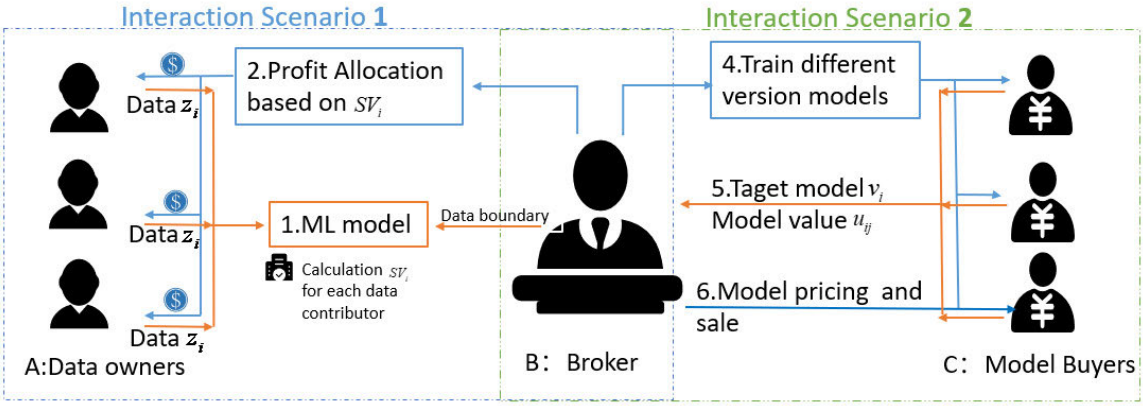


FIGURE 1. Three agents in the data market.

three types of succinct pricing functions and studied the corresponding revenue maximization problems. **3. Model-based pricing and Arbitrage-free**. The key challenges in data markets were discussed in [29], which shows that the most crucial problem is arbitrage-free pricing. Balazinska *et al.* [29] formulated arbitrage-free pricing functions defined by arbitrary queries and proposed new necessary conditions for avoiding arbitrage. Combining arbitrage-free properties and model-based pricing, Chen *et al.* [15] proposed the first MBP framework, which directly prices ML model instances with different accuracy options via noise injection. They provided solutions for sellers to determine prices of models under different market scenarios. Later, Liu *et al.* [13] proposed an end-to-end model marketplace named Dealer from the perspective of current data marketplace. The Dealer markets formulate the data owners' compensation function based on the SV and privacy sensitivity. Efficient dynamic programming algorithms DPP and DPP+ are employed to solve the corresponding RM problem, which actually cannot capture the different demand and value curves in the market research. To summarize, there is still broad room for exploring new methods to handle the aforementioned issues.

B. THE SHAPLEY VALUE AND DATA VALUATION

The Shapley Value: The SV has been recently applied for data valuation and according to the contribution of data records to the models [8], [9], [12], [30]. The SV is originated in cooperative game theory for measuring the contribution of each participant fairly, which was named in honour of Lloyd Shapley [31]. On the one hand, the SV uniquely possesses properties and supports different utility functions flexibly. On the other hand, empirical studies prove that the SV is capable of giving more insights into the importance of each data point and identifying outliers or corrupted data [8]. Apparently, the SV has been widely applied in data pricing due to its unique advantages. We introduced it to measure data value, and establish WTP and utility function for data pricing.

Calculating the exact SV requires enumerating all subsets and retraining $O(2^n)$ models to get the marginal contribution. Many approximation methods have been developed to overcome the computational hardness. The most representative method is the Truncate Monte-Carlo method (TMC-Shapley) [32], [33]. This method calculates the approximate SV based on the random sampling of permutations. Similarly, a series of effective algorithms are proposed, such as group test based method and heuristic method based on influence function, which are suitable for all ML models. While, many approximate algorithms based on a unique model have been proposed, for instance, the exact computation for KNN classifiers [9], q-fair for federated learning [34], etc. Distributional Shapley [35] further considered statistical aspects of the data, and the data value was measured in the context of underlying data distribution. Since data and data products can be reproduced at a marginal cost close to zero, and adverse data replication may lead to grossly undesirable revenue divisions [33], the Shapley-Robust algorithm [12] is used to protect against adversarial replication and maintain the conditions of Shapley fairness. In this paper, to facilitate the calculation, we use the TMC-Shapley algorithm to solve the SV.

III. BACKGROUND AND PRELIMINARIES

In this section, the background and preliminaries of the Dealer data market are introduced. The participants and their interactions are discussed in Section III-A. The SV is adopted to measure the data value in Section III-B, and the importance of arbitrage-freeness in the data market is explained in Section III-C. For convenience, the notations are summarized in Table 1 and the acronyms are summarized in Table 2.

A. DATA MARKET AND INTERACTIONS

The existing tripartite data market [14], [23] mainly includes the data owners, the broker, and the model buyers, as shown in Fig.1. In this subsection, we formalize the data market dynamics, which consists of the interactions between data

TABLE 1. A list of notations.

Notation	Definition
D	Dataset $D = \{(\tilde{\mathrm{x}}_i, \mathrm{y}_i)\}_1^N$.
z_i	$z_i = (\tilde{\mathrm{x}}_i, \mathrm{y}_i)$, a piece of data in D.
N	The size of D
SV_i	The Shapley value of z_i.
n	The data boundary, means the SV_i of the data record is ranked n_{th} in D.
m	The number of potential model buyers.
Ssv_n	The sum of SV_i of the top n data, $Ssv_n = \sum_i^n SV_i$.
η	The sensitivity coefficient, $WTP = \eta \cdot Ssv_{\mathrm{n}}$.
c	The collection cost per unit SV.
p_s	The service subscription fee from the broker to model buyers.
$\Pi(\cdot)$	The revenue function of the broker.
$\langle v_1, \cdots, v_N \rangle$	The different model versions in different accuracy levels.
$\langle v_i, u_{ij} \rangle$	v_i denotes the target model and u_{ij} denotes the model value for the j_{th} model buyers.
$p(v_i)$	The price for the model version v_i set by the broker.

TABLE 2. A list of acronyms.

Acronym	Definition
RM	Revenue Maximization problem in (18).
RM-ILP	The process that transform the RM problem into an equivalent Integer Linear Programming problem.
MBP	Model-based Pricing market.
WTP	The willingness-to-pay for model buyers.
SV	The Shapley value.
ML	Machine Learning.
DPP	The dynamic programming algorithm with just the model value prices space.
DPP+	The dynamic programming algorithm with complete solution space.
TMC-Shapley	The Truncate Monte-Carlo method for the Shapley value.
LND	Log-normal distribution.

owners and a broker (scenario 1), a broker and model buyers (scenario 2). Two revenue optimization frameworks are formed separately to investigate the two challenges mentioned in Section I.

1) AGENTS

- **Data Owners.** The data owners provide the relational dataset D for sale. We assume that data record z_i have undergone a certain degree of data processing to achieve data desensitization and data privacy protection. They are of legal origin, with no illegal privacy, and meet the data analysis, storage and application of the transaction conditions. Without losing generality, we assume that one data owner D_i possesses only one data record z_i, making it convenient to expand to group data records.
- **Broker.** The broker purchases the dataset and trains different versions of models with different accuracy level options for service subscriptions or direct model sales. The broker then allocates the revenue based on the SV to different data owners.
- **Model Buyers.** The buyer specifies an ML model, which learns over the dataset D with their preferences for accuracy levels. Two transaction modes can be seen below in interaction scenario 2.

2) INTERACTIONS

- **Scenario 1: interactions between data owners and a broker.** AS presented in the blue box in Fig.2, the data owners provide different data records to the broker. We suppose that data owners offer data records to combine datasets with training different versions of ML models for selling. Then the broker allocates the revenue fairly, using the SV as their contributions. In this scenario, the broker makes a trade-off in data utility and cost, purchasing some valuable data records rather than the whole dataset. A maximized revenue function is set based on SV of the dataset and used to find the optimal data boundary and data records.
- **Scenario 2: interactions between the broker and model buyers.** As shown in the green box in Fig.3, the broker supplies a menu of different kinds of ML models with different accuracy levels, and we propose that there are two kinds of transaction modes:

1) Service subscription. The model buyers subscribe to the service provided by the broker to meet business needs and pay the subscription fee denoted by p_s. Whether to subscribe to the service or not depends on their WTP for the service.
2) Single model purchase. As with the same practice of selling digital commodities in several versions, broker provides a set of models for sale with different levels of accuracy based on the SV coverages for model buyers. The broker sets the price $p(v_i)$ for every model version [8], [15], and model buyers purchase only one models they need.

B. THE SHAPLEY VALUE

The data owners submit their data records to the broker in our three-agent system. Then the broker trains ML models on data owners' data record and provides service to the model buyers, as shown in Fig.2. Two challenges are proposed here. One is how to measure the contributions and usefulness of each data for the ML model, and the other is how to distribute the payment from the broker back to the data owners. A natural way of tackling those challenges is to adopt a game-theoretic viewpoint, where each data contributor is modeled as a player in a coalitional game. The SV is a classic method in cooperative game theory to distribute the total gains generated by the coalition of all players and has been applied to problems in various domains, especially in the ML domain [8], [13], [35].

We modeled the cooperative game as follows. Consider N data owners $D_1, \ldots, D_N$ and data owner D_i owns data record $z_i (1 \leq i \leq N)$. We assume a utility function $U(S) (S \subseteq \{1, \ldots, N\})$ that evaluates the utility of a coalition S, which consists of data records from multiple data owners. Their cooperation aims to train the corresponding ML model and provide service to model buyers. The SV_i for the user is defined as (1) to measure the marginal utility improvement,

then averaged over all possible coalitions of the data owners [8], [13].

$$SV_i = \sum_{s \subseteq \{z_1, z_2 \cdots z_N\} \setminus z_i} \frac{V(s \cup \{z_i\}) - V(s)}{\binom{N-1}{|s|}}, \tag{1}$$

where s is the subset of D, $|s|$ denotes the size of the selected subsets, $V(\cdot)$ is the utility function.

The key to obtaining SV_i is to select an appropriate ML algorithm and calculate the marginal value of the data record z_i according to the dataset D. The importance of the SV stems from the fact that it is the unique value division scheme satisfied the following desirable properties [15].

1) **Group rationality.** The value of the entire dataset is completely distributed among all users and satisfies $SV(D) = \sum_{i \in D} SV_i$;
2) **Fairness**. Fairness is reflected in two aspects. On one hand, if the user i and j satisfy $V(s \cup \{z_i\}) = V(s \cup \{z_j\})$, $\forall S \subseteq D \backslash \{z_i, z_j\}$, then $SV_i = SV_j$. It implies that if two users are identical to all utility of subsets, they should have the same value. On the other hand, if the user satisfies $s \subseteq D \backslash \{z_i\}$ and $V(s \cup \{z_i\}) = V(s)$, then $SV_i = 0$. It signifies that if the marginal contribution of a certain subset to all other subsets is zero, the corresponding SV should be zero;
3) **Additivity**. Consider two utility functions $U(\cdot)$ and $V(\cdot)$, satisfying $SV(U, i) + SV(V, i) = SV(U + V, i)$, $\forall z_i \in D$. For more general case, the sum of SV based on different utility functions is equal to the SV based on the sum of utility functions [8].

Intuitively, for an ML task, the dataset D, the ML algorithm, and the utility function $V(\cdot)$ are three main components in computing the SV. The utility function can be considered from **test-prediction**, **test-loss** and **self-loss** aspects. Existing literatures mainly use evaluation metrics as the utility function. We take the accuracy evaluation index as the value function in our classification test. For a more convenient and quick calculation, the TMC-Shapley [8] is adopted to distribute revenue to data owners, as shown in Algorithm 1. The computational complexity of TMC-Shapley is based on the model we are training, and more details can be found in those articles [8], [32], [36].

C. ARBITRAGE-FREENESS

Suppose there are two options for a model buyer pursuing high accuracy. Option1: buy a model directly at a high price. Option 2: buy a set of low-accuracy models at low prices and "combine" them to achieve the desired accuracy. If the cost of option 1 is greater than that in option 2, then there is room for arbitrage transactions, which will complicate the interface between buyers and the broker. The buyers may need to reason carefully about their purchase behavior to achieve the lowest price, while the broker may not achieve the maximum revenue. Therefore, it is necessary to design arbitrage-free pricing functions, which are expected to meet the following properties [13], [15], [37]:

Algorithm 1 Truncated Monte Cralo Shaply

Input: Train dataset $D = \{(x_i, y_i)\}_1^N$, Learning algorithm A, Utility function $V(\cdot)$
Output: the Shapley value (SV) of all training point z_i

Initialize $SV_i = 0$ for $i = 1, \ldots, N$ and $t = 0$
while Convergence criteria not met **do**
  $t \leftarrow t + 1$
  π^t: Random permutation of train data points
  $v_0^t \leftarrow V(z_i, A)$
  for $j \in \{1, \ldots, N\}$ **do**
    if $\left|V(D) - v_{j-1}^t\right| <$ Performance Tolerance **then**
      $v_j^t = v_{j-1}^t$
    else
      $v_j^t = V(\{\pi^t[1], \ldots, \pi^t[j]\}, A)$
    end if
    $SV_{\pi^t[j]} \leftarrow \frac{t-1}{t} SV_{\pi^{t-1}[j]} + \frac{1}{t}(v_j^t - v_{j-1}^t)$
  end for
end while

- **Monotonicity**: Given a function $f : (R^+)^k \rightarrow R^+$, it is monotone if and only if for any two vectors $x, y \in (R^+)^k : x \leq y$, we have $f(x) \leq f(y)$;
- **Subadditivity**: Given a function $f : (R^+)^k \rightarrow R^+$, it is subadditive if and only if for any two vectors $x, y \in (R^+)^k : x \leq y$, we have $f(x + y) \leq f(x) + f(y)$.

IV. PROBLEM FORMULATION

This section describes two revenue optimization problems for the broker. To begin with, WTP based on the sum of SV is introduced to establish a revenue optimization framework to find the data boundary (Section IV-A). Later, the RM-ILP process is proposed for single model pricing (Section IV-B).

A. DATA BOUNDARY

"The more, the better" principle does not always hold for data trading when considering the cost, and will lead to an inefficiency in the data market by contrast [10]. Data boundary n is the data records where the marginal gain is higher than the marginal cost when the SV of the data is sorted from largest to smallest. We form a revenue optimization framework to find the data boundary according to WTP of model buyers.

For model convenience, the data market's mechanism is based on the service subscription, and the interactive process is presented in the blue box in Fig.2. Data owners provide data records and combine a dataset D. The SV is given by the third-party organization and calculated by Algorithm 1. The broker chooses an optimal data boundary n and then offers model buyers with diverse ML models. Correspondingly, model buyers subscribe to the service and pay a subscription fee denoted by p_s to the broker.

We sort the SV_i for D and let Ssv_n denote the sum of SV given by the top n data, that is $Ssv_n = \sum_{i=1}^{n} SV_i$. Ssv_n means

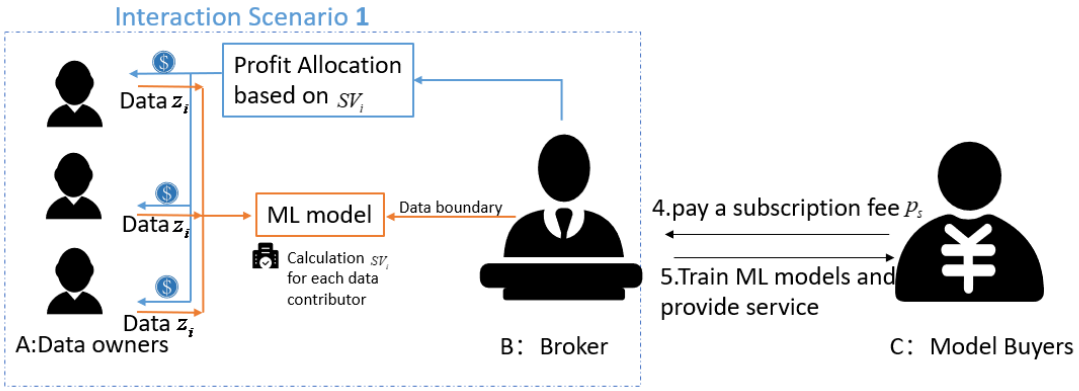


FIGURE 2. Agents and their interactions in secenario 1.

the accuracy level that the model can reach after adding n_{th} data record. It is believed that the broker have sufficient reasons to purchase data records from high to low according to SV to maximize profits. Model buyers have different WTP, depending on the Ssv_n, and increase as it grows. In general, we assume $WTP = \eta \cdot Ssv_n(\eta \geq 0)$, and the utility function satisifies:

- $\frac{dSsv_n}{dn} > 0$: the utility is an increasing function;
- $\frac{d^2Ssv_n}{d^2n} \leq 0$: the marginal utility of the utility function decreases.

Equation (2) in [19], [22] is used to fit the correlation between Ssv_n and top n data records. Apparently, it satisfies the assumptions of the utility function.

$$Ssv_n = \sum_{i}^{n} sv_i = \beta_1 - \beta_2 \cdot \exp(-\beta_3 \cdot n). \tag{2}$$

To estimate parameters in (2) for different datasets, we calculate the SV_i of raw datasets and get the experiment points $(n_1, Ssv_1), \ldots, (n_N, Ssv_N)$. The non-linear least squares algorithm optimizes $\beta = (\beta_1, \beta_2, \beta_3)$ by minimizing the sum of square errors as

$$\min_{\beta_1,\beta_2,\beta_3} \sum_{i=1}^{N} (Ssv_i - (\beta_1 - \beta_2 \cdot \exp(-\beta_3 \cdot n_i)))^2. \tag{3}$$

Model buyers have different WTP for the service. Concretely, if the WTP of a model buyer is higher or equal to the subscription fee p_s charged by the broker, the model buyer will subscribe to the service, otherwiset he will not purchase. Let w denote the actual WTP of a model buyer and w' denote a nominal WTP. The actual WTP depends on the Ssv_n and let $w = w' \cdot Ssv_n$. Thus, w' is a positive random variable, reflecting the heterogeneity of model buyers. Let the probability density is denoted by $f(w')$, and two distributions are put up for different cases.

1) CASE 1. UNIFORM DISTRIBUTION

Let W' denote the maximum nominal WTP, and thus we have $W = W' \cdot Ssv_n$ for the maximum actual WTP. Given a set of model buyers, and the probability density $f(w')$ follows a uniform distribution $[0, W']$. It implies that model buyers will subscribe in a specific interval with the same probability. The cumulative distribution function $F(p_s)$ means the probability that W' is less or equal to p_s. Therefore, the probability of subscription is expressed as

$$p_r = 1 - F(p_s) = W - p_s. \tag{4}$$

Then, the broker's profit can be presented by (5). The first term is the revenue gained from the model buyers who are willing to pay higher or eaqul to the p_s, and the second term is the cost paid to the data owners.

$$\begin{aligned} \Pi(p_s, n) &= \underbrace{p_s \cdot m \cdot p_r}_{\text{revenue}} - \underbrace{c \cdot n}_{cost}, \\ &= p_s \cdot m \cdot (W - p_s) - c \cdot n, \\ &= sp_s \cdot m \cdot \left(W' \cdot Ssv_n - p_s\right) - c \cdot n, \end{aligned} \tag{5}$$

where $Ssv_n = \sum_{i}^{n} SV_i = \beta_1 - \beta_2 \cdot \exp(-\beta_3 \cdot n)$, $(\beta_1, \beta_2, \beta_3 \geq 0)$.

Till now, an optimization problem (6) can be formulated to obtain the optimal subscription fee p_s^* and data boundary n^*.

$$\begin{aligned} &\max_{p_s,n} \Pi(p_s, n) = p_s \cdot m \cdot \left(W' \cdot Ssv_n - p_s\right) - c \cdot n \\ &\text{s.t. } p_s \geq 0, \\ &\quad n \geq 0. \end{aligned} \tag{6}$$

The constraints of (6) ensure non-negative solution of p_s and n. By differentiating $\Pi(p_s, n)$ with respect to n or p_s, we obtain

$$\frac{\partial \Pi(p_s, n)}{\partial p_s} = m \cdot \left(W' \cdot Ssv_n - 2p_s\right). \tag{7}$$

$$\frac{\partial \Pi(p_s, n)}{\partial n} = p_s \cdot m \cdot W' \cdot \beta_2 \cdot \beta_3 \cdot \exp(-\beta_3 \cdot n) - c. \tag{8}$$

The second derivative $\Pi(p_s, n)$ that n or p_s is fixed, as shown in (9) and (10). Therefore, the solutions of the uniform distribution are globally optimal. We can obtain

the solutions of n^* and ${p^*}_s$ by solving $\frac{\partial \Pi(p_s,n)}{\partial n} = 0$ and $\frac{\partial \Pi(p_s,n)}{\partial p_s} = 0$.

$$\frac{\partial^2 \Pi(p_s,n)}{\partial^2 p_s} = -2 \cdot m < 0. \tag{9}$$

$$\frac{\partial^2 \Pi(p_s,n)}{\partial^2 n} = -\beta_2 {\beta_3}^2 \cdot \left(p_s \cdot m \cdot W'\right) \cdot \exp\left(-\beta_3 \cdot n\right) < 0. \tag{10}$$

2) CASE 2. lOG-NORMAL DISTRIBUTION

Since a lot of natural phenomena follow the log-normal distribution (LND [38]), it is helpful for practically modelling this process. Define that w' follows an LND distribution, and the probability density is $f(w')$. It impliying that most model buyers will subscribe at low prices. However, those who cancel subscriptions will rely more on price fluctuations near reasonable price. Recall that the cumulative distribution the function $F(p_s)$ means the probability where W' is less or equal to p_s. Therefore, the probability of subscription is expressed as

$$p_{\mathrm{r}} = 1 - F\left(p_s\right) = \frac{1}{2} - \frac{1}{2} erf\left(\frac{\ln p_s - \mu}{\sqrt{2}\sigma}\right), \tag{11}$$

where $erf(x)$ is the error function used in measure theory.

Then, the model buyer's profit can be formulated as

$$\begin{aligned}
&\Pi\left(p_s, n\right) \\
&= \underbrace{p_s \cdot m \cdot p_r}_{\text{revenue}} - \underbrace{c \cdot n}_{cost} \\
&= p_s \cdot m \cdot \left(\left(\frac{1}{2} - \frac{1}{2} erf\left(\frac{\ln p_s - \mu}{\sqrt{2}\sigma}\right)\right) \cdot Ssv_n\right) - c \cdot n \\
&= \left(\frac{p_s \cdot m}{2} \cdot \left(1 - erf\left(\frac{\ln p_s - \mu}{\sqrt{2}\sigma}\right)\right) \cdot Ssv_n\right) - c \cdot n,
\end{aligned} \tag{12}$$

where $Ssv_n = \sum_i^n SV_i = \beta_1 - \beta_2 \cdot \exp\left(-\beta_3 \cdot n\right), \left(\beta_1, \beta_2, \beta_3 \geq 0\right)$.

The first term of (13) is the revenue gained from the model buyers who are willing to pay higher or eaqul to the p_s, and the second term is the cost paid to the data owners.

The LND distribution can be used flexibly by adjusting the μ and σ values according to different situations. For the sake of convenience, we now assume $\mu = 0$ and $\sigma = 1$, and an optimization problem can be formulated as

$$\begin{aligned}
&\max_{p_s,n} \ \Pi\left(p_s, n\right) = \frac{p_s \cdot m}{2} \cdot \left(1 - erf\left(\frac{\ln p_s}{\sqrt{2}}\right)\right) \cdot Ssv_n - c \cdot n \\
&\text{s.t. } p_s \geq 0, \\
&\quad n \geq 0,
\end{aligned} \tag{13}$$

where the constraints ensure non-negative solution of p_s and n.

By differentiating $\Pi\left(p_s, n\right)$ concerning p_s and n, we obtain

$$\frac{\partial \Pi\left(p_s, n\right)}{\partial p_s} = \frac{m \cdot Ssv_n}{2} \cdot \left(1 - erf\left(\frac{\ln p_s}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} \cdot \exp\left(-\frac{\ln^2 p_s}{2}\right)\right). \tag{14}$$

$$\frac{\partial \Pi\left(p_s, n\right)}{\partial n} = \left(\frac{p_s \cdot m}{2} \cdot \left(1 - erf\left(\frac{\ln p_s}{\sqrt{2}}\right)\right)\right) \cdot \left(\beta_2 \cdot \beta_3 \cdot \exp\left(-\beta_3 \cdot n\right)\right) - c. \tag{15}$$

The second order derivatives of $\Pi\left(p_s, n\right)$ with respect to p_s or n are fixed shown in (16) and (17), which non-positive. Therefore, the solutions of the LND distribution are globally optimal. We can obtain the solutions of n^* and ${p^*}_s$ by solving $\frac{\partial \Pi(p_s,n)}{\partial n} = 0$ and $\frac{\partial \Pi(p_s,n)}{\partial p_s} = 0$.

$$\frac{\partial^2 \Pi\left(p_s, n\right)}{\partial^2 p_s} = -\frac{m \cdot Ssv_n}{\sqrt{2\pi} \cdot p_s} \cdot \left(1 + \ln p_s\right) \cdot \exp\left(-\frac{\ln^2 p_s}{2}\right) < 0. \tag{16}$$

$$\frac{\partial^2 \Pi\left(p_s, n\right)}{\partial^2 n} = -\left(\beta_2 {\beta_3}^2\right) \left(\frac{p_s \cdot m}{2} \cdot \left(1 - erf\left(\frac{\ln p_s}{\sqrt{2}}\right)\right)\right) \cdot \exp(-\beta_3 \cdot n) < 0. \tag{17}$$

The following section will check the unique maximum point $({p_s}^*, n^*)$ and maximum value $\Pi\left({p_s}^*, n^*\right)$ on two real datasets.

B. MODEL-BASED PRICING

We further focus on the interaction between broker and model buyer, as shown in the green box in Fig.3. Market research is introduced to pay more attention to model buyers' purchase intention. The broker uses the dataset D and trains different versions of models $\langle v_1, v_2 \cdots v_N \rangle$ for sale, ordered with the accuracy levels. According to the target model v_i and model value u_{ij} obtained by market research and characteristics of data market's principles, we established a fair model pricing mechanism and maximized the broker's revenue. Assume that the model buyers are interested in a model v_i with model value u_{ij}, model buyers will buy the model only if $p(v_i) \leq u_{ij}$ and the broker will get the revenue $p(v_i)$. The RM problem can be written as

$$\begin{aligned}
&\max_{\langle p(v_1), p(v_2) \cdots p(v_{\mathrm{N}}) \rangle} \ \sum_i^N \sum_j^M p(v_i) I(p(v_i) \leq u_{ij}) \\
&\text{s.t. } p(v_i) > p(v_j) \geq 0, \quad v_i \geq v_j, \\
&\quad p(\mathrm{v}_i) + p(v_j) \geq p(v_i + v_j), \quad \mathrm{v}_i, v_j \geq 0.
\end{aligned} \tag{18}$$

where $p(v_i)$ are integer variables and $I(p(v_i) \leq u_{ij})$ is an indicator variable that takes value one when $p(v_i) \leq u_{ij}$ otherwise takes value zero. Constraints ensure the arbitrage-free properties.

Problem (18) is challenging to solve due to the existence of the indicator function $I(p(v_i) \leq u_{ij})$. The indicator variable means that the objective function is a piecewise function. We use the binary integer variable x_{ij1}, x_{ij2} and new constraints to transform the nonlinear objective function into a

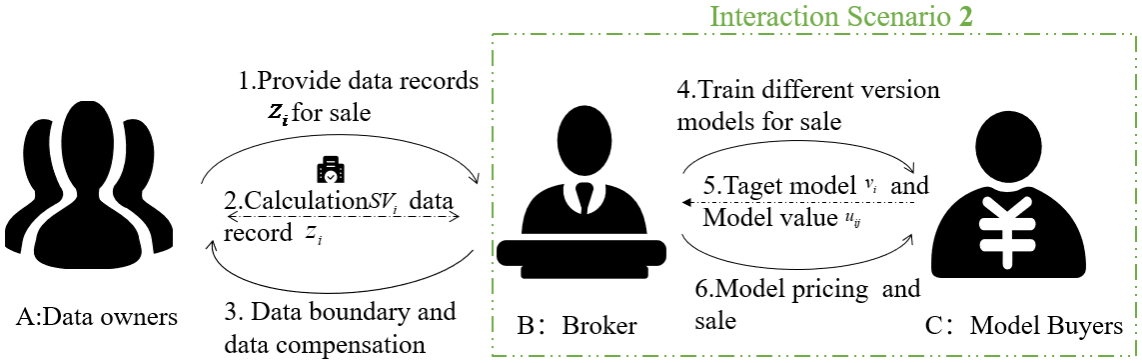


FIGURE 3. Agents and their interactions in secenario 2.

quadratic integer programming problem, rewritten as (19).

$$\begin{aligned}\max_{\langle p(v_i),x_{ij1},x_{ij2}\rangle} \quad & \sum_{i}^{N}\sum_{j}^{M} p(v_i)\cdot x_{ij1}\\ \text{s.t. } & p(v_i) > p(v_j) \geq 0, \quad v_i \geq v_j,\\ & p(v_i)+p(v_j) \geq p(v_i+v_j), \quad \forall v_i, v_j,\\ & x_{ij2}\cdot u_{ij} \leq p(v_i) \leq x_{ij1}\cdot u_{ij} + x_{ij2}\cdot L,\\ & x_{ij1}+x_{ij2}=1,\\ & 0 \leq x_{ij1}, x_{ij2} \leq 1, \end{aligned} \tag{19}$$

where L is infinity and $p(v_i), x_{ij1}, x_{ij2}$ are all integer variables. The first and second constraints ensure the arbitrage-free properties, and the other constraints ensure the indicator function's transformation.

The following is a conditional analysis for j_{th} model buyer for the model i, and the model value is denoted as u_{ij}.

Case 1. when $x_{ij1}=1, x_{ij2}=0$**.**

The objective function changed into $p(v_i)$, and the third constraint changed into $0 \leq p(v_i) \leq u_{ij}$, which means that the price for the model v_i is lower than its model value, the model buyer buys this model, and the broker gets revenue $p(v_i)$.

Case 2. when $x_{ij1}=0, x_{ij2}=1$**.**

The objective function changed into 0, and the third constraint changed into $u_{ij} \leq p(v_i)(L \textit{ is infinity})$, which means that the price for the model v_i is higher than its model value, and the model buyer does not buy.

Further, since the objective function is characteristic of a binary integer variable x_{ij1}, we convert the quadratic linear programming into an ILP problem, shown as (20).

$$\begin{aligned}\max_{\langle p(v_i),x_{ij1},x_{ij2},y_{ij}\rangle} \quad & \sum_{i}^{N}\sum_{j}^{M} y_{ij}\\ \text{s.t. } & p(v_i) > p(v_j) \geq 0, \quad v_i \geq v_j,\\ & p(v_i)+p(v_j) \geq p(v_i+v_j), \quad \forall v_i, v_j,\\ & x_{ij2}\cdot u_{ij} \leq p(v_i) \leq x_{ij1}\cdot u_{ij} + x_{ij2}\cdot L,\\ & x_{ij1}+x_{ij2}=1,\\ & 0 \leq x_{ij1}, \quad x_{ij2} \leq 1,\\ & \mathrm{y}_{ij} \leq x_{ij1}\cdot L,\\ & \mathrm{y}_{ij} \leq p(v_i),\\ & \mathrm{y}_{ij} \geq p(v_i) - L\cdot(1-x_{ij1}),\\ & \mathrm{y}_{ij} \geq 0, \end{aligned} \tag{20}$$

where L is infinity and $p(v_i), x_{ij1}, x_{ij2}, y_{ij}$ are all integer variables. The first and second constraints ensure the arbitrage-free properties, the third to fifth constraints ensure the transformation of the indicator function, and the other constraints ensure the ILP transformation

The following is a conditional analysis for the ILP transformation.

Case 1. when $x_{ij1}=1$

The constraints $\mathrm{y}_{ij} \leq x_{ij1}\cdot L(L \textit{ is infinity})$ changed into $\mathrm{y}_{ij} \leq L$, and the constraints $\mathrm{y}_{ij} \geq p(v_i) - L\cdot(1-x_{ij1})$ changed into $\mathrm{y}_{ij} \geq p(v_i)$. Meanwhile, constraint ensures that $\mathrm{y}_{ij} \leq p(v_i)$. So the combination of three constraints ensures that $\mathrm{y}_{ij} = p(v_i)$. The objective function changed into $p(v_i)$, and the model buyer buys this model.

Case 2. when $x_{ij1}=0$

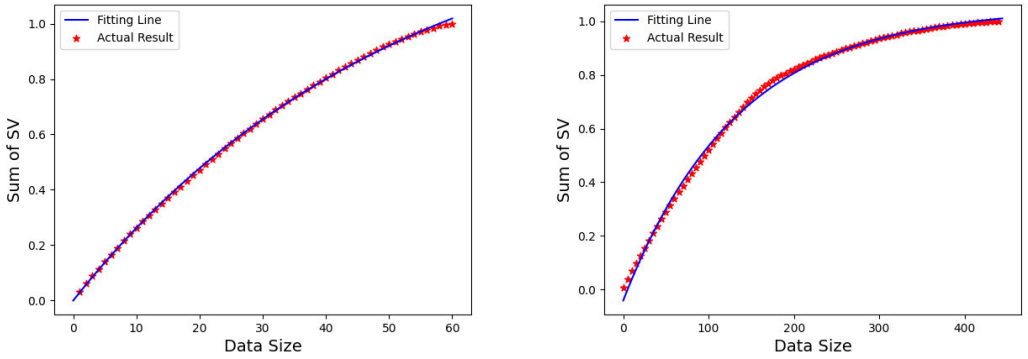
The constraints $\mathrm{y}_{ij} \leq x_{ij1}\cdot L(L \textit{ is infinity})$ changed into $\mathrm{y}_{ij} \leq 0$, and the constraints $\mathrm{y}_{ij} \geq p(v_i) - L\cdot(1-x_{ij1})$ changed into $\mathrm{y}_{ij} \geq -L$. Meanwhile, constraint ensures that $\mathrm{y}_{ij} \geq 0$. So the combination of three constraints ensures that $\mathrm{y}_{ij} = 0$. The objective function changed to 0, and the model buyer does not buy this model.

The above method that transforms the revenue maximization problem into an ILP problem is named as RM-ILP process and an effective solution can be obtained compared with other benchmarks.

V. EXPERIMENT

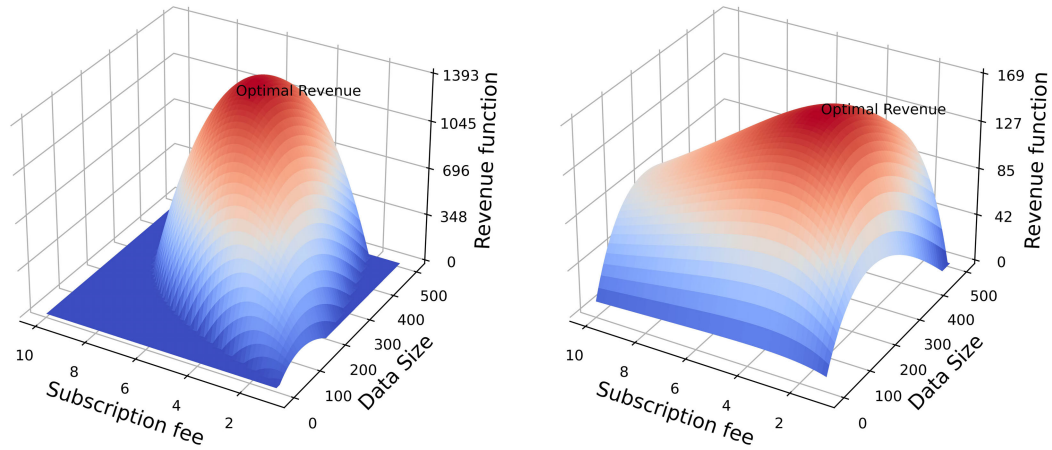
Experimental Setup: All experiments were run on a machine with 4 Intel i5-1021U 1.60 GHz cores, 16 GB RAM, and 500 GB disk with Windows 10 as the OS. We have prototyped the RM-ILP process and use Python with the package "Gurobi" to solve the ILP problem.

We present a numerical study of the proposed approaches on datasets, evaluating our method concerning



(a) Fitting curve for Anderson's Iris dataset (b) Fitting curve for Breast Cancer dataset

FIGURE 4. Estimation of the data Shapley value function of top *n* data using support vector machine. (a) is for the Anderson's Iris dataset and $\beta_1 = 1.475$,, $\beta_2 = 1.475$ and $\beta_3 = 0.0195$. (b) is for Breast Cancer dataset and $\beta_1 = 1.050$, $\beta_2 = 1.091$ and $\beta_3 = 0.007$.



(a) Revenue of the broker where WTP is based on uniform distribution (b) Revenue of the broker where WTP is based on Log-normal distribution

FIGURE 5. Revenue of the broker under varied data boundaries and subscription fees based on the Breast Cancer dataset.

scenario 1 and 2. Analysis of the data boundary is presented in Section V-A. The SV of two real datasets are evaluated, and the maximum revenue function is set up to get optimal data boundary and subscribe fee. Further,the RM-ILP process is conducted and compared with other benchmarks in Section V-B.

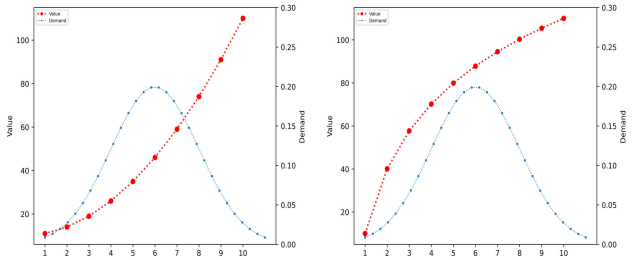
A. EXPERIMENT ON DATA BOUNDARY

We use Anderson's Iris and Breast Cancer datasets [39] and compute the SV for each data record separately. Then, the SV of D is normalized, and those with negative values are eliminated. We use (3) to fit the correlation between Ssv_n and top n data, shown in Fig.4.

We obtain the representative numerical results of the subscription fee p_s based on the SV of Breast Cancer dataset. Fix the number of potential model buyers to be 100, and both collection cost c per unit data records and the sensitivity coefficient η are equal to 1. The optimization problem can be verified visually in Fig.5. The maximum revenue can be achieved when the optimal data size and subscription fee are applied. Note that different classification ML algorithms and utility functions of SV can get similar results.

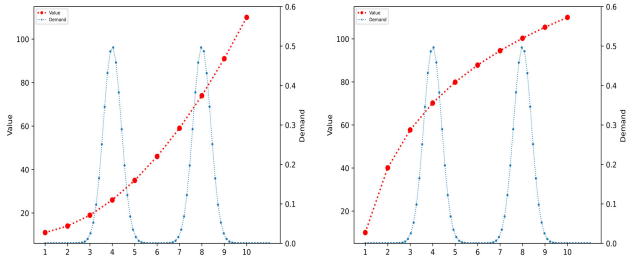
B. EXPERIMENT ON MODEL-BASED PRICING

Different versions of the model should have different expected means, shown as a curve of values. Moreover, we can capture "how many" buyers are interested in a particular model, which can be seen as a curve of demands distribution. The curves of values and demands can be obtained via market research to simulate the interactivity of the data market. we use $\langle v_i, u_{ij} \rangle$ to denote each market research point, where the v_i denotes the target model and u_{ij} denotes the model value for the j_{th} model buyer. We experimentally study the revenue maximization of our proposed method on four different datasets and compare them with four pricing approaches.



(a) Convex demand and normal distribution value

(b) Concave demand and normal distribution value



(c) Convex demand and bimodal distribution value

(d) Concave demand and bimodal distribution value

FIGURE 6. The (a) shows when the curve of demands is the normal distribution($\mu = 5$, $\delta = 3$), and the curve of values is convex; the (b) shows when the curve of demands is the normal distribution ($\mu = 5$, $\delta = 3$), and the curve of values is concave; the (c) shows when the curve of demands is bimodal distribution ($\mu_1 = 2$, $\mu_2 = 7$, $\delta = 3$), and the curve of values is convex; the (d) shows when the curve of demands is bimodal distribution ($\mu_1 = 2$, $\mu_2 = 7$, $\delta = 3$), and the curve of values is concave.

1) DATASETS

Four virtual datasets with different demands (normal distribution and bimodal distribution) and different values (concavity and convexity) are proposed. We first fix the demands and vary the values, as shown in Fig.6(a) and Fig.6(b), then fix values and vary demands, as shown in Fig.6(c) and Fig.6(d). Normal and bimodal distribution are set for the curves of demands, aiming to distinguish model buyers who are interested in medium accuracy and extremely low or high accuracy levels. For different model versions, the demand d_i and the value u_i are combined to generate a model value point u_{ij} which follows the independent distribution with range $[u_i - \mathrm{t}, u_i + \mathrm{t}]$. With t decreases, the range becomes narrower.

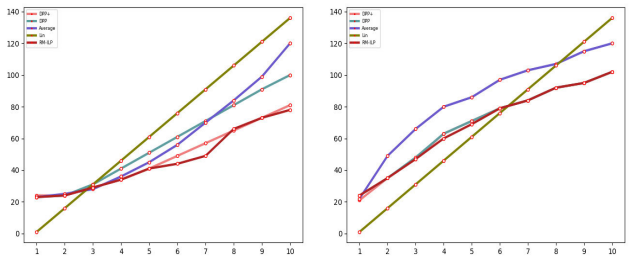
Two virtual datasets with various demands and versions based on uniform distribution are proposed to demonstrate the run time performance. We fixed the version number or value number to 15, and then set another one to change from 10 to 30,as shown in Fig.10(a) and Fig.10(b).

2) BENCHMARK METHODS

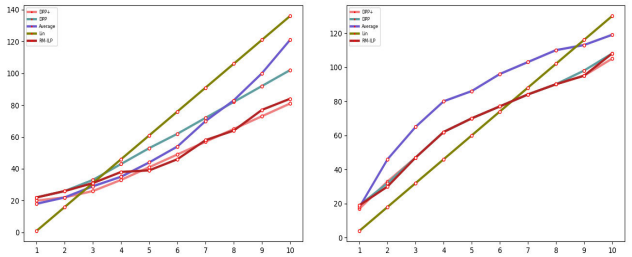
Lin: a linear approach that uses the lowest and the highest model value to form a linear function. The final prices are the interpolations of the linear function.

Average: the prices of models are the average of values in market research for different versions.

DPP: different prices are set for all versions, and the final prices are computed by the dynamic programming algorithm [13] with just the model value prices space.

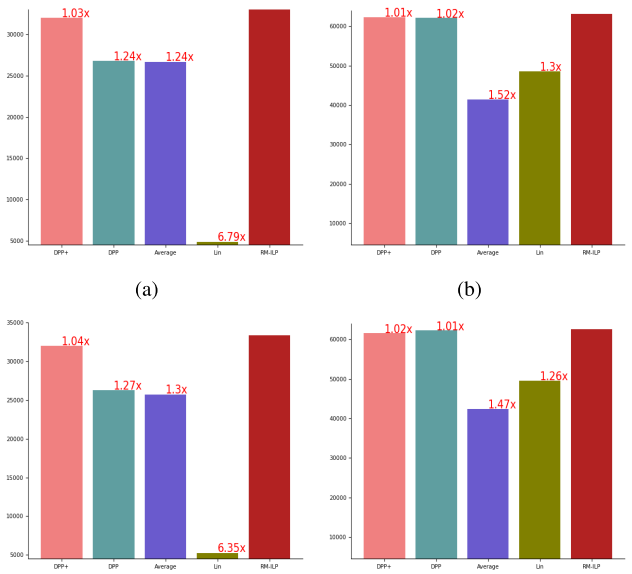


(a) (b)



(c) (d)

FIGURE 7. The price curves for curves of various values and demands. (a) The price curves for convex demand and normal distribution value. (b) The prices curves for concave demand and normal distribution value. (c) The prices curves for convex demand and bimodal distribution value. (d) The prices curves for concave demand and bimodal distribution value.



(a) (b) (c) (d)

FIGURE 8. The revenue results for curves of various values and demands. (a) The revenue results for convex demand and normal distribution value. (b) The revenue results for concave demand and normal distribution value. (c) The revenue results for convex demand and bimodal distribution value. (d) The revenue results for concave demand and bimodal distribution.

DPP+: different prices are set for all versions, and the final prices are computed by the dynamic programming algorithm with complete solution space [13].

Fig.7, Fig.8 and Fig.9 show the final model prices, revenue, and affordability ratio (fraction of the model buyers that can afford to buy a model). Fig.7 shows that the RM-ILP process

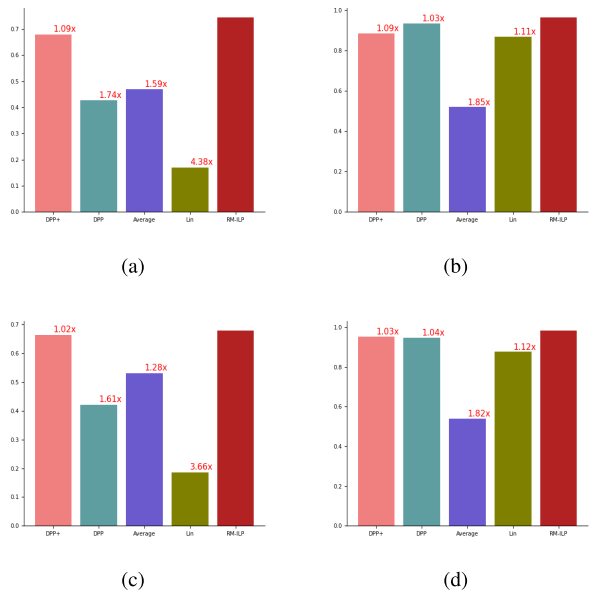


FIGURE 9. **The affordable ratio for curves of various values and demands. (a) The affordable ratio for convex demand and normal distribution value. (b) The affordable ratio for concave demand and normal distribution value. (c) The affordable ratio for convex demand and bimodal distribution value. (d) The affordable ratio for concave demand and bimodal distribution.**

has more accurate prices, especially in the curves of convex values. Fig.8 shows that the RM-ILP process can achieve up to 6.79× revenue gains compared to the four benchmarks and outperforms the DPP and DPP+ by 3% in the curves of convex value and 1% in the curves of concave value. Fig.9 shows that the RM-ILP process has the highest affordability ratio and outperforms the other algorithms by 2%.

DPP and DPP+ are compared with RM-ILP process firstly. They are algorithms based on dynamic programming and their results can be expressed as RRM (Relax Revenue Maximization). DPP is just based on the market research, while DPP+ is based on complete solution space given by additional interpolation points. The exact solution of RM problem in (17) is expressed as RM (Revenue Maximization). Liu *et al.* [13] proves that $\frac{RM}{2} \leq RRM \leq RM$. Compared with the RM-ILP process, neither DPP nor DPP+ algorithms capture the curves of convex values, as shown in Fig.8(a) and Fig.8(c). We still note that our method RM-ILP process can get more accurate solutions in different datasets. Lin and Average algorithms are compared with RM-ILP process secondly. As shown in Fig.8(a) and Fig.8(b), when the curve of values is convex, the Lin and Average approaches miss the opportunities to sell model instances with a medium accuracy level. When the curve of values becomes concave, as shown in Fig.8(b) and Fig.8(d), Lin and Average algorithms show a little better than the convex value curve. Generally speaking, all approaches show a slightly better concave curve than the convex value curve. The principal reason is that the curve of concave values is consistent with the characteristics of subadditivity [40].

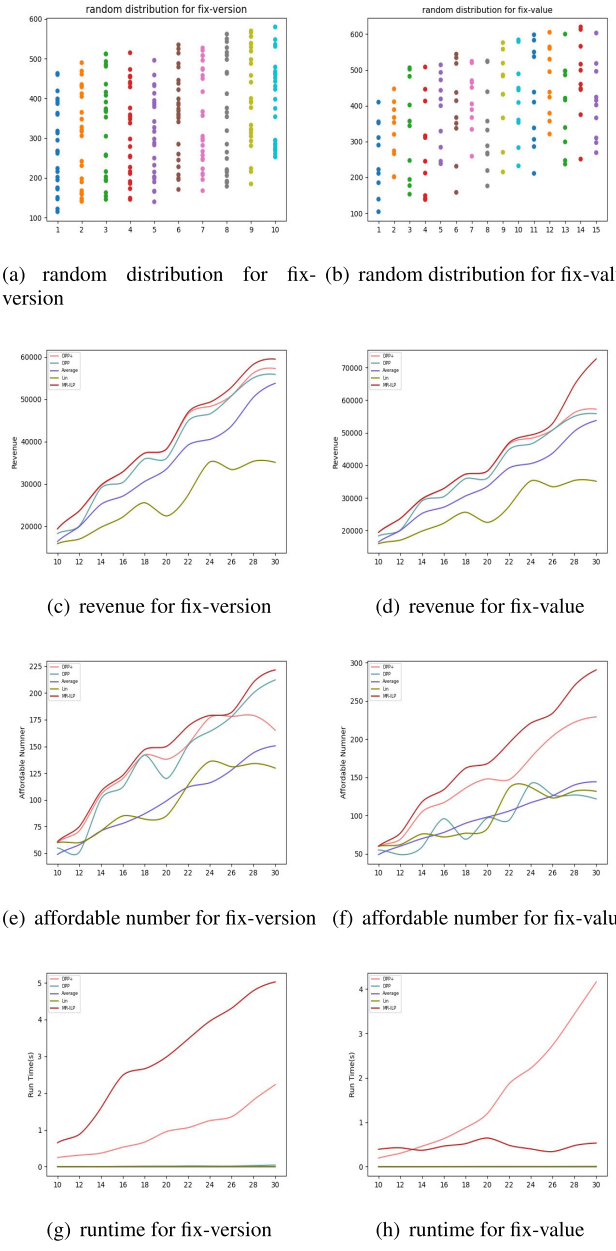


FIGURE 10. **Performance of RM-ILP process for random distribution data.**

Fixing the version numbers, we vary the number of value numbers and compare revenue gains, affordability ratio, and the runtime in Fig.10(c), Fig.10(e) and Fig.10(g), respectively. The RM-ILP process consistently produces a pricing curve with the highest affordability ratio and revenue while consistently slower than the benchmarks. This is because the RM-ILP process requires solving ILP exponentially many times. Fixing the value number, we vary the number of version numbers and compare revenue gains, affordability ratio, and the runtime in Fig.10(d), Fig.10(f) and Fig.10(h), respectively. Our proposed process consistently produces a pricing curve with the highest affordability ratio and revenue. Meanwhile, the RM-ILP process is faster than the DPP+ when various version numbers. This is because the

DPP+ algorithm requires constructing complete price space with the constraints of arbitrage-free properties, and it is a time-consuming process.

The RM-ILP process produces a pricing curve with the highest affordability ratio and revenue while slower than the benchmarks in many cases. Although the solution is time-consuming, we believe that the time complexity can be omitted compared with the loss of profit, especially in the offline application scenario of data transactions.

VI. CONCLUSION

In this work, we learned lessons from data markets that directly sell ML models instead of raw data in data markets. We defined the three-agent data market and introduced the SV to measure data value fairly. Under this data market, we assumed that the broker's decisions are derived from two revenue optimization problems: one is based on the sum of the SV of data records for data boundary, and the other utilized market research for model pricing. We proposed the RM-ILP process to obtain a more accurate and higher affordability solution. Further, extensive experiments based on curves of demands and values demonstrate that the proposed algorithm outperforms the benchmarks.

There are still limitations and broad room for improvement. First, understanding data value is a prerequisite for pricing data in different data trading scenarios. We focus on valuation in the particular context of training dataset in specific data transactions, and the SV has apparent shortcomings. A faster and more accurate measurement of data value is required. Second, privacy disclosure and compensation are also core future challenges. Making a fair privacy compensation and promoting a legal and healthy development of the data market is the point we should pay more attention to. Third, Some references utilize machine learning (mainly Reinforcement Learning) to solve dynamic pricing problems [41], [42]. Dynamic pricing allows enterprises to set flexible prices for information goods according to real-time demand. It is necessary to explore the dynamic pricing mechanism with the utilize of machine learning. Finally, there is only one broker in our market, but multiple brokers co-exist in practical applications, and the versions of models can be further optimized.

REFERENCES

[1] J.-Y. Cheng and P.-W. Dong, "Thinking of corporation financial management innovation in the era of big data," in *Proc. Int. Conf. Manage. Sci. Manage. Innov.* Paris, France: Atlantis Press, 2015, pp. 450–455.

[2] R. Raskar, P. Vepakomma, T. Swedish, and A. Sharan, "Data markets to support AI for all: Pricing, valuation and governance," 2019, *arXiv:1905.06462*.

[3] J. Pei, "A survey on data pricing: From economics to data science," *IEEE Trans. Knowl. Data Eng.*, early access, Dec. 21, 2020, doi: 10.1109/TKDE.2020.3045927.

[4] C.-L. Yeh, "Pursuing consumer empowerment in the age of big data: A comprehensive regulatory framework for data brokers," *Telecommun. Policy*, vol. 42, no. 4, pp. 282–292, 2018.

[5] P. Bajari, V. Chernozhukov, A. Hortaçsu, and J. Suzuki, "The impact of big data on firm performance: An empirical investigation," *AEA Papers Proc.*, vol. 109, pp. 33–37, May 2019.

[6] M. Farboodi, R. Mihet, T. Philippon, and L. Veldkamp, "Big data and firm dynamics," *AEA Papers Proc.*, vol. 109, pp. 38–42, May 2019.

[7] C. I. Jones and C. Tonetti, "Nonrivalry and the economics of data," *Amer. Econ. Rev.*, vol. 110, no. 9, pp. 2819–2858, 2020.

[8] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2242–2251.

[9] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. J. Spanos, and D. Song, "Efficient task-specific data valuation for nearest neighbor algorithms," 2019, *arXiv:1908.08619*.

[10] X. L. Dong, B. Saha, and D. Srivastava, "Less is more: Selecting sources wisely for integration," *Proc. VLDB Endowment*, vol. 6, no. 2, pp. 37–48, Dec. 2012.

[11] E. Junqué de Fortuny, D. Martens, and F. Provost, "Predictive modeling with big data: Is bigger really better?" *Big Data*, vol. 1, no. 4, pp. 215–226, Dec. 2013.

[12] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *Proc. ACM Conf. Econ. Comput.*, Jun. 2019, pp. 701–726.

[13] J. Liu, J. Lou, J. Liu, L. Xiong, J. Pei, and J. Sun, "Dealer: An end-to-end model marketplace with differential privacy," *Proc. VLDB Endowment*, vol. 14, no. 6, pp. 957–969, Feb. 2021.

[14] A. Rieke, H. Yu, D. Robinson, and J. van Hoboken, "Data brokers in an open society," Dept. Law, Inst. Inf. Law, Open Soc. Found., London, U.K., Nov. 2016, Tech. Rep., p. 64.

[15] L. Chen, P. Koutris, and A. Kumar, "Towards model-based pricing for machine learning in a data marketplace," in *Proc. Int. Conf. Manage. Data*, Jun. 2019, pp. 1535–1552.

[16] X. Li, J. Yao, X. Liu, and H. Guan, "A first look at information entropy-based data pricing," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jun. 2017, pp. 2053–2060.

[17] Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang, "A pricing model for big personal data," *Tsinghua Sci. Technol.*, vol. 21, no. 5, pp. 482–490, 2016.

[18] H. Yu and M. Zhang, "Data pricing strategy based on data quality," *Comput. Ind. Eng.*, vol. 112, pp. 1–10, Oct. 2017.

[19] J. Yang, C. Zhao, and C. Xing, "Big data market optimization pricing model based on data quality," *Complexity*, vol. 2019, pp. 1–10, Apr. 2019.

[20] M. Zhang, A. Arafa, J. Huang, and H. V. Poor, "Pricing fresh data," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1211–1225, May 2021.

[21] M. Zhang, A. Arafa, J. Huang, and H. V. Poor, "How to price fresh data," in *Proc. Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*, Jun. 2019, pp. 1–8.

[22] D. Niyato, M. A. Alsheikh, P. Wang, D. I. Kim, and Z. Han, "Market model and optimal pricing scheme of big data and Internet of Things (IoT)," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[23] H. Oh, S. Park, G. M. Lee, J. K. Choi, and S. Noh, "Competitive data trading model with privacy valuation for multiple stakeholders in IoT data markets," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3623–3639, Apr. 2020.

[24] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-Market demonstration: Pricing for online data markets," *Proc. VLDB Endowment*, vol. 5, no. 12, pp. 1962–1965, Aug. 2012.

[25] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with QueryMarket," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 613–624.

[26] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," *J. ACM*, vol. 62, no. 5, pp. 1–44, Nov. 2015.

[27] S. Deep and P. Koutris, "QIRANA: A framework for scalable query pricing," in *Proc. ACM Int. Conf. Manage. Data*, May 2017, pp. 699–713.

[28] S. Chawla, S. Deep, P. Koutrisw, and Y. Teng, "Revenue maximization for query pricing," *Proc. VLDB Endowment*, vol. 13, no. 1, pp. 1–14, Sep. 2019.

[29] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," *Proc. VLDB Endowment*, vol. 4, no. 12, pp. 1482–1485, 2011.

[30] O. A. Alzubi, J. A. Alzubi, M. Alweshah, I. Qiqieh, S. Al-Shami, and M. Ramachandran, "An optimal pruning algorithm of classifier ensembles: Dynamic programming approach," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 16091–16107, Oct. 2020.

[31] L. S. Shapley, "17. A value for n-person games," in *Contributions to the Theory of Games (AM-28)*, vol. 2, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 2016, pp. 307–318, doi: 10.1515/9781400881970-018.

[32] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *J. Comput. Oper. Res.*, vol. 36, no. 5, pp. 1726–1730, 2009.
[33] M. Ancona, C. Oztireli, and M. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley value approximation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 272–281.
[34] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," 2019, *arXiv:1905.10497*.
[35] A. Ghorbani, M. Kim, and J. Zou, "A distributional framework for data valuation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3535–3544.
[36] T. P. Michalak, K. V. Aadithya, P. L. Szczepanski, B. Ravindran, and N. R. Jennings, "Efficient computation of the Shapley value for game-theoretic network centrality," *J. Artif. Intell. Res.*, vol. 46, pp. 607–650, Apr. 2013.
[37] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," *ACM Trans. Database Syst.*, vol. 60, no. 12, pp. 79–86, 2014.
[38] A. Alberini, "Efficiency vs bias of willingness-to-pay estimates: Bivariate and interval-data models," *J. Environ. Econ. Manage.*, vol. 29, no. 2, pp. 169–180, Sep. 1995.
[39] O. L. Mangasarian and W. H. Wolberg, "Cancer diagnosis via linear programming," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep., 1990.
[40] J. Peetre, "Concave majorants of positive functions," *Acta Mathematica Academiae Scientiarum Hungaricae*, vol. 21, nos. 3–4, pp. 327–333, Sep. 1970.
[41] M. R. Kummara, B. R. Guntreddy, I. G. Vega, and Y. H. Tai, "Dynamic pricing of ancillaries using machine learning: One step closer to full offer optimization," *J. Revenue Pricing Manage.*, vol. 20, pp. 646–653, Jun. 2021, doi: 10.1057/s41272-021-00347-6.
[42] J. O. Kephart, J. E. Hanson, and A. R. Greenwald, "Dynamic pricing by software agents," *Comput. Netw.*, vol. 32, no. 6, pp. 731–752, May 2000.

YINGJIE TIAN received the bachelor's degree in mathematics from Shandong Normal University, Jinan, China, in 1994, the master's degree in applied mathematics from the Beijing Institute of Technology, Beijing, China, in 1997, and the Ph.D. degree in management science and engineering from China Agricultural University, Beijing.

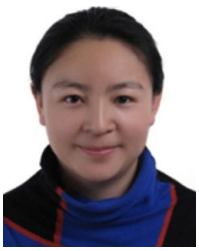
He is currently a Professor with the Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences. His research interests include artificial intelligent, machine learning, and optimization.

YURONG DING received the bachelor's degree in automation from the Beijing Institute of Technology, Beijing, China, in 2017. She is currently pursuing the master's degree in management science and engineering with the University of Chinese Academy of Sciences.

Her research interests include data value, data pricing, and data science.

SAIJI FU received the bachelor's degree in information management and information system from the University of Science and Technology, Beijing, China, in 2017, the master's degree in financial engineering from the University of Chinese Academy of Sciences, Beijing, in 2019, where she is currently pursuing the Ph.D. degree in management science and engineering.

Her research interests include machine learning, data mining, and medical image analysis.



DALIAN LIU received the B.S. degree in mathematical education from Hebei Normal University, Shijiazhuang, China, in 2001, the M.S. degree from Xidian University, Xi'an, China, in 2004, and the Ph.D. degree in computer science and technology from Beijing Jiaotong University, Beijing, China, in 2017.

She is currently an Associate Professor with Beijing Union University. Her current research interests include optimization and data mining.