DataPrice: An Interactive System for Pricing Datasets in Data Marketplaces

Yiding Zhu
Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, Zhejiang University
zhuyd@zju.edu.cn

Hongwei Zhang
Zhejiang University
Hangzhou, China
hongweizhang@zju.edu.cn

Jiayao Zhang
Zhejiang University
Hangzhou, China
jiayaozhang@zju.edu.cn

Jinfei Liu*
Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, Zhejiang University
jinfeiliu@zju.edu.cn

Kui Ren
Zhejiang University
Hangzhou, China
kuiren@zju.edu.cn

ABSTRACT

With the flourishing of data-driven applications, data marketplaces, which can dramatically facilitate data utilization, have emerged recently. However, determining the appropriate price for datasets presents a significant challenge due to the intangible nature of data. *DataPrice* alleviates the challenge by providing an interactive system based on a pricing model trained on real datasets collected from commercial data marketplaces. The pricing model can estimate an appropriate price according to the dataset description as a reference for both data sellers and data buyers. By leveraging Shapley values to evaluate the contribution of each attribute in metadata on the estimated price, *DataPrice* offers an explanation of the pricing process in a text form to enhance user trust.

PVLDB Reference Format:
Yiding Zhu, Hongwei Zhang, Jiayao Zhang, Jinfei Liu, and Kui Ren. DataPrice: An Interactive System for Pricing Datasets in Data Marketplaces. PVLDB, 17(12): 4433 - 4436, 2024.
doi:10.14778/3685800.3685893

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/ZJU-DIVER/DataPrice.

1 INTRODUCTION

Data is becoming increasingly important in large-scale data-driven services, with many people referring to it as the "new oil". The growing demand for data has given rise to many commercial data marketplaces such as AWS Data Exchange [1] and DataRade [3]. These marketplaces act as brokers, connecting data sellers who supply data with data buyers who need it. They provide technical support and services, ensuring transparency, efficiency, and trust in data transactions. However, pricing data is still a significant challenge. Both data sellers and buyers are concerned with the proper price of data to make informed market decisions.

Unlike traditional physical products, data is intangible and can be duplicated, transported, and accessed at almost zero cost. These distinctive characteristics of data make pricing mechanisms designed for physical goods ineffective in pricing data [10]. Therefore, a series of data pricing methods have been proposed [5–8]. Some infer data prices based on given price points to satisfy requirements like arbitrage freeness and prevention of double charging [5, 7], while others evaluate data prices according to the fundamental properties such as usability, integrity, and accuracy [6, 8]. However, the application and effectiveness of these approaches in real-world scenarios remain largely unexplored and untested. To bridge this gap, Azcoitia et al. [4] recently conducted a comprehensive study on commercial data marketplaces. Nevertheless, their research only considers data marketplaces in the United States and Europe and lacks a practical model for price prediction. We address these gaps and demonstrate *DataPrice*, an interactive system for pricing datasets in data marketplaces.

DataPrice develops an innovative system for participants in data marketplaces to determine the appropriate price of a dataset based on metadata that is considered critical in influencing data prices. *DataPrice* first extracts metadata from user-entered descriptions of datasets, including data category and sentence embedding. Users can further adjust and enrich the metadata to better reflect the characteristics of datasets. Based on the collected metadata, *DataPrice* uses a pricing model trained on real datasets to predict three pricing plans. Shapley value is employed to measure the contribution of each attribute in metadata to the predicted prices and generate textual explanations, which ensures transparency and understanding of the pricing process [9, 11, 12]. Using *DataPrice*, data sellers can receive suggested pricing plans, and data buyers can compare the listed price with the price generated by *DataPrice* to make decisions. *DataPrice* can also be used as a reference to measure the value of datasets, providing insight into how to improve data quality.

2 SYSTEM OVERVIEW

In this section, we introduce the architecture of *DataPrice*. An overview of *DataPrice* is shown in Figure 1. *DataPrice* consists of

*Corresponding author.

Proceedings of the VLDB Endowment, Vol. 17, No. 12 ISSN 2150-8097.
doi:10.14778/3685800.3685893

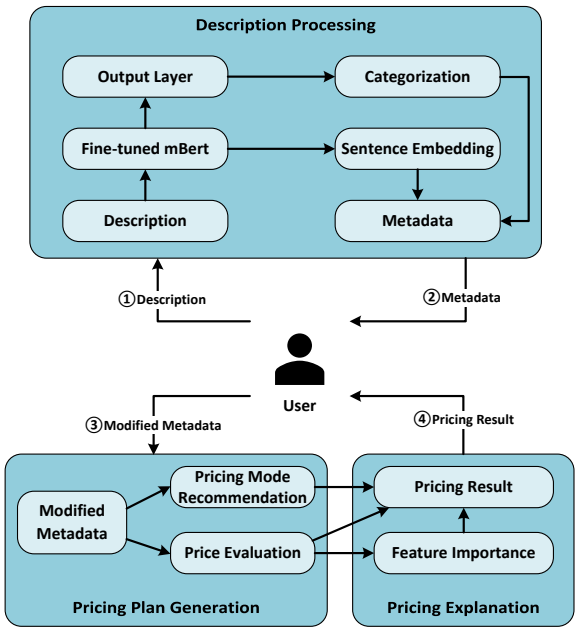

Figure 1: The Architecture of *DataPrice*.

three components: description processing, pricing plan generation, and pricing explanation. The description processing component extracts information from user-input dataset descriptions (Section 2.1). Based on the information and other user-input attributes, the pricing plan generation component predicts dataset pricing plans and estimates corresponding recommendation scores (Section 2.2). The pricing explanation component explains the reasons for the resulting prices (Section 2.3). These components combine to provide a user-friendly system for determining the price of a dataset based on its metadata and elucidate the explanation behind the valuations.

2.1 Description Processing

DataPrice first processes the description input by users in natural language regarding the dataset to be valued. The dataset description contains various information, including content details, data source, and format. The description processing component extracts the metadata that captures the key characteristics of the dataset from the description for subsequent pricing plan generation. This component employs mBERT [2], a multilingual BERT model pre-trained on a diverse corpus that includes texts from multiple languages. In the fine-tuning phase, mBERT is trained for the task of dataset description analysis. This specialized training enables the model to identify key terms that indicate the category of the dataset such as "daily transactions" and "climate change". This component consists of the following two modules.

Sentence embedding. To obtain a high-dimensional representation of the dataset description, this module extracts word embeddings from the final layer of mBERT and averages these word embeddings to get a sentence embedding. The full descriptive context is effectively condensed into a vector form. This vector captures syntactic and semantic nuances, containing a distilled summary of the information. It is used as an input to the underlying pricing model along with other metadata, which is an important basis for pricing.

Dataset categorization. This module categorizes datasets based on the embeddings from mBERT. A linear layer is integrated into mBERT as a classifier adept at dataset categorization. The classifier can learn from the subtleties within embeddings, allowing it to differentiate between different types of datasets. For example, it can accurately classify a "comprehensive report of market trends" under the "Financial" category and an "annual summary of climate change" under the "Environmental" category.

2.2 Pricing Plan Generation

In commercial data marketplaces, three data pricing modes are observed [1, 3]: **subscription** (data buyers are charged a recurring fee on a monthly or annual basis), **one-off payment** (data buyers pay a one-time fee once to access the data permanently), and **volume-based payment** (data buyers are charged based on the amount of data they consume such as the volume of data downloaded or the number of API calls). Leveraging the metadata obtained from the description processing component and additional input provided by the user, the pricing plan generation component estimates appropriate prices of datasets for the three established pricing modes and calculates the respective recommendation scores. It consists of the following two modules.

Price evaluation. A pricing model is utilized to predict the price of a dataset in three pricing modes: subscription, one-off payment, and volume-based payment. The pricing model receives as input the metadata generated in the description processing component and additional attributes provided by the user (including the dataset size, how frequently it is updated, and the geographic scope of the dataset). According to the research by Azcoitia et al. [4], these attributes have the greatest impact on the value of the dataset. For example, a large dataset that is updated regularly and contains consumer behavior information for a specific market has a higher value than a smaller, static dataset that contains common public information. The pricing model is trained on real datasets on data pricing and data transactions collected from commercial data marketplaces. After tuning hyperparameters and comparing the performance of several machine learning models, the random forest model shows the best results. Therefore, the random forest model is employed as the pricing model.

Pricing mode recommendation. This module computes the recommendation score for each pricing mode, guiding users to the most suitable option. The recommendation score $Score_i$ for pricing mode i is calculated as $Score_i = \frac{P_i}{K}$, where K is the total number of the closest data points considered, and P_i is the count of pricing modes i in K closest data points. We utilize the K-Nearest Neighbors algorithm to identify the K closest data points in our training dataset, indicating that they possess similar pricing scenarios to the dataset requiring pricing. For example, if three pricing modes occur 30, 10, and 10 times out of the 50 nearest data points in our training dataset, respectively, the recommendation scores are 0.6, 0.2, and 0.2, respectively.

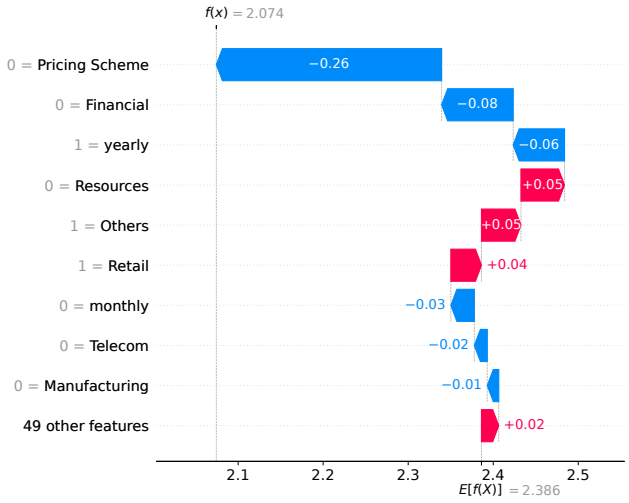

Figure 2: Use Shapley values to explain how each attribute contributes to pushing the model output from the base value $E[f(x)]$ to the actual output $f(x)$.

2.3 Pricing Explanation

Users may have inquiries about the rationale behind the proposed pricing plans. To tackle the issue, *DataPrice* provides explanations for each pricing plan. Through textual descriptions, it offers users insights into how the attributes in metadata contribute to the price, enhancing understanding and trust in the valuation process.

Shapley value [11], a well-celebrated notion in game theory proposed by Lloyd Shapley, provides a way to fairly measure the individual contributions to a collective outcome. Given a pricing model f and metadata input $x = (x_1, \ldots, x_d)$ with d features, the utility of an attribute subset $S \subseteq \{x_1, \ldots, x_d\}$ is defined as the predicted price of model on S, denoted by $f(S) \in \mathbb{R}$. The Shapley value $SV_i \in \mathbb{R}$ for the attribute $x_i \in \{x_1, \ldots, x_d\}$ is given by

$$SV_i = \frac{1}{d} \sum_{S \subseteq \{x_1,\ldots,x_d\} \setminus \{x_i\}} \frac{f(S \cup \{x_i\}) - f(S)}{\binom{d-1}{|S|}}.$$

The pricing explanation component utilizes Shapley values to evaluate the impact of metadata on pricing. Figure 2 provides an example of how metadata influences the data price. Using the volume-based pricing model and updating yearly are the primary factors that lower the price. The dataset belongs to the "Others" and "Retail" categories but not the "Financial" category, so the price is relatively lower. According to the computed Shapley values, the system uses a predefined text template to generate explanations, highlighting which attributes increase or decrease the price.

3 SYSTEM DEMONSTRATION

In this section, we demonstrate *DataPrice*. Figure 3 is a screenshot of *DataPrice*. The user can observe three panels including "Descriptio", "Metadata", and "Pricing Plans". The user can use *DataPrice* to price a dataset in the following steps.

Step 1 (Description processing). The user first needs to enter a description of the dataset they want to price in the text box within the "Description" panel (see Figure 3-1). After clicking the "GENERATE" button, *DataPrice* analyzes the information from the entered description and predicts the category of the dataset. The predicted category is then displayed in the "Metadata" panel (see Figure 3-2) for further adjustment or confirmation by the user.

DataPrice serves both data sellers and buyers in the data marketplaces. Sellers should provide a comprehensive description of their datasets when using *DataPrice* to determine the prices. This allows *DataPrice* to obtain the accurate and necessary metadata to assess the value of datasets. Buyers can directly copy the description of the dataset into *DataPrice* to check whether its selling price is reasonable. In addition, users can also use *DataPrice* to investigate prices of datasets with certain characteristics. By providing a dataset description in natural language, such as "A dataset that contains information about the top-ranking restaurants in the United States...", *DataPrice* will automatically extract key metadata.

Step 2 (Metadata adjustment). The user can check and adjust the generated metadata in the "Metadata" panel. Each dataset is assigned a main category while being associated with various tags to collectively represent its finer classification. Both "Category" and "Tags" labels draw from well-known data marketplaces such as AWS Data Exchange [1] and DataRade [3]. They are then integrated into unified terms to avoid redundancy. For example, similar concepts of "Health" and "Healthcare" are labeled as "Health". The User can select multiple relevant tags from a dropdown menu "Tag", which dynamically updates to display tags associated with the selected category.

Moreover, data volume, freshness, and coverage are also significant factors in determining the price of datasets. To make the pricing more reliable, the user is encouraged to input this supplementary information. When a data sample is available, the user can upload it using the "UPLOAD" button. *DataPrice* will automatically generate a data dictionary based on the data sample, which is used in estimating the price. For data coverage, besides the select box, *DataPrice* provides the user with an interactive world map. On this map, the user can mark the countries or regions where the data is collected. The user can also quickly select all countries or regions on a continent by checking the checkbox below the select box. Selected areas will be highlighted on the map, giving the user a visualization of the geographical scope.

The collected metadata in *DataPrice* plays an essential role in determining the price of a dataset. Metadata may be incomplete caused by a lack of available data samples or unknown data coverage. However, *DataPrice* can still complete dataset pricing and offer a relatively wide range of prices, ensuring the user has a basic understanding of the value of the dataset.

Step 3 (Pricing plans). Upon finishing the metadata adjustment in the "Metadata" panel, the user can click on the "SUBMIT" button to send metadata to the backend of *DataPrice* for generating the pricing results. The pricing results are then listed in the "Pricing Plans" panel (see Figure 3-3). It contains three different pricing plans: subscription, one-off payment, and volume-based payment. Each pricing plan has a "Fee" that shows price information, a "Reason" that provides insights into the pricing rationale, and a "Score" that indicates the recommendation score of the plan. For subscription-based pricing plans, the pricing information is displayed in the

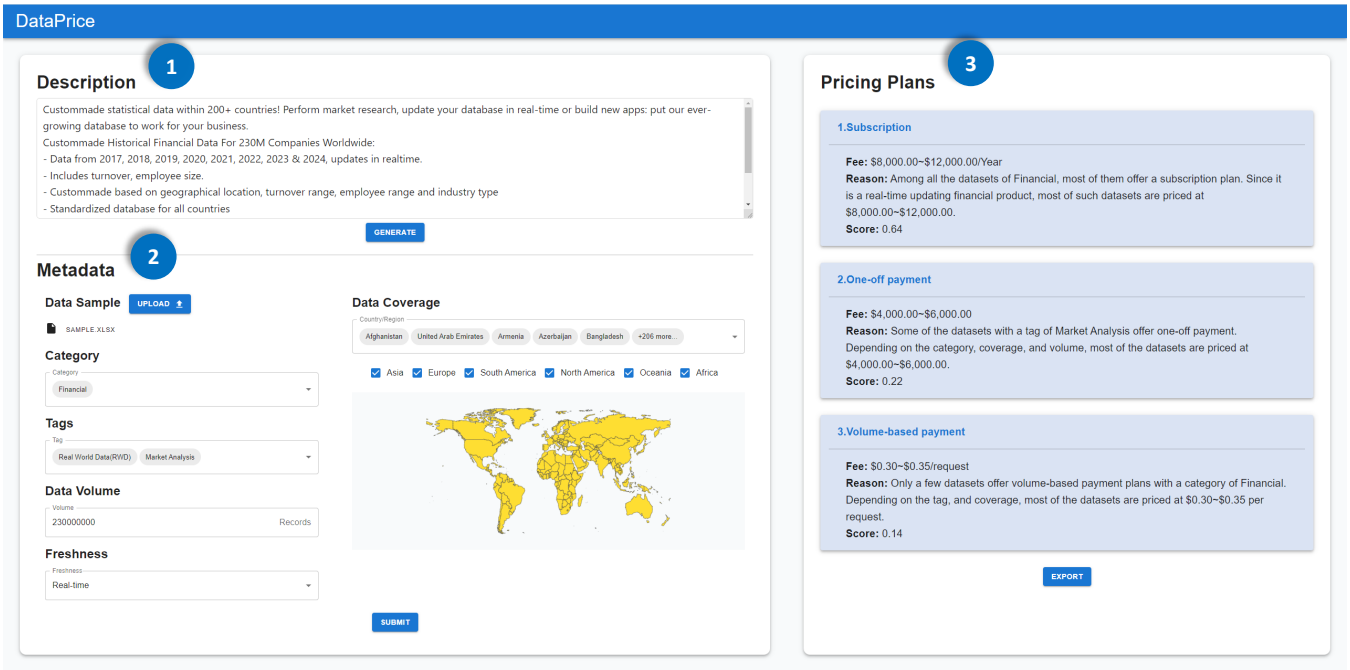

Figure 3: A screenshot of *DataPrice*.

form of a range of subscription fees and their payment cycle (e.g. "$500~$1000/month"). A one-off payment plan directly shows its price, while a volume-based payment plan displays a per-request charge.

To assist users in selecting pricing plans, *DataPrice* displays recommendation scores that signify the suitability of the plans. A higher recommendation score suggests that the pricing mode is the most commonly utilized within the historical data most similar to the current data.

In addition to providing recommendation scores, *DataPrice* enhances user understanding by presenting detailed explanations. These explanations delve into how the attributes of the metadata influence prices, such as categorization, volume, and update frequency. Thus, *DataPrice* clarifies the pricing mechanics, making it clear why a particular plan is the most suited to the user needs based on historical data.

DataPrice also provides an "EXPORT" button that allows the user to download the pricing plans along with the inputs. This file can serve as a reference for the actual data marketplaces.

4 CONCLUSION

In this paper, we demonstrated *DataPrice*, an interactive and user-friendly system for pricing datasets. We introduced the system pipelines, implementation details, and user scenarios of *DataPrice*. *DataPrice* only requires users to input several lines of dataset description and some key metadata. Based on this, *DataPrice* generates three different pricing plans along with explanation and recommendation scores. Both data sellers and buyers can benefit from utilizing *DataPrice* to make decisions in data marketplaces. *DataPrice* is in its prototype stage, and there are many ways for improvements. These enhancements include training the pricing model on a larger scale of the dataset to improve pricing accuracy, and applying a large language model to the pricing explanation module for more natural expressions.

ACKNOWLEDGMENT

This work was supported in part by the National Key RD Program of China (2021YFB3101100) and NSFC grants (62102352, U23A20306).

REFERENCES

[1] 2024. AWS Data Exchange. https://aws.amazon.com/data-exchange.
[2] 2024. BERT Base Multilingual Cased. https://huggingface.co/bert-base-multilingual-cased.
[3] 2024. Datarade. https://datarade.ai/.
[4] Santiago Andrés Azcoitia, Costas Iordanou, and Nikolaos Laoutaris. 2023. Understanding the Price of Data in Commercial Data Marketplaces. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 3718–3728.
[5] Shaleen Deep and Paraschos Koutris. 2017. QIRANA: A framework for scalable query pricing. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 699–713.
[6] Judd Randolph Heckman, Erin Laurel Boehmer, Elizabeth Hope Peters, Milad Davaloo, and Nikhil Gopinath Kurup. 2015. A pricing model for data markets. *iConference 2015 Proceedings* (2015).
[7] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. 2015. Query-based data pricing. *Journal of the ACM (JACM)* 62, 5 (2015), 1–44.
[8] Jinfei Liu, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. 2021. Dealer: an end-to-end model marketplace with differential privacy. *Proceedings of the VLDB Endowment* 14, 6 (2021).
[9] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
[10] Jian Pei. 2020. A survey on data pricing: from economics to data science. *IEEE Transactions on knowledge and Data Engineering* 34, 10 (2020), 4586–4608.
[11] Lloyd S Shapley et al. 1953. A value for n-person games. (1953).
[12] Jiayao Zhang, Qiheng Sun, Jinfei Liu, Li Xiong, Jian Pei, and Kui Ren. 2023. Efficient Sampling Approaches to Shapley Value Approximation. *Proc. ACM Manag. Data* 1, 1 (2023), 48:1–48:24. https://doi.org/10.1145/3588728