Doorda Report

The Importance and Impact of Using External Data to make Critical Decisions

Research shows a huge appetite for UK external data, but that it can be without ease of discovery and diligence expected

DOORDA
TRUSTED DATA, ON DEMAND

Contents

About This Report

Doorda has uncovered the prevalence, impact, opportunities and potential risks of using external data when making critical business decisions.

Given that most businesses are now incorporating technologies that encompass complex internal and external data analytics, the automation of business processes, Artificial Intelligence (AI) and the optimisation of customer engagement; this report reviews the potential and impact of integrating and using external data in today's digital world. Conducted in the UK by OnePoll, it reviews and reports on the importance of a diligent approach and the exponential impact of using external data to innovate and capitalise on new commercial opportunities.

Commercial and consumer behaviours are changing, whether it's revenue growth, personalisation, assessing and managing risk, security or sustainability considerations, businesses are looking to data analytics and data science to differentiate, compete and thrive. Understanding how best to add the huge potential value of external data to analytical processes and models is therefore critical to success.

Methodology

Between 28th June – 5th July 2023 OnePoll conducted a survey of 300 employed adults with roles responsible for data science, data analytics and predictive models.

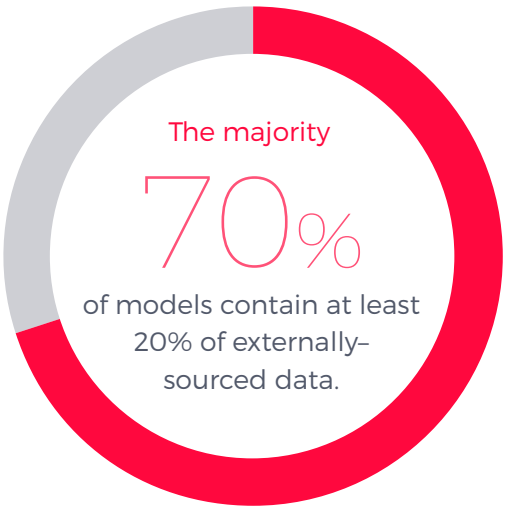
Sectors included in the survey: financial services, insurance and pensions, financial technology, insurance technology and professional services.

Executive Summary

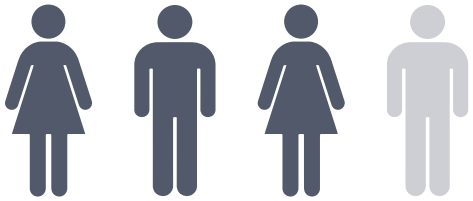
This survey of UK data scientists reveals that the use of external data in data science has quietly become immensely important, with 92% of companies seeing significant improvements to outcomes thanks to it.

Businesses must now take account of the importance of external data, and look for improvements that can be made to external data discovery and integration to further drive innovation and the associated analytical benefits.

The survey revealed just how important external data is to improving data science outcomes with:



Data science requires research, experimentation and some "trial and error". Results revealed that you don't necessarily get the best results the first time around by using external data, but when it's relevant and applied correctly, results shine through.



3 out of 4 respondents have had results which exceeded their expectations.

Finding new data is key to continuing the improvement and innovation of new and existing data models. Introducing fresh and trustworthy data is becoming increasingly important with the rise and dominance of Large Language Models (LLMs) and AI. Models such as ChatGPT learn from a vast amount of diverse data, hence it's crucial to get inputs right. The importance of curating and monitoring data quality that might affect outputs and bias is perhaps why:

Responses show that more could be done from a due diligence POV when it comes to the data supply-chains. Regularly acquiring quality data from trusted sources in a timely manner is a clear priority but external data supply chains are often fragile and under-vetted compared with internal data governance procedures.

2 in every 5 respondents found useful data, only to be let down later by its quality or availability.

When sourcing new external data, going to known suppliers and asking trusted colleagues are, unsurprisingly, the most popular approach, but those who only work with one or limited external data suppliers may be missing out.

All expected and proposed avenues of exploration were represented – there is no single, perfect way of finding good external data.

This survey of UK data scientists revealed an astonishing 140 different job titles where the respondent's role is principally data science and data analytics. The number and inconsistency of job titles shows that data science and analytics are still evolving and raises the question whether companies realise just how widespread and important data science has become within their own organisations.

> "there is no single, perfect way of finding good external data"

Research Highlights

External data is the elephant in the room

70% of survey respondents reported that their models contain at least 20% of externally-sourced data.

This is enormously revealing and important as while sourcing, cleaning and validating any data can be complex and time consuming, external data presents more of these concerns than with internal data, but still plays a significant part in most models.

Both internal and external data should undergo due diligence in data modelling processes, but internal data is often considered more familiar and controllable. Lack of familiarity and/or previous experiences with external data that may not have yielded immediate results or required justification often cause extreme unease for data practitioners. But external data introduces potential insights that those tasked with performance now need to take on board in the competitive, data-led digital world.

> External data cannot be overlooked or avoided

Increasingly customers and regulators also expect to know more about business partners in the chain, their values, any risks, or 'unacceptable' business practices or ESG related considerations. Verifying correctness and trustworthiness is key. Whilst the survey shows it is important to fully consider all options for augmenting internal with external data, the use of modelled external data (e.g., "scores") that are opaque in origin need to be weighed carefully against using fully documented raw data variables whose provenance is certain.

Curious minds, but well–trodden paths

As target markets and behaviours shift, businesses need to digitally market to, or personalise products for their customers and partners, hence the requirement to regularly investigate and improve existing models grows.

Most respondents now recognise that they need additional data to satisfy evolving business requirements, and that newly available data sources could improve existing data models.

There is no single, perfect way of finding good external data, with results spread across a number of source potentials, but existing data suppliers, established market–places and internal colleagues are unsurprisingly the main ways of finding out about external data.

Given the high usage of external data, this suggests that organisations are relying on what they believe to be 'safe' routes, or are lacking in new, known sourcing options, or possibly they are not approaching external data sourcing and integration rigorously enough to draw proper value.

The report throws fuel on the quiet debate that the search and integration of external data is not always subject to the same governance attention and weighting to that of internal data.

Data science strives for accuracy, but the process is not always exact

When responding to the question of 'has the positive improvement of adding external data ever exceeded your expectations' 78% was the big margin winner.

> 78% stated that the positive effect of adding external data had exceeded their expectations

Clearly when data scientists find that extra gem and new insights emerge, then the business rewards and personal recognition are palpable, but there is still a certain amount of trial and error in reaching these accomplishments. Science is not always exact the first time and there will always be a need to mine for diamonds in rough terrain.

The big theme of 'our time' is AI which is a subset of the field of data science, and choosing appropriate machine learning algorithms or statistical techniques that align with problems and data characteristics are part of the data scientist's day job. Data science involves iterative processes and a structured approach, but experimentation within this is key – trial and error is the only route to learning.

The importance of setting aside a budget for research, integration and data trialling may present some business challenges as it requires experience and budget redirection, but the potential benefits clearly justify experimentation and occasional failure. **Those businesses that think more creatively about what they can do differently and who broaden their searches for external data in order to achieve innovation will leap frog familiar outcomes and energise results.**

Pose hard questions about external data availability... early

Disregarding 'cost', a broad range of criteria are used for deciding whether to use a new external dataset. 54% of respondents cite 'quality' with ease of access, supplier knowledge and experience being the next-best top 3 decision factors.

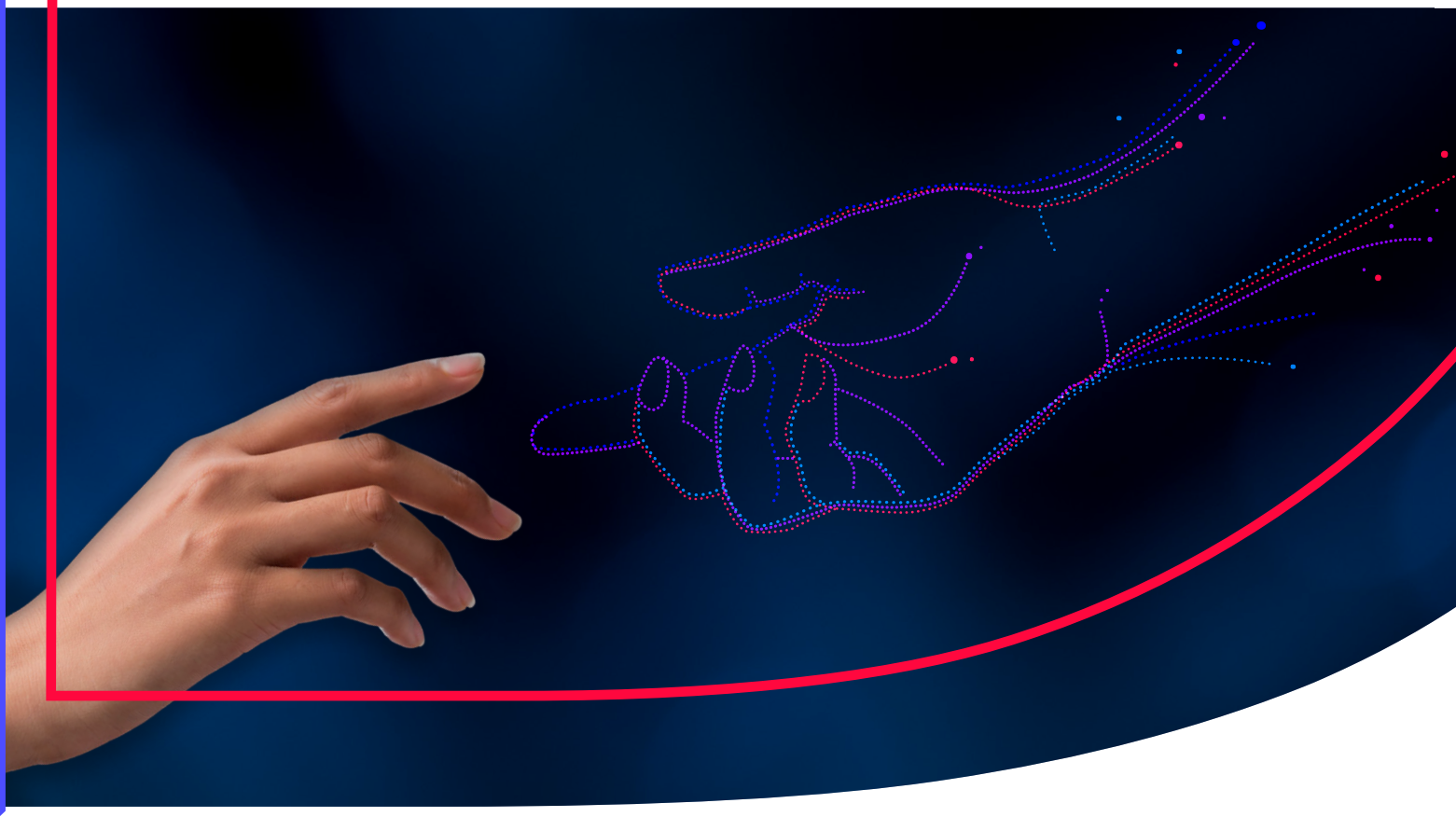
Few will be surprised that 'quality' is top of the pile–it's arguably an easy catch-all, correct response especially as data quality measures can and should encompass completeness, consistency, validity, timeliness and relevance. Aligned to this is the fact that 40% of respondents say they have been let down at some point by untrustworthy or unavailable data from their supplier.

Users and businesses will not succeed in their data analysis endeavours if they are let down by data quality or availability. **Protecting and governing a 'data supply chain' so that it remains trustworthy and available is of primary importance and should be evaluated as part of the 'quality' of each new source.**

When looking for external data, pose the hard questions at the start of engagement:

- Where's it sourced from?
- Can it be acquired regularly?
- How do you ensure it's complete and valid on a consistent basis?
- How frequently is it refreshed?

> "Assumptions should not be made that because there is good data one week, there will be good data another week."



Official does not mean “standardised”

The challenges faced when integrating external data are varied and complex with the top 3 ‘biggest challenges’ again being poor data quality or other attributes of ‘quality’ such as incompleteness and incompatibility.

External data often comes with the classification and description of ‘official’ – a badge of honour that can imply a certain level of legitimacy and credibility, but this is not a guarantee of absolute trustworthiness.

Official sources are typically recognised, endorsed and provided by relevant authorities, institutions or organisations. This can suggest a higher level of authority, but a certain level of critical thinking needs to be applied to its application and regular format standardisation is rarely a given. It’s especially important to note that:

- Official doesn’t mean standardised. There are different levels of aggregation and formatting.
- Official data is published in awkward, inconsistent ways that can often not be easily linked with other data.

Away from the data itself, it’s generally junior data practitioners who are tasked with collecting, integrating and cleaning data sources. Often these less experienced members of data teams are unknowingly not questioning data integrity and it’s a case of just getting the ‘job done’ for their seniors.

Across organisations, many with data responsibilities are applying different approaches, duplicating efforts and relying on potentially stale or poor-quality data sources.

A consistent approach that encourages better governance is needed every time data is taken from every external source to decide whether the data can be trusted and used to make the most of it.

If the external data is supplied by a third party, it may make sense to address governance in partnership with them so that both compliance needs can be met and innovation can thrive.

About Doorda

Doorda provides high quality, analytics–ready data, enabling organisations to obtain new insights, faster.

Data about UK Properties, Businesses and Geo–demographics are gathered from over 1,500 Official sources, categorised and uniquely linked, updated daily and made available in the Doorda Cloud Data Platform.

www.doorda.com

Appendix –
Full Survey Results

Data Scientists Survey

Commissioned by Doorda
Conducted by OnePoll

Date: 28th June – 5th July 2023

Sample: 300

Demographic: UK: Employed adults who are mainly responsible for Data Science, Data Analytics, or Predictive Models in their role.

1. What percentage of the model data you use is externally sourced?	
0%	2%
1%- 5%	1%
6 - 10%	8%
11% - 20%	17%
21% - 40%	32%
41% - 60%	26%
61% - 80%	10%
81% - 100%	2%
Not sure	2%

2. How important is it to regularly investigate external data that could improve an existing model?	
Very important	59%
Somewhat important	40%
Not very important	2%

3. When new external data has been added to models you've worked on, what has been the best improvement you have ever seen?	
The model was improved a lot	29%
The model was improved somewhat	63%
The model was not improved much	7%
Not sure	1%

4. Has the positive improvement of adding new external data ever exceeded your expectations?	
Yes	78%
No	17%
Not sure / can't remember	5%

5. What, if any, are the main ways you find out about potentially interesting or new external data? [Select up to three]	
Existing data suppliers	46%
Data marketplace	45%
Internal Colleagues	35%
Existing services / software suppliers	34%
Web search	30%
Big-name data providers	26%
External Forums / peer group	23%
Not sure / no way in particular	2%
Other (please specify)	0%

6. Disregarding "cost", what are your main criterias for deciding whether to use a new external dataset? [Select up to three]	
Data Quality	54%
Ease of access	35%
Knowledge & experience of data supplier	28%
Results of testing the data in my data model	24%
Well known data supplier	22%
Data is new / unique / never seen before	21%
Clear permitted use and compliance statements	20%
Data provenance / origin	20%
Available from a Data Marketplace	15%
Colleague recommendation	11%
Forum / peer-group recommendation	9%
Not sure / none in particular	1%

7. Have you ever used data in a model that later proved untrustworthy or became unavailable?	
Yes	40%
No	51%
Not sure / can't remember	8%

8. What, if any, are the biggest challenges you face when integrating external data? [Select up to three]	
Poor data quality	34%
Gaps / missing data	33%
Incompatible data	31%
Large volumes	26%
Compliance regulations	25%
Linking to existing data	24%
The data ingestion processes	22%
Inadequate data documentation	20%
Freshness / keeping data up to date	17%
Not sure / none in particular	3%
Other (please specify)	0%

9. What are the main ways that you approach licensing and compliance with new external data sources? [Select up to three]	
Engage compliance team	44%
Read all licensing rules	39%
Ask supplier	32%
Use only proprietary sources	27%
Use only Open Data	27%
Licencing and compliance is not an issue for me	20%
Rely on colleagues to resolve	18%
Not sure / no way in particular	3%