ANALYTICS

The Value of Big Data Predictive Analytics

by Jacob LaRiviere, Preston McAfee, Justin Rao, Vijay K. Narayanan, and Walter Sun

The "Big Data" revolution is upon us: Firms are scrambling to hire a new brand of analysts dubbed "data scientists" and universities have responded to this demand by introducing data science courses into degrees ranging from computer science to business. Survey-based reports find that firms are currently spending an estimated $36 billion on storage and infrastructure and that is expected to double by 2020. Once companies are logging and storing detailed data on all their customer engagements and internal processes, what's next? Presumably, firms are investing in big data infrastructure because they believe that it offers a positive return on investment. However, looking at the surveys and consulting reports it is unclear what are the precise use cases which will drive this positive ROI from big data.

Our goal in this article is to offer specific, real-world case studies to show *how* big data has provided value for companies that have worked with Microsoft's analytics teams. These cases reveal the circumstances in which big data predictive analytics are likely to enable novel and high value solutions, and the situations where the gains are likely to be minimal.

Predicting demand. The first use case involves predicting demand for consumer products that are in the "long tail" of consumption. Firms value accurate demand forecasts because inventory is expensive to keep on shelves and stock-outs are detrimental to both short-term revenue and long-term customer engagement. Aggregated total sales are a poor proxy because firms need to distribute inventory geographically, necessitating hyper-local forecasts. The traditional way of solving this problem is to use time-series econometrics with historical sales data. This method works well for popular products in large regions but tends to fail when data gets thin because random noise overwhelms the underlying signal.

A big data solution to this problem is to use anonymized and aggregated web search and/or sentiment data linked to each store's location on top of the existing time-series data. Microsoft data scientists have employed such an approach to help a forecasting firm predict auto sales. Building machine learned models with web search data as one of the inputs reduces Mean Absolute Forecast Error, a standard measure of prediction accuracy, for monthly national sales predictions on the order of 40% from baseline for auto makes with relatively small market shares, compared to traditional time series models. While the gains were smaller for the most popular models at the national level, the relative improvement increases as one drills down to the regional level.

In this case, the big data solution leverages the previously unused artifact that people do a considerable amount of research online and social inquiry before buying a car. The increased prediction accuracy, in turn, makes it possible to achieve large increases in operational efficiency — having the right inventory in the right locations.

Anonymized web search data has proven to be helpful for other forecasts as well since online activity often is a good leading proxy for purchases and actions of the general public. Having the additional data is insufficient on its own. Processing search data and combining it with traditional sources is vital in creating a successful prediction: We found that raw search query volume is insufficient in parsing out the signals that correlate to true product demand.

Being intelligent about which signals to draw from big data requires care and best practices can be case specific. For example, single queries from a user might be less important that multiple queries from a user. Although we used search data in this case study, a firm could just as easily use the location of users visiting their website or by linking detailed sales data to a customer's location.

Improved pricing. Using a single price is economically inefficient because part of the demand curve that could be profitably served is priced out of the market. As a consequence, firms regularly offer targeted discounts, promotions, and segment-based pricing to target different consumers. E-commerce websites have a distinct advantage in pursuing such an approach because they log detailed information on customer browsing, not just the goods they end up purchasing, and aggressively adjust prices over time. These price adjustments are a form of experimentation and, jointly with big data, allow firms to learn more about their customers' price responsiveness.

Offline retailers can mimic e-commerce's nuanced pricing strategies by tracking consumers through smartphone connectivity and logging which customers enter the store, what type of goods they looked at, and if they purchased or not. Machine learning applied to these data can algorithmically generate customer segments based on price responsiveness and preferences, which generally offers a large improvement on traditional demographic-based targeting. Our experience with pricing advertising on the Bing search engine is that using big data can produce substantial gains through better matching advertisers to consumers. The success of algorithmic targeting has been well documented and is a key driver of revenue in online advertising market. Advances in measurement technology increasingly allow offline firms to benefit from these types of gains through more efficient pricing.

Predictive maintenance. Smoothly operating supply chains are vital for stable profits. Machine downtime imposes a cost to firms due to forgone productivity and can be particularly disruptive in both complex manufacturing supply chains and consumer products. Executives in asset-intensive industries often state that primary operational risk to their businesses is unexpected failures of their assets. A wave of new data generated by the "internet of things" (IoT) can provide real-time telemetry on detailed aspects of production processes. Machine learning models trained on these data allow firms to predict when different machines will fail before they fail.

Airlines are particularly interested in predicting mechanical failures in advance so that they can reduce flight delays or cancellations. Microsoft data scientists from the Cortana Intelligence Suite team are able to predict the probability of aircrafts being delayed or canceled in the future based on relevant data sources such as maintenance history and flight route information. A machine-learning-based solution based on historical data and applied in real time predicts the type of mechanical issue that will result in a delay or cancellation of a flight within the next 24 hours, allowing the airlines to take maintenance actions while the aircrafts are being serviced and thus prevent possible delays or cancellations.

Similar predictive-maintenance solutions are also built in other industries — for example, tracking real-time telemetry data to predict the remaining useful life of an aircraft engine, using sensor data to predict the failure of an ATM cash withdrawal transaction, employing telemetry data to predict the failure of electric submersible pumps used to extract crude in the oil and gas industry, predicting the failures of circuit boards at early stages in the manufacturing process, predicting credit defaults, and forecasting energy demand in hyper-local regions to predict the overload situations of energy grids. Machine learning will make supply chains less brittle and reduce the effects of disruptions for many goods and services.

These cases help highlight a few general principles:

1) The value derived from the analytics piece can greatly exceed the cost of the infrastructure. This indicates there will be strong growth in big data consulting services and specialized roles within firms.

2) Big data is less about size and more about introducing fundamentally new information to prediction and decision processes. This information matters most when existing data sources are insufficient to provide accurate or actionable predictions — for example, due to small sample sizes or coarseness of historical sales (small effective regions, niche products, new offerings, etc.).

3) The new information is often buried in detailed and relatively unstructured data logs (known as a "data lake"), and techniques from computer science are needed to extract insights from it. To leverage big data it is vital to have talented data engineers, statisticians, and behavioral scientists working in tandem. The term *data scientist* is often used to refer to someone who has these three skills. But in our experience single individuals rarely possess all three.

Radically new applications. The cases that we've discussed concern how big data can be employed to improve existing processes (e.g., more precise demand forecasts, better price sensitivity estimates, better predictions of machine failure). But it also has the potential to be applied in ways that disrupt existing processes. For example, machine learning models taking massive data sets as inputs coupled with clever designs that account for patient histories have to the potential to revolutionize how certain diseases are diagnosed and treated. Another example involves matching distributed electricity generation (e.g., solar panels on roofs) to localized electricity demand, unlocking huge value by equating electricity supply and demand with more efficient generation.

The value described from predicting demand more accurately, better pricing, and predictive maintenance are the specific use cases which easily justify large firms' investments in big data infrastructure and data science. These uses are likely to drive value of the same order of magnitude as the investments. The value of radically new applications is challenging to understand ex ante and speculative by nature. It is reasonable to expect losses for many firms due to uncertain and higher risk investments with a few firms earning spectacular profits.

Jacob LaRiviere is an economist at Microsoft Research, an adjunct professor at the University of Tennessee, and an affiliate faculty member at the University of Washington.

Preston McAfee is a corporate vice president and the chief economist at Microsoft.

Justin Rao is an economist at Microsoft Research and an affiliate faculty member at the University of Washington.

Vijay K. Narayanan leads the Algorithms and Data Science Solutions unit of the Data Group at Microsoft.

Walter Sun is the founder of Bing Predicts and a partner data scientist at Microsoft. He is an affiliate faculty member of the University of Washington and an adjunct professor at Seattle University.