A MATTER OF FACT?

Practical steps for making Big Data work

By Svetlana Gogolina | July 2020

IPSOS VIEWS

GAME CHANGERS

Ipsos

USING DATA FOR MARKETING INSIGHTS

Data science has been a part of insight generation and targeting for years. Consumer behavioural data has been used to predict future purchases and switching behaviour. Internet search data informs targeted advertising of similar products.

Today, the task in hand is to make better use of a vast amount of what we call “unrequested data” or real-world evidence (RWE) data: data that is collected for purposes other than market research, such as electronic patient records data or traffic data.

Mastering these new various data sources and combining them to generate insight is a new challenge for the market research industry and for market insight teams. This journey certainly offers exciting new opportunities and the possible pitfalls along the way can be avoided - if you know where to look.

For those of us who are involved in market research and marketing insight, this paper sums up some best practices for generating maximum insight from diverse data sources, and avoiding data disasters.

BIG OPPORTUNITIES

The Big Data revolution, powered by fast computers, mighty computing algorithms and a proliferation of data sources, is transforming the industries surrounding us. The speed of change is staggering and the changes are incredibly profound.

Exciting developments have delivered some great success stories, ranging from image recognition technologies being successfully used for patient diagnostics to the first prototypes of driverless cars. Big claims are being made about data science and AI: that their development is one of the most important events in the history of humanity.

Most of us in business insight functions believe that Big Data and AI-powered multi-source data analytics can and should be used for business insight generation. Multi-data source analytics has recently became a buzz word in our industry, and rightly so.[1]

Many companies across industries are investing in data analytics teams and data science capabilities. We all are on the same exciting, challenging and often painful journey of discovery.

BIG CHALLENGES

Data quality is the key problem when it comes to Big Data. After an initial phase of complete excitement and wildly unrealistic expectations, it has become clear that Big Data doesn't automatically translate into Big Insights. To make it work, there are a few things that we need to get right first.

Suboptimal data quality is enemy number one when it comes to the widespread and profitable application of machine learning algorithms to data sources for extracting insights.

The notion of "rubbish-in, rubbish-out" has been known in analytics forever, however it has a special sinister meaning for machine learning. Machine learning has strict demands for quality because bad data affects the process in two ways: firstly in the historical data that is used to train the predictive model, and then in the new data used by that model to make future decisions.

To properly train a predictive model, historical data must satisfy both broad coverage and high quality standards. The data must be right. It must be correct, properly labelled, and de-duplicated.

But you must also have the *right* data. This means lots of unbiased data over the entire range of inputs that the predictive model is being developed for.

Increasingly complex problems demand not just more data, but more diverse, comprehensive data. And this brings more quality problems...

Unfortunately, today most data fails to meet basic quality standards. This is due to a range of factors that include data creators not understanding what is expected, patchy coverage, lots of missing poorly recorded data or human error.

It comes without saying that data scientists should "clean" the data before training the predictive model. It is time-consuming, tedious work (taking up to 80% of data scientists' time). And – unfortunately – even with such efforts, cleaning neither detects nor corrects all the errors. There is also no way to understand the impact on data cleaning techniques on the predictive model.

Furthermore, increasingly complex problems demand not just more data, but more diverse, comprehensive data. And this brings more quality problems. For example, handwritten notes and local acronyms have massively complicated IBM's efforts to apply machine learning to cancer treatment.

Data quality is equally a great challenge for implementation when it comes to applying the trained model to new data. While the data science team (who developed the predictive model) may have done a good job cleaning the training data, it can still be compromised by any bad data that it is applied to.

To give just one example, when working with social text data, our Social Intelligence Analytics team spends a significant amount of time cleansing the data from commercial advertising and reposts of press articles. The team also use specialised dictionaries to train natural language processing (NLP) algorithms to recognise technical terms such as molecule names or technical jargon when working with healthcare texts, for example. The algorithms are trained to recognise shorthand and abbreviations used.

In order to derive insights that you can trust, a surprising amount of work often needs to take place in the background.

We must remember that if the data is bad or the coverage is poor, even the best AI algorithms are also useless.

DATA INTEGRATION

PRACTICAL SOLUTIONS

How can we avoid 'garbage in -garbage out' situations and make data integration a success? Human intelligence is still needed, at least for the time being…

Here are eight rules that, in our experience, help to optimise the process:

1. Whenever a model is built, it should be guided by **clearly articulated business objectives** and hypotheses. Mining data without a sense of direction won't lead you to insights.

2. **Do not believe in 'magic algorithms'**. Algorithms do not work on their own.

3. A **sound knowledge of data sets** is essential. This includes the coverage, the limitations, the meaning of fields and variables, the codes, and what the data set should (and should not) be used for.

4. Most secondary data sources give you answers to the **'what'** questions but they cannot always tell you **'why'**. You need to use other types of data for these kinds of inquiries - and not unrequested data, which is rarely used for causal models because of its limitations.

5. **Primary market research can be used to fill the gaps** in secondary data sources and to verify any assumptions about secondary data sources.

6. Use qualitative and observational research to bring out the **human story** from the data.

7. Data scientists need to be **familiar with the subject matter**. This is the case for specialised areas such as healthcare, but also applies to media measurement and the public sector.

8. In order for analytics to work, the data science team needs to work as part of a broader **multi-disciplinary team** of subject experts and business stakeholders.

THE KNOWLEDGE GAP

In more specialised areas like healthcare, it is important to recruit medically trained data scientists to work on extremely complex patient records or medical claims data. Similarly, teams working on audience measurement projects are all specialists in audience measurement and know most of the existing data sources very well. The same of course applies to data scientists working with public sector data.

This poses a problem: where do you find data scientists who are also experts in a certain very specialised area (such as oncology healthcare), but also have a good understanding of business challenges? There are few such people around who combine all these desired skills in one person.

PRINCIPLES FOR HANDLING DATA

While there are some shared principles for handling data, different types have their own unique considerations. It is important to know what you are working with and for what purpose, to remove bias, and not to make assumptions.

Specific subject knowledge is important so that data scientists can identify variables for data fusion. And this needs to be done in line with a longer view of what goals the insights from the fused data set will serve.

Having just technical data science skills is rarely enough. It is important to remove bias from each data set before different sources are integrated or fused.

For example, when measuring audiences, the greatest challenge is to provide the industry with a reliable and objective measure of performance while dealing with data sources that are inevitably biased towards certain media owners. Bias in print readership will have different causes from bias in data sources supporting out-of-home or digital advertising.

To give a second example, working with patient real-world data is extremely challenging in many ways, but especially in terms of regulations and compliance. When collecting patient chart data provided by physicians, we cannot collect any information that potentially allows to identify the patient. Data from electronic patient records comes to us fully anonymised so it is not possible to match patients in our sample with the same patient in electronic patient records - the data needs to be fused. We need to find a 'patient X' in one data set who probabilistically likely to look very similar to 'patient Y' in the other data set. This is all in terms of their disease treatment, clinical characteristics, potential treatment pathway and outcomes.

With any type of modelling, assumptions usually need to be made, and it is important to verify the validity of these assumptions. When this is done without knowledge and consideration, the most powerful algorithms produce the most misleading results. But, when these steps are followed, integrated data sources offer amazing opportunities.

THE "HUMAN" PERSPECTIVE

We must not forget about the importance of the human perspective. There are still many areas where, compared to a machine, humans still have the upper hand: we have experience, adaptability, empathy – and all of these are needed when it comes to working out how best to join different data streams.

Real-world evidence data sets do not provide an understanding of the human element: the "human stories" behind the numbers. A wide range of tried and tested primary qualitative and quantitative research techniques remain the best way of understanding the perceptions, needs, motivations and emotions of the key stakeholders. These techniques are the essential complement of the insights generated from unrequested Big Data. Qualitative insight is one of the many available tools that will need to work together with other data sources.

Qualitative research also allows insights developed from different data sources to be embedded much more effectively within an organisation.

Even from the perspective of a data person, I firmly believe that the more muti-data source analytics we do, the more important the human insight will become.

THE FUTURE

Big Data and data science will continue to disrupt market insights. However, as powerful and as helpful as data science is, it will not be able to address *all* industry information needs - at least not for a while.

At Ipsos, we believe there remains a firm need for primary research, as the optimal way to identify and understand the real person behind the data.

But in today's multi-data source reality, primary research is likely to become shorter, more agile and focused, and will be used in combination with other sources more often.

The unlikely combination of data science and qualitative research will be key for marketing insights professionals of the future, with qualitative research likely to become the best friend of a data scientist.

And there will always be the need for a person who 'tells the story' behind the data.

FURTHER READING

1. Multi-source healthcare data, Ipsos Views, April 2020: https://www.ipsos.com/en/multi-source-healthcare-data

A MATTER OF FACT?

Svetlana Gogolina UK Head of Data Science, Ipsos

The **Ipsos Views** papers
are produced by the
Ipsos Knowledge Centre.

www.ipsos.com
@Ipsos

GAME CHANGERS