Poster. What Is the Price of Data? A Measurement Study of Commercial Data Marketplaces

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org
IMDEA Networks Institute
Leganes, Madrid, Spain

Costas Iordanou
kostas.iordanou@cut.ac.cy
Cyprus University of Technology
Limassol, Cyprus

Nikolaos Laoutaris
nikolaos.laoutaris@imdea.org
IMDEA Networks Institute
Leganes, Madrid, Spain

ABSTRACT

Data-driven decision making powered by ML is changing how the society and the economy work and is having a profound positive impact on our daily life. A McKinsey report predicted that data-driven decision-making could reach US$2.5 trillion globally by 2025, whereas the European Data Strategy estimates a size of 827 billion euro for the EU27. ML is driving up the demand for data in what has been called the fourth industrial revolution.

A large number of Data Marketplaces (DMs) have appeared in the last few years to help owners monetise their data, and data buyers fuel their marketing processes, train their ML algorithms, and make data-driven decisions. In this poster, we present some preliminary findings of what is, to the best of our knowledge, the first systematic measurement study of DM for data products. This ecosystem, despite being quite vibrant commercially, remains completely unknown to the scientific community. Very basic questions such as "*What is the range of prices of data traded in modern DMs?*", "*Which categories of products command the highest prices?*", "*Are the observed prices consistent across DMs?*", "*Which features correlate with the most expensive data products?*" appear to have no answer and evade most meaningful speculations.

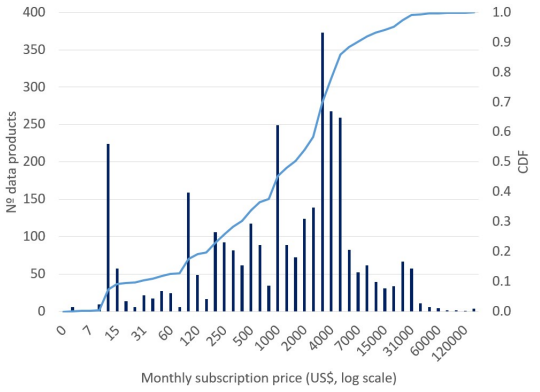
To answer such questions, we first conducted an extensive survey for compiling a catalogue with more than 180 DMs. We then selected 38 of them that fulfill necessary criteria for a measurement study. For these DMs we developed custom crawlers for retrieving information about the products they trade. Using these crawlers we obtained information for more than 213,964 data products and 2,015 data providers. We also developed ML classifiers for identifying data products of similar categories to compare prices across DMs, and executed 9 different regression models to understand which features are driving the prices of data products.

Preliminary findings: Analysing the collected data we observed that the majority of data products were either given for free, or did not carry a fixed price, but rather were up for direct negotiation between the seller and interested buyers. Regarding the 4,200 products that carried a price, we observed the following:

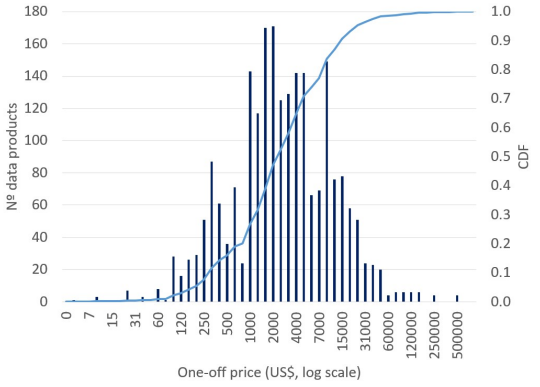
- Prices vary in a wide range from a few US-dollars up to several hundred of thousands. The median price for data products sold under a *subscription* model is US$1,400 per month, and US$2,200 for those sold as an *one-off* purchase.

Table 1: Summary of scraped DMs

Marketplace	#Data Products	#Paid products	#Sellers
Advaneo	198,743	1	N/A
AWS	4,013	2,515	262
DataRade	1,592	1,592	1,262
Knoema	158	158	142
DAWEX	160	160	79
Carto	8,182	5,283	42
Crunchbase	16	14	15
Veracity	115	95	38
Refinitiv	214	185	76
Other data providers	771	769	29



(a) Subscription-based



(b) Fixed price (one-off)

Figure 1: Histogram and CDF of data products

Table 2: Top 10 most relevant features by category and algorithm

Financial			Marketing			Healthcare			All		
RF	**kNN**	**GBR**	**RF**	**kNN**	**GBR**	**RF**	**kNN**	**GBR**	**RF**	**kNN**	**GBR**
units	units	units	units	units	csv	units	csv	wordlist	units	units	DelMethod
entities	Email	S3Bucket	locationdata	History	units	people	units	DelMethod	yearly	idSessions	S3Bucket
S3Bucket	Download	wordmonthli	Weight	USA	yearly	wordhealth	daily	wordhospit	Download	Retail	units
wordsubmit	daily	wordstock	USA	idSessions	people	wordtrend	wordmarket	wordidentifi	wordreport	USA	entities
Download	IdCompanies	worddeliv	IdCompanies	NCountries	RESTAPI	wordmedic	wordgo	wordamerica	entities	IdCompanies	requests
people	USA	people	txt	Financial	wordqualiti	wordglobal	Limitations	wordhealth	people	worduser	people
txt	wordmarket	DelMethod	daily	Others	wordaccur	csv	locationdata	wordreport	wordmarket	Others	yearly
wordedgar	Retail	txt	S3Bucket	people	wordidentifi	DelMethod	wordpopul	wordstudi	monthly	wordconsum	pdf
wordcustom	wordcontact	wordneed	wordmonthli	wordcontact	wordwebsit	wordinsight	wordprofil	wordupdat	wordcontact	Canada	RESTAPI
wordlist	realtime	wordsubmit	wordvp	Email	UIExport	wordreport	wordinsight	wordcontact	wordwebsit	wordcompani	wordlist

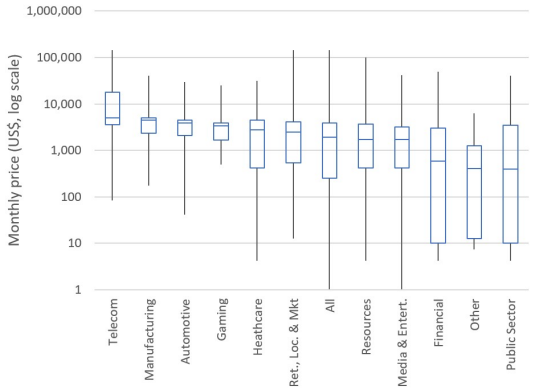


Figure 2: Box plot of subscription-based prices in AWS

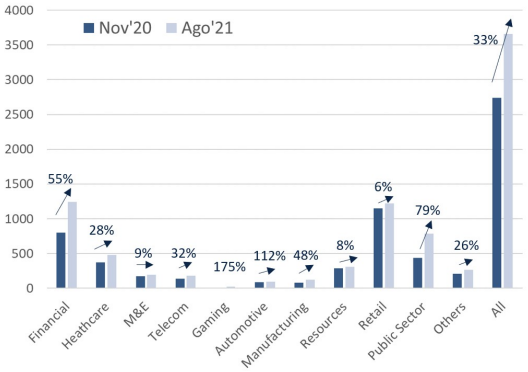


Figure 3: N° products by category in AWS

Table 3: Results of price prediction models

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
DNN	0.73	0.33	0.35	0.77	0.30	0.22	0.68	0.26	0.18			

- Focusing on AWS' DM we found that those related to *telecoms*, *manufacturing*, *automotive* and *gaming* command the highest median prices. We also studied temporal aspects of DMs and noticed that DMs such as AWS have been growing with a significant 3% monthly rate from Dec'2020 to Aug'21.

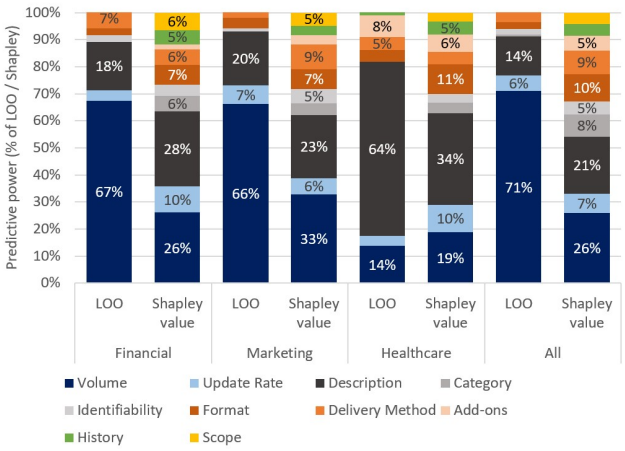


Figure 4: Predicting power by feature group

- Using classifiers, we enriched our sample by consistently labelling products across DMs.
- Using regression models, we managed to fit the prices of commercial products from their features with R^2 above 0.84.
- Features related to volume and units, and domain-specific characteristics of data products captured by their descriptions and categories are responsible for 66% of this score.
- Due to the heterogeneity of the sample there is no single feature that drives the prices, but instead we spotted meaningful features that proved to be conclusive in specific domains: stems like '*custom*', '*edgar*' or '*market*', and '*contact*', '*identifi*' or '*accur*' appear in the top 10 for *financial* and *marketing* products respectively. Interestingly, data update rate seems to be a key price driver for *financial* and *healthcare*-related products, whereas the ability to provide exact locations and the possibility of connecting data points from the same owner are for *marketing* data.

Future work: The significant growth rate we have seen at AWS and other DMs makes us believe that in the future their paid catalogue is bound to grow and therefore, we will continue monitoring them to see how they evolve. We are also working on improving our feature importance analysis and extending it to become a price recommendation tool for new data products.