An Interpretable Market-based Data Price Prediction Tool

Santiago Andrés Azcoitia
Universidad Politécnica de Madrid
Madrid, Spain
santiago.andres@upm.es

Alicia Cabrero Jiménez
Universidad Politécnica de Madrid
Madrid, Spain
alicia.cjimenez@alumnos.upm.es

ABSTRACT

Artificial intelligence (AI) and machine learning (ML) are having a profound impact on the economy but require huge amounts of data, which is partially generated by increasingly digitalised organisations but often acquired from third parties. This has resulted in a rampant demand for data in emerging data markets that face daunting challenges derived from the nature of data as an economic good (freely replicable, non-rival) and its elusive value. Despite the appearance of data marketplaces (AWS, Snowflake, Nokia DM) aimed to facilitate data transactions, data holders find it difficult to set a price for their data assets, and data consumers have trouble estimating a fair price to pay for data.

This paper presents an interpretable market-based data pricing tool designed to help with these tasks by estimating the price of data based on the prices of commercial data products observed in data marketplaces. Resorting to sentence transformers, neural networks, sensitivity analysis, and novel two-step SHapley Additive exPlanations (SHAP), not only does our tool provide insightful user-friendly reporting and interpretation of price predictions using different price schemes, but it also improves the accuracy, the robustness, and the generalisability of state-of-the-art (SOTA) models.

PVLDB Workshop Format:
Santiago Andrés Azcoitia and Alicia Cabrero Jiménez. An Interpretable Market-based Data Price Prediction Tool. VLDB 2025 Workshop: 3rd Data EConomy Workshop (DEC).

1 INTRODUCTION

AI/ML will have a huge impact on the economy and will transform labour markets and many other aspects of our daily lives [9, 11, 12, 21]. Economists trust that this technology will act as a tailwind to propel productivity and economic growth in the next decades by addressing critical challenges such as climate change or the ageing workforce [19, 30]. For this, AI/ML requires huge amounts of data.

However, data assets are often owned and controlled by third parties who are reluctant to share them. This is due to the elusive nature of information assets–data is freely replicable, reusable, etc. [13]–, the fear of inadvertently sharing a competitive advantage with potential competitors and the risk of fines for accidentally disclosing sensitive personal information. As a result, data-driven companies tend to integrate horizontally throughout the value chain, amassing and fiercely protecting critical data to provide digital services to their customers [37], and data exchanges rely on *ad hoc* sharing agreements with selected partners, even in kind.

Enabling effective data markets is key to making the most of AI/ML. Data transactions will require money to be exchanged, which leads to the problem of agreeing on a price. However, data providers find it difficult to set the price of data assets, and data consumers struggle to anticipate the cost of acquiring data in the market. As a result, failure to agree on a price has become the key reason why data transactions are falling apart nowadays [23]. In addition, regulatory and competition bodies are committed to increasing the transparency of data markets to properly tax digital services, including data transfers [7, 40, 43], and to promote trust and ensure fairness in the data-driven economy [33].

Data pricing has long been an active research topic at the intersection of economics and computer science. Different schools have approached the problem from various perspectives such as data auctions, the value of privacy, the utility brought to specific AI/ML models, the quality of data assets or the amount of information in database queries [34]. Data marketplaces (DM) have appeared to mediate data transactions [39]. In addition to helping with data discovery and swift data delivery, DMs can play a role in price negotiations. As data markets grow, pricing based on market references is becoming a convenient solution to this problem.

Following the approach of quality-based pricing, previous works developed ML models to classify data products and predict their prices based on their metadata [3, 47]. However, they showed limitations in how they process inputs, the pricing schemes supported, and their interpretability by their users.

In this paper, we present a tool that predicts the price of a data product based on those observed in commercial DMs. Not only do we present tailored neural network (NN) classifiers and regressors that outperform pre-existing models, but we also provide user-friendly interpretations of their predictions. The pricing tool uses this functionality to assist providers in setting the price of their data assets under different schemes (volume-based, subscription-based, or one-off download price), and data consumers in estimating the cost of acquiring data in the market. *Our Contributions* include:

- Developing a generic data pricing tool to help data holders and users agree on the price of data products.
- Designing and optimising more generalisable data product classifiers and regressors based on transformers.
- Improving the accuracy of existing classifiers (min. +0.1 F1 score) and price regressors (-15-40% log MAE).
- Augmenting data to increase robustness (-60% MAE in pricing data products with equivalent descriptions).
- Developing a methodology to support a variety of pricing schemes often found in DMs.
- Applying interpretable AI techniques to fully understand features and keywords affecting the price of data products.

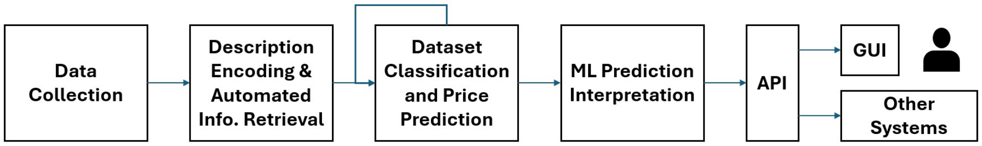
Data Collection
Description Encoding & Automated Info. Retrieval
Dataset Classification and Price Prediction
ML Prediction Interpretation
API
GUI
Other Systems

Figure 1: Architecture and Methodology.

The remainder of the paper is structured as follows. In Sect. 2, we introduce the architecture of the tool, and explain key concepts used throughout the paper. Section 3 then compares the performance of the models with the state of the art, and derives features and feature groups that influence the classification of a data product and its price. Section 4 shows how we improved the robustness of prediction models through data augmentation. Finally, Sect. 5 discusses related work and Sect. 6 concludes and points to future research on the topic.

2 ARCHITECTURE AND METHODOLOGY

Figure 1 summarises the architecture and methodology of our interpretable market-based data price prediction tool. They address the limitations of pioneering work on the topic [3, 47] namely by using more generalisable processing of data product descriptions that allows multiple languages, by supporting different pricing schemes, and by providing human-interpretable price predictions.

First, a flexible automated crawler allows *scraping* information about data products sold by commercial DMs. We follow common scraping good practices, such as avoiding repeated visits to the same product, setting up random wait times from 30 seconds to 1 minute after requesting a web page from a server in order to avoid flooding it with requests. Section 2.1 presents the dataset used.

Second, we run open-source large language models (LLMs) locally to automate information retrieval and transform HTML scraped data into structured metadata feature matrices. As shown in Sect. 2.2, we test different pre-trained sentence transformers, some of them supporting multiple languages, to encode descriptions. This solves the limitations of previous ML models based on bag of words (BoW), term frequency (TF) and inverse document frequency (IDF).

Taking description embeddings and relevant metadata features of commercial data products as input, we test and optimise ML models to classify data products and enrich their metadata and to feed price regressors that predict the logarithm of their monthly subscription price based on market references. Section 2.3 presents the analysis we made to design and optimise our models. We reuse existing datasets to compare to the state of the art [5].

One of the key aspects of the development of ML models is transparency and interpretability. To this end, Sect. 2.4 discusses a novel two-step SHAP analysis that returns the importance of metadata and description tokens in determining the price of a data product or in classifying such product in a certain category.

Moreover, we develop a methodology to deliver price predictions for the most usual pricing schemes, namely one-off pricing, subscription-based pricing and volume-based pricing in Sect. 2.5. Finally, Sect. 2.6 shows how data augmentation helped to increase the robustness of price predictions.

2.1 Datasets

We trained classifiers for standard data categories of AWS for which we have sufficient data product descriptions. We followed the methodology presented in previous works on this topic to gather information on data products and their classifications in this DM [3], a summary of which is shown in Tab. 1.

Table 1: Summary of Data used to train classifiers

Category	0	1	Total	Category	0	1	Total
Financial	8,817	4,260	13,077	**Healthcare**	11,199	1,506	12,705
M&E	11,694	1,007	12,701	**Telecom**	12,013	907	12,920
Gaming	12,602	91	12,693	**Automotive**	12,292	461	12,753
Manufacturing	12,292	645	12,937	**Resources**	11,443	13,56	12,799
Retail	8,424	5,616	14,041	**Public Sector**	10,687	20,05	12,692
Others	11,029	1,695	12,724				

To train data product price regressors, we used a dataset of 8,379 price references of 4,103 products obtained from commercial marketplaces in 2021–2022 [5]. It contains product descriptions in English, asking prices and relevant metadata features that appear to drive them, such as time and geographical scope, volume, format, delivery methods, limitations, add-ons, etc [3].

2.2 Encoding descriptions

Previous approaches to predicting the price of data relied on BoW and TF-IDF techniques that ignore the meaning of descriptions. This solution is not generalisable to descriptions typed in by users, which do not necessarily resemble the ones used in the training.

To circumvent this shortcoming, we used pre-trained text transformers [41] and sentence BERT [35] to encode data product descriptions into sentence embeddings that carry semantic information. To evaluate the new approach, we replaced the word features of existing datasets with BERT encodings and maintained the rest of the metadata before training the models in $\mathbf{X} = (\mathbf{X}_{\mathbf{meta}} \mid \mathbf{X}_{\mathbf{enc}})$. We tested different pre-trained transformers (all-mpnet-base-v2 [48], multi-qa-mpnet-base-dot-v1, all-distilroberta-v1 [36], multi-qa-distilbert-cos-v1, and paraphrase-multilingual-mpnet-base-v2) on our classifiers and regressors with 10 different 80/20 train/test splits.

2.3 Designing and optimising models

We developed classifiers and regressors based on neural networks using TensorFlow [1] and Keras [24] and optimised them to maximise their performance. As a result, these models did perfectly interpret sentence transformer embeddings, and outperformed SOTA regression models like Random Forest, Gradient Boosting, XGBoost, kNeighbors, lightGBM and CatBoost regressors.

When designing and testing the models, we followed all the recommended good practices by first standardising the input data

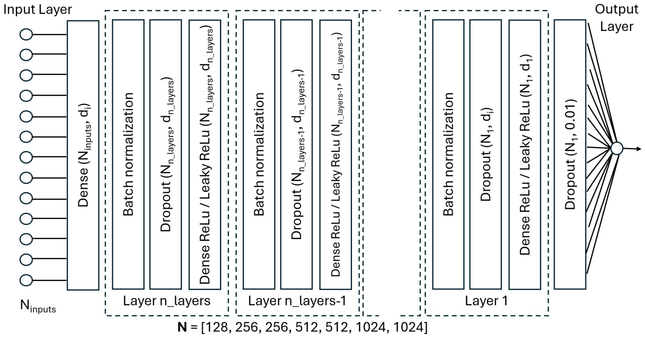
Input Layer
Dense (N_{inputs}, d_i)
Batch normalization
Dropout (N_{n_layers}, d_{n_layers})
Dense ReLu / Leaky ReLu (N_{n_layers}, d_{n_layers})
Batch normalization
Dropout ($N_{n\_layers-1}$, $d_{n\_layers-1}$)
Dense ReLu / Leaky ReLu ($N_{n\_layers-1}$, $d_{n\_layers-1}$)
Batch normalization
Dropout (N_1, d_i)
Dense ReLu / Leaky ReLu (N_1, d_1)
Dropout (N_1, 0.01)
Output Layer
N_{inputs}
Layer n_layers
Layer n_layers-1
Layer 1
N = [128, 256, 256, 512, 512, 1024, 1024]

Figure 2: Neural network architecture.

and removing highly correlated features. Figure 2 shows the architecture of the neural networks we used for classifiers and regressors [29]. We used linear and sigmoid activation functions for the output layer of regressors and classifiers, respectively. As a loss function, we used mean absolute error (MAE) for the regressors and binary cross-entropy for the classifiers. To avoid overfitting, we randomly applied dropout between training epochs and to avoid dying/exploding neurons, we also applied batch normalisation between all layers. We used the Adam optimiser [25] with a tuned learning rate decay to train the model faster at the beginning and then decrease the learning rate with further epochs to make training more precise. Finally, we used callbacks to stop the training at the optimal epoch.

We optimised the number of hidden layers and training epochs considering the trade-off between accuracy and training time, we tested rectified linear unit (ReLU) and leaky ReLU activation functions for hidden layers, different dropout percentages, and different batch sizes and learning decay profiles to accelerate the learning time and optimise the results. Figure 3 shows the accuracy achieved and the training time (average of 10 different runs with different 80/20 train test splits of the input data) for a range of values of these parameters and highlights those chosen for the classifier.

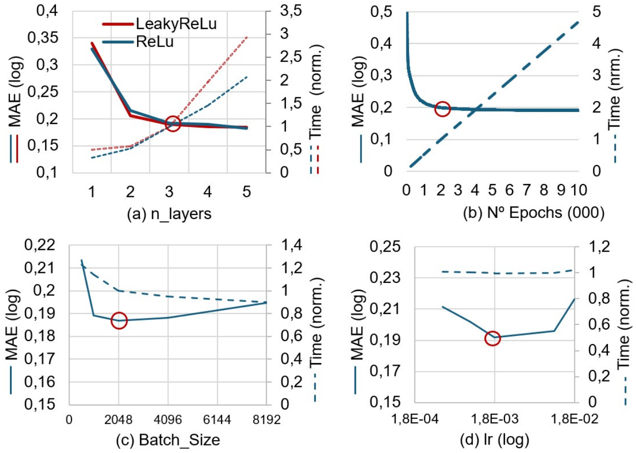

Figure 3: Parameter tuning

2.4 Interpreting predictions

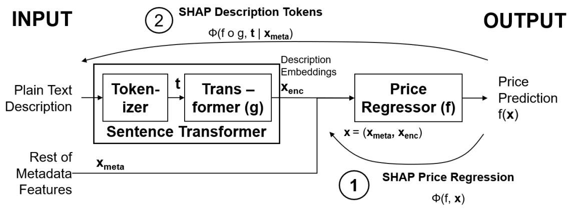

Figure 4: Two-step SHAP analysis for better interpretability

We used SHAP to provide the user with an interpretation of price predictions at the level of feature and feature group [28]. As input, we need the metadata of the data product and the encoding of its description: $\mathbf{x} = (\mathbf{x}_{\text{meta}}, \mathbf{x}_{\text{enc}})$. A direct calculation of the SHAP values returns $\phi(f, x_i), x_i \in \mathbf{x}$. Since the model is trained to produce the logarithm of the price, the SHAP values, which add to $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$, can be interpreted as price multipliers to move from the predicted base price if no information on the data product was known $p_{\text{base}} = 10^{\mathbb{E}[f(\mathbf{X})]}$ to the actual prediction $p_{\text{pred}} = 10^{f(\mathbf{x})}$.

$$p_{\text{pred}} = p_{\text{base}} \cdot \prod_{x_i \in \mathbf{x}} 10^{\phi(f, x_i)} \tag{1}$$

However, SHAP values of description encodings are difficult to interpret because users cannot connect them to specific words in descriptions. Carrying out a SHAP analysis on both the model and the sentence transformer $(f \circ g)$ would significantly increase the dimensionality of the problem, leading to very long processing times. Therefore, we implemented a two-step SHAP, as follows:
1. We aggregate SHAP values of description encodings to calculate the aggregate importance of the description in the prediction.

$$\phi(f, \mathbf{x}_{\text{enc}}) = \sum_{x_i \in x_{\text{enc}}} \phi(f, x_i) \tag{2}$$

2. We estimate the SHAP values by description token in the entire model (encoding + price prediction) assuming that $\mathbf{x}_{\text{meta}}$ remains constant, that is, $\phi(f \circ g, \mathbf{t} | \mathbf{x}_{\text{meta}})$, or $\phi(f \circ g, \mathbf{t})$. Note that the sum of the token SHAP values does not necessarily match $\phi(f, \mathbf{x}_{\text{enc}})$.
3. We transform the token SHAP values to sum $\phi(f, \mathbf{x}_{\text{enc}})$, preserving the sign of $\phi(f \circ g, \mathbf{t})$ and the proportion between the tokens driving the prediction in the same direction, and minimising the mean square difference between ϕ' and ϕ, as follows:

$$\phi'(f \circ g, \mathbf{t}) = \begin{cases} w^+ \cdot \phi(f \circ g, \mathbf{t}) & \phi(f \circ g, \mathbf{t}) \geq 0 \\ w^- \cdot \phi(f \circ g, \mathbf{t}) & \phi(f \circ g, \mathbf{t}) < 0 \end{cases} \tag{3}$$
$$argmin_{w^+, w^-} \sum_t \phi'(f \circ g, \mathbf{t})^2 - \phi(f \circ g, \mathbf{t})^2$$

subject to:

$$w^+, w^- \geq 0, \tag{4}$$
$$\sum_t \phi'(f \circ g, \mathbf{t}) = \phi(f, \mathbf{x}_{\text{enc}})$$

4. We replace the SHAP values of the encoding features $\phi(f, \mathbf{x}_{\text{enc}})$ with those of the tokens used in the description $\phi'(f \circ g, \mathbf{t})$.

Since they refer to specific words or stems, token SHAP values are easily interpretable by users. Furthermore, we maintain the desirable additivity of SHAP values, namely the sum of SHAP values of description tokens and metadata features is the gap between the base price $\mathbb{E}[f(\mathbf{X})]$ and the prediction we aim to explain $f(x)$.

$$f(x) - \mathbb{E}[f(\mathbf{X})] = \phi'(f \circ g, t) + \sum_{x_i \in \mathbf{x}_{\text{meta}}} \phi(f, x_i). \tag{5}$$

2.5 Converting price between pricing schemes

Even though our model predicts subscription prices (US$ / month), the dataset that feeds it includes price references that belong to different pricing schemes, and users may be interested in estimating one-off and volume-based prices. Therefore, converting between pricing schemes can be useful both to make more training data available and to produce outputs for different pricing schemes.

To do so, we make use of *consumer indifference*, a concept widely used in microeconomics. We say consumers are indifferent to two options of delivering and/or pricing a data product if both options provide the consumer with the same level of satisfaction. For example, a consumer who wants a snapshot of a dataset may indifferently download–and pay a one-off tariff p_{off} for–it from a provider's webpage, or subscribe to an AWS data product serving the same dataset for the minimum permissible time, typically one month, to download the information and then unsubscribe, paying the monthly tariff p_{sub} once. Generally, downloading a one-off data product worth p_{off} every T_u months is equivalent to subscribing to a data product worth p_{off}/T_u per month for period T_u.

Volume-based tariffs allow consumers to download a volume of data v over a certain period T_v paying a certain price p_v. This would be equivalent to paying for a subscription worth p_v/T_v per month for a period T_v to a dataset that contains the same data points v. However, APIs usually charge by volume and let users tailor the data they download from a wider database to their needs, as opposed to other methods that deliver the whole database that users query or trim locally. The model captures the premium for being able to select data through the API delivery feature.

2.6 Augmenting the training data

Users do not necessarily type descriptions that match those used to train the models, nor do they necessarily use the same words. In this paper, we use data augmentation techniques to improve the robustness of price predictions to small variations of the inputs, such as their descriptions or update rate. We show results that corroborate this improvement in Sect. 4.

To make the model more robust, we prompt LLMs to paraphrase descriptions while maintaining all the information they contain. Then we create new training data including the encoding of paraphrased descriptions and maintaining other metadata features.

To ensure predictions show a correct sensitivity to update rate, we again resort to the concept of consumer indifference. For example, data users interested in a product worth p that is updated after a period T_u can save money by downloading data less frequently. Consumers interested in refreshing the data every n periods would not need to pay more than p/n. This also holds for subscription-based products whose contract duration is T_u or shorter.

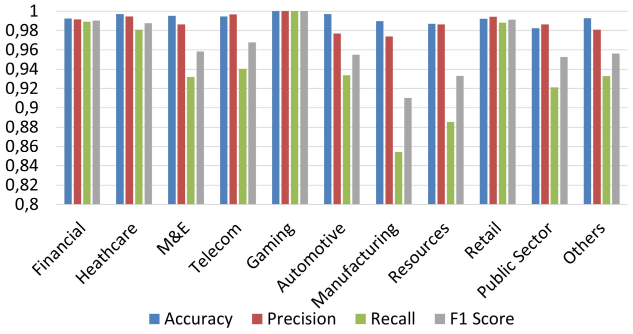

Figure 5: Performance of NN classifiers

Table 2: Comparison of results vs SOTA

	Accuracy	Precision	Recall	F_1 Score
Financial [3]	0.93	0.97	0.81	0.88
Financial (new)	0.99	0.98	0.98	0.98
Retail [3]	0.95	0.96	0.88	0.91
Retail (new)	0.97	0.97	0.96	0.96

3 EVALUATING NEURAL NETWORKS

In this section, we evaluate NN classifiers and regressors against SOTA datasets and models [3]. In addition to being more generalisable and hence more suitable for a price prediction tool, we show that the new classifiers and regressors outperform existing models.

3.1 Evaluating classifiers

Figure 5 summarises the average performance achieved in the test set by NN classifiers for ten different train/test splits of our dataset. Most of them achieve high accuracy and F_1 scores above 0.9. Heterogeneous lower-performing categories like 'Manufacturing', 'Public Sector' or 'Resources' even require more ground data to learn to better classify products. Compared to the results obtained in previous work, the new classifiers based on sentence transformers significantly improve key evaluation metrics, as shown in Tab. 2. These results are similar regardless of the sentence transformer used. We obtained negligible standard deviation (< 0.008) and maximum difference (0.02) between F1 scores across transformers.

To understand how the models work, we ran a SHAP analysis on the entire corpus to identify the most relevant tokens that classifiers rely on. We used TF-IDF statistics to weight the average SHAP values and consider the importance of the tokens in the training data. Table 3 presents the top ten tokens for some categories with and without TF-IDF corrections. These results show that algorithms generally use meaningful words to classify data products.

3.2 Evaluating price regressors

Regarding price prediction, we tested the NN model with ten different train/test data splits of the whole dataset, and then of data filtered by category (*Financial*, *Retail*, and *Healthcare*). In addition to other advantages they brought, the new models using sentence transformers significantly outperformed SOTA classifiers in all

Table 3: Top-10 relevant tokens for data product classifiers

Category	Most relevant tokens by average SHAP·TF·IDF	Most relevant tokens by average SHAP
Financial	financial, trading, investment, exports, companies, business, company, market, securities, asset	pricing, banking, import, valuation, financing, contracts, stocks, worth, deposit, merger
Healthcare	data, medical, patient, co, health, clinical, vid, disease, 19, patients	hospitals, iology, doctors, esthesia, dna, physicians, diagnosed, surgeons, medication, blood
Telecom	data, mobile, users, location, marketing, electronic, tr, network technology, market	bandwidth, telecom, phone, smartphone, network, route, connectivity, users, tower, trace
Gaming	gaming, nfl, player, sports, football, basketball, data, mobile, players, fantasy	gaming, football, nfl, player, sports, basketball, metric, fantasy, antiques, stats
Retail	intent, data, sentiment, consumer, location, companies, visitation, product, sales, locations	properties, lawyers, san, lawyers, dealers, smartphone, mosaic, purchase, retailer, venues

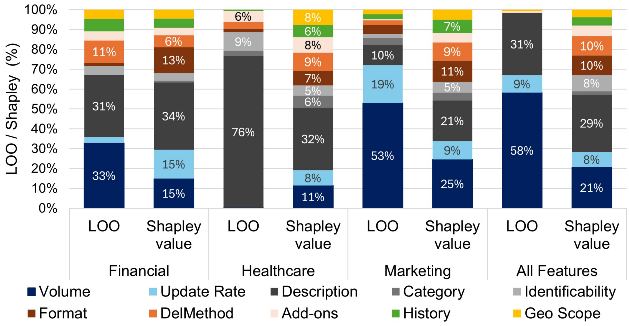

Figure 6: LOO and Shapley values (%) by feature group

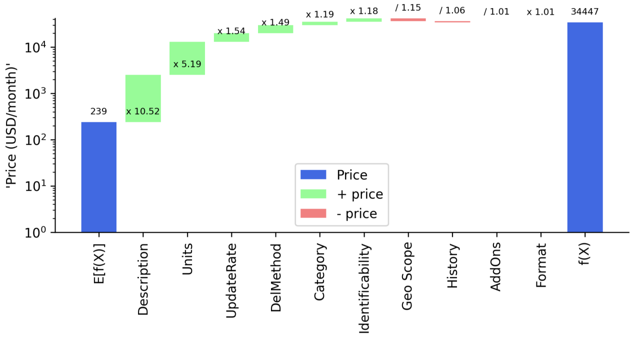

Figure 7: SHAP waterfall by feature group

categories, as shown in Tab. 4. According to these results, training a single general model covering all types of data products is preferable to training a specific model for each category.

We performed importance analysis to ensure that the predictions were based on meaningful metadata. Figure 6 plots the groups of features that turned out to be more meaningful to make predictions using two separate techniques to measure the impact of removing specific groups of features from the training data: leave-one-out (LOO), and Shapley [14, 38]. We reused the grouping in SOTA work [3], arranged together and coloured in similar tones feature groups that respond to similar questions about data products. Interestingly, it is *what* (in grey, descriptions, categories, and identificability) and *how much* (in blue, volume and update rate) data a product contains that determine almost 2/3 of its price. *How* data is delivered (in orange, format, delivery methods, and add-ons) appeared to be also important, whereas data time span, or history (in green, answering to *when*) and geographical scope (in yellow, answering to *where*) proved to be less relevant.

Furthermore, an interpretable data pricing tool must offer understandable explanations of its predictions. Using a two-step SHAP, the tool produces waterfall charts to explain why a prediction is above or below the average price observed in the training sample, both at the level of individual features and feature groups. Figure 7 shows a screenshot of a waterfall plot for a credit card transaction dataset. It clearly shows that the high price results from a combination of what the data is about (its description) and how much data it contains (its volume and update rate).

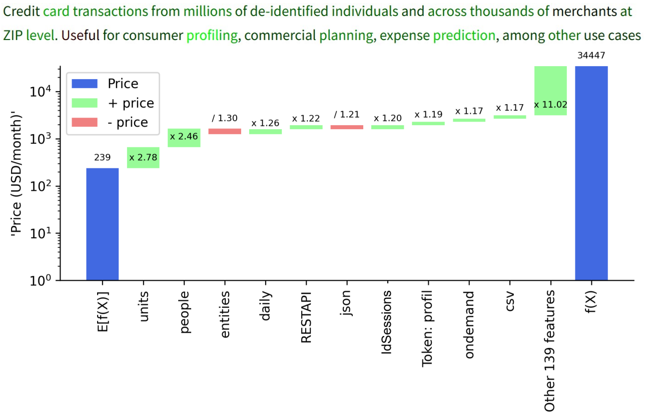
Credit card transactions from millions of de-identified individuals and across thousands of merchants at ZIP level. Useful for consumer profiling, commercial planning, expense prediction, among other use cases

Figure 8: SHAP waterfall by feature and token importance

An initial SHAP analysis at the feature level includes opaque encoding features that users cannot interpret. After a second step, the tool highlights the tokens and words in the product description driving the predicted price up (in green scale) and down (in red scale), and provides a customisable ranking of the most relevant features. Figure 8 shows screenshots of the result for a credit card transaction dataset, and clearly reflects that the fact that such data is "*de-identified*", concerns "*card transactions*" and allows "*consumer profiling*" is what makes it more valuable. In fact, the last concept appears in the waterfall plot at the feature level after this analysis.

Table 4: Comparison of regression models performance vs SOTA [3]

Model	Financial			Marketing			Healthcare			All		
	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE
RF	0.85	0.2	0.14	0.86	0.21	0.13	0.78	0.25	0.15	0.84	0.23	0.16
kN	0.78	0.31	0.26	0.74	0.33	0.24	0.77	0.26	0.17	0.69	0.37	0.31
GB	0.82	0.23	0.16	0.8	0.28	0.19	0.73	0.27	0.19	0.79	0.3	0.22
NN	0.73	0.33	0.35	0.77	0.30	0.22	0.68	0.26	0.18	0.72	0.33	0.28
new NN	0.89	0.17	0.10	0.89	0.17	0.10	0.82	0.19	0.12	0.91	0.15	0.08

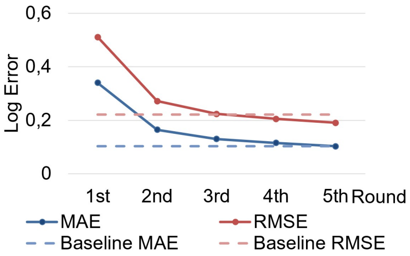

Figure 9: Price prediction error for paraphrased data products

Previous works focused on identifying the characteristics of commercial data products that drive their price in the market, while a tool that predicts such a price must respond to any input requested by their users, not necessarily matching those seen at training time. For example, equivalent descriptions of the same data product should lead to the same or similar prices. The price of data products that bring the same utility to end users should also be similar. Hence, a consumer interested in yearly updates of a dataset would likely pay 1/12 the price of another consumer interested in monthly updates of the same product (e.g., by subscribing and paying one month and then unsubscribing to the data service, or just paying for downloading the dataset once a year instead of every month).

Therefore, improving robustness becomes critical when training price prediction models. We resort to data-augmentation techniques to help with this.

4 IMPROVING ROBUSTNESS THROUGH DATA AUGMENTATION

To show how data augmentation can improve the model, we paraphrase the descriptions of the original data products without losing information. We then predict the price of the paraphrased input and measure the error, which should ideally be the same but turned out to be ×3 the baseline error obtained for the original data products. We add this new *paraphrased* dataset to the training data and retrain the model.

Figure 9 shows the prediction error (MAE and RMSE) measured in the new paraphrased version of the dataset after subsequent rounds of augmentation and retraining. After enriching–and multiplying—the training data five times, the model achieves the baseline accuracy in a sixth paraphrased version.

As a result of this process, the resulting model becomes more robust to small variations in descriptions. Future work will test data augmentation with descriptions in different languages, or changes in the desired update rate, which we just tested empirically so far.

5 RELATED WORKS

Data pricing has long been a relevant research line at the intersection of economics and computer science [34, 46]. In fact, the lack of empirical data on the price of data is considered a key challenge in DM surveys [16, 23, 34, 45]. Several different approaches have been proposed to price data assets. Some authors proposed auction designs applicable to data products [17, 18], others have defined pricing strategies and marketplaces based on differential privacy [15, 27] or queries to a database [8, 26]. Novel AI/ML DM architectures have been proposed under the concept of value-based pricing [2, 10, 32] and the value of privacy [31].

Quality-based pricing sets the price of data by weighting its quality features [20, 44]. Following this approach, previous work measured the price of data in commercial marketplaces, identified key metadata features that drive it, and suggested a design for a market-based price prediction tool [6]. Furthermore, a posterior work fit price regressors to understand the importance of those features but lacked the generalisability required for a price prediction tool [3]. This work, which we use as a baseline to compare our models, also published a dataset that we extend in this paper [5]. A previous data price prediction tool was developed on these data, which does not include a fully human-interpretable explanation of predictions, as we propose in this paper [47].

6 CONCLUSION AND FUTURE WORK

Our work proposes a novel market-based data pricing tool that combines fully-interpretable price predictions with generalisable tailored neural networks that outperform SOTA models. We rely on multilingual sentence transformers to support multiple languages, and we have developed a methodology based on consumer indifference to convert between pricing schemes. Finally, we have pointed to data augmentation techniques that can help further improve the robustness and generalisability of the tool.

This work represents the first step of an effort to build and operate an observatory of the data economy. We are working on updating, extending and enriching our training data by streamlining data ingestion processes to refresh market information more frequently, and to validate and further improve the robustness of our models. Moreover, we look forward to ingesting and integrating information about real data transactions, and to actively interfacing with ongoing standardisation initiatives, such as the International Data Spaces [22] and the Gaia-X project [4] for data exchanges, or W3C Data CATalog vocabulary [42] for metadata specification.

ACKNOWLEDGMENTS

This work was partially developed by the first author while working at LSTech España S.L. and partially supported by the European Union's HORIZON project UPCAST (101093216).

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proc. of ACM EC*. https://doi.org/10.1145/3328526.3329589

[3] S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. 2023. Understanding the Price of Data in Commercial Data Marketplaces. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. 3718–3728. https://doi.org/10.1109/ICDE55515.2023.00300

[4] Gaia-X Association. [n.d.]. Together Towards a Federated and Secure Data Infrastructure. https://gaia-x.eu/. Last accessed: May 2025.

[5] Santiago Andrés Azcoitia. 2023. Data Pricing Tool. https://gitlab.com/sandresazcoitia1/data-pricing-tool

[6] Santiago Andrés Azcoitia, Costas Iordanou, and Nikolaos Laoutaris. 2022. Measuring the Price of Data in Commercial Data Marketplaces *(DE '22)*. ACM, 1–7.

[7] D. Bunn and E. Asen. 2022. What European Countries Are Doing about Digital Services Taxes. https://taxfoundation.org/data/all/eu/digital-tax-europe-2022/. Last accessed: Feb'24.

[8] Shuchi Chawla, Shaleen Deep, Paraschos Koutris, and Yifeng Teng. 2019. Revenue maximization for query pricing. *Proc. of the VLDB Endow.* 13 (2019). https://doi.org/10.14778/3357377.3357378

[9] McKinsey & co. 2018. Notes from the AI Frontier: Modeling the Impact of AI on the World Economy. (2018).

[10] Z. Cong, X. Luo, J. Pei, F. Zhu, and Y. Zhang. 2022. Data Pricing in Machine Learning Pipelines. *Knowl. Inf. Syst.* 64, 6 (2022), 1417–1455.

[11] PWC Consulting. 2017. Sizing the Prize - What's the Real Value of AI for your Business and How Can you Capitalise? (2017).

[12] IDC & The Lisbon Council. 2023. European DATA Market Study 2021–2023. (2023). https://digital-strategy.ec.europa.eu/en/library/results-european-data-market-study-2021-2023

[13] Diane Coyle and Annabel Manley. 2023. What is the Value of Data? A Review of Empirical Methods. *Journal of Economic Surveys* 38 (08 2023), 1317–1337.

[14] Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. *Proc. of the ICML* (2019).

[15] Arpita Ghosh and Aaron Roth. 2011. Selling Privacy at Auction. In *Proc. of the ACM EC '11* (San Jose, California, USA). https://doi.org/10.1145/1993574.1993605

[16] L. Giaretta, T. Marchioro, E. Markatos, and Š. Girdzijauskas. 2022. Towards a Decentralized Infrastructure for Data Marketplaces: Narrowing the Gap between Academia and Industry. In *Proc. of the 1st Workshop on Data Economy* (Rome, Italy) *(DE '22)*. ACM, 49–56.

[17] Andrew V. Goldberg and Jason D. Hartline. 2003. Competitiveness via Consensus. In *Proc. of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms* (Baltimore, Maryland) *(SODA '03)*. Society for Industrial and Applied Mathematics, USA, 215–222.

[18] Andrew V. Goldberg, Jason D. Hartline, and Andrew Wright. 2001. Competitive Auctions and Digital Goods. In *Proc. of the ACM-SIAM Symposium on Discrete Algorithms* (Washington, D.C., USA). Society for Industrial and Applied Mathematics.

[19] Mohammed El-Erian Gordon Brown and Michael Spence With Reid Lidow. 2023. *Permacrisis. A Plan to Fix a Fractured World.* Simon & Schuster.

[20] J. R. Heckman, E. Boehmer, E. H. Peters, Milad Davaloo, and Nikhil G Kurup. 2015. A Pricing Model for Data Markets. In *iConference 2015 Proceedings*.

[21] Nicolaus Henke and Jacques Bughin et al. 2016. The Age of Analytics: Competing in a Data-driven World. *McKinsey Global Institute* (2016). https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world#

[22] IDSA. 2022. International Data Spaces Reference Architecture Model. *International Data Spaces Association* (2022).

[23] Javen Kennedy, Pranav Subramaniam, Sainyam Galhotra, and Raul Castro Fernandez. 2022. Revisiting Online Data Markets in 2022: A Seller and Buyer Perspective. *SIGMOD Record* 51, 3 (Nov. 2022), 30–37. https://doi.org/10.1145/3572751.3572757

[24] Keras. 2015. Simple. Flexible. Powerful. https://keras.io/. Last accessed: Jul'25.

[25] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc of ICLR '15*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[26] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. 2012. QueryMarket demonstration: Pricing for Online Data Markets. *Proc. of the VLDB Endow.* 5 (2012).

[27] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. 2015. A Theory of Pricing Private Data. 39, 4, Article 34 (2015), 28 pages. https://doi.org/10.1145/2691190.2691191

[28] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proc, of NIPS* (Long Beach, California, USA) *(NIPS'17)*. Red Hook, NY, USA, 4768–4777.

[29] J. Majewski. 2020. Training Neural Networks for Price Prediction with TensorFlow, [Online]. Available: https://medium.com/data-science/training-neural-networks-for-price-prediction-with-tensorflow-8aafe0c55198. Repository: https://github.com/Jan-Majewski/Project_Portfolio/blob/master/03_Real_Estate_pricing_in_Warsaw/03_04_Training_Neural_Networks.ipynb.

[30] V. Masterson. 2024. 9 Ways AI is Helping Tackle Climate Change. (2024).

[31] Chaoyue Niu, Zhenzhe Zheng, Fan Wu, Shaojie Tang, Xiaofeng Gao, and Guihai Chen. 2018. Unlocking the Value of Privacy: Trading Aggregate Statistics over Private Correlated Data. In *Proc. of ACM SIGKDD*. https://doi.org/10.1145/3219819.3220013

[32] Olga Ohrimenko, Shruti Tople, and Sebastian Tschiatschek. 2019. Collaborative Machine Learning Markets with Data-Replication-Robust Payments. *CoRR* (2019). arXiv:1911.09052 [cs.GT]

[33] European Parliament and the Council of Europe. 2022. Regulation (EU) 2022/868 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act).

[34] Jian Pei. 2020. Data Pricing – From Economics to Data Science. In *Proc. of the ACM SIGKDD* (Virtual Event, CA, USA). 3553–3554. https://doi.org/10.1145/3394486.3406473

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[36] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[37] Carl Shapiro and Hal R. Varian. 2000. *Information Rules: A Strategic Guide to the Network Economy.* Harvard Business School Press.

[38] Lloyd S. Shapley. 1952. *A Value for n-Person Games.* RAND Corporation. https://www.rand.org/pubs/papers/P0295.html

[39] Markus Spiekermann. 2019. Data Marketplaces: Trends and Monetisation of Data Goods. *Intereconomics* (2019), 9.

[40] J. Ulloa. 2019. Newsom wants companies collecting personal data to share the wealth with Californians. https://www.latimes.com/politics/la-pol-ca-gavin-newsom-california-data-dividend-20190505-story.html. Last accessed: Feb'24.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in NIPS*, Vol. 30. Curran Associates, Inc.

[42] W3C. [n.d.]. Data Catalog Vocabulary (DCAT) - Version 3. https://www.w3.org/TR/vocab-dcat-3/. Last accessed: May 2025.

[43] J. Walczak. 2020. New York Lawmakers Float New Data Tax Proposal. https://taxfoundation.org/blog/new-york-data-tax-proposal/. Last accessed: Feb'24.

[44] Haifei Yu and Mengxiao Zhang. 2017. Data pricing Strategy based on Data Quality. *Computers and Industrial Engineering* 112 (2017), 1–10. https://doi.org/10.1016/j.cie.2017.08.008

[45] Jiayao Zhang, Yuran Bi, Mengye Cheng, Jinfei Liu, Kui Ren, Qiheng Sun, Yihang Wu, Yang Cao, Raul Castro Fernandez, Haifeng Xu, Ruoxi Jia, Yongchan Kwon, Jian Pei, Jiachen T. Wang, Haocheng Xia, Li Xiong, Xiaohui Yu, and James Zou. 2024. A Survey on Data Markets. arXiv:2411.07267 [cs.GT] https://arxiv.org/abs/2411.07267

[46] Mengxiao Zhang, Fernando Beltrán, and Jiamou Liu. 2023. A Survey of Data Pricing for Data Marketplaces. *IEEE Transactions on Big Data* 9, 4 (2023), 1038–1056. https://doi.org/10.1109/TBDATA.2023.3254152

[47] Yiding Zhu, Hongwei Zhang, Jiayao Zhang, Jinfei Liu, and Kui Ren. 2024. DataPrice: An Interactive System for Pricing Datasets in Data Marketplaces. *Proc. VLDB Endow.* 17, 12 (Aug. 2024), 4433–4436. https://doi.org/10.14778/3685800.3685893

[48] Zilliz. 2024. The Guide to all-mpnet-base-v2. *Zilliz.com.* [Online]. Available: https://zilliz.com/ai-models/all-mpnet-base-v2. Technical documentation for sentence embedding model.