Exploring the Increased Significance of Third Party Data within Big Data

Written by Richard Boire at Boire Analytics

The data science industry is consumed by ever-growing amounts of data. Sometimes what gets overlooked when talking about the notion of Big Data is that, it is comprised of two components: first-party data and third-party data. First-party data represents information gathered directly by an entity that intends to use it. While first-party data are valuable, third-party data are the more significant components of Big Data. Third-party data consists of external data that are unrelated to a given company and all its stakeholder interactions.

In the early days of data science, geographic data was the exclusive source of third-party data. At the time, various government institutions like the U.S. Census Bureau and Statistics Canada were the only organizations that compiled these data. Third-party data can be a very powerful resource for data scientists. My very first predictive model at one of the leading-edge financial institutions is a great example of this. For this acquisition model, I was able to extract names and addresses from a list source and append demographic information from Stats Can, using postal codes as the link.

The acquisition model considered demographics, such as income, education, ethnicity, etc., as potential model variables and it performed well, yielding a 50 percent improvement in performance. While extraordinary at the time, it wasn't really surprising given this was the first time that a predictive model was used in a marketing campaign by this financial institution.

Most businesses today now understand the tremendous value of predictive models, as demonstrated by the explosive demand for data scientists in virtually all sectors. But the Internet and digital evolution have simply expanded the availability of data sources and ultimately third-party data. The business applications are limitless.

Sports is an obvious place to start. Let's say I'd like to help the Toronto Maple Leafs use data more effectively such as looking at advanced player stats and team standings. These kind of data can be used to determine salaries, what players to utilize during certain game situations as well as other lesser-known factors that drive wins. Michael Lewis' book "Moneyball" provides another great illustration of how data are used in baseball. Lewis in his book discusses the finding garnered from baseball statisticians that an attempt to steal second base with nobody out tended to produce fewer runs than having the base runner remain at first base. In virtually all sports now, access to this open source data has created a growth industry for analytics practitioners.

In insurance, digital platforms are emerging for consumers, which empower the consumer to seek the best policy available given the price and the resulting desired benefits. This is quickly transforming the nature of the insurance broker or sales agent where the so-called "middleman" role is now being replaced by a platform. Yet, for data scientists, these digital platforms are collecting data about consumers, which can be used by a variety of different insurance providers. One good example of this is the development of acquisition models in targeting consumers from these platforms who are not already existing policyholders.

The proliferation of telematics into our everyday lives represents another source of information that is disrupting the way we evaluate insurance risk by looking at driving behavior. Historically, insurance companies gathered data about drivers when they applied for coverage, which largely consisted of basic information such as age, gender and miles driven to work and/or for pleasure. Today, telematic devices within the car now record the actual observed driving behaviour such as speed, quickness of turns, brake speed and more, which all represent valuable information when building models to predict the

likelihood of a claim. This observed driving behavior, as well as the fact that the information reflects recent behavior, can provide much more powerful inputs when trying to build insurance risk models. This type of driving information is far superior to the information that was typically collected at the time of application.

Internet of Things (IoT) data represents another form of third-party data that didn't exist until recently. As devices become connected, they can be used for many different business applications. Wearable technology now collect data about our bodily reaction to certain events. This can provide insurance and health practitioners with tremendous inputs into models that are trying to predict a certain health condition. But these types of data, which also look at how various machine processes are interconnected, are more commonly used in the prediction of machine defects.

Weather data are another source of third-party data being used by practitioners. Think of the retail industry where knowledge of the weather can be used to make marketing decisions on what items to display at a given store. At the same time, this information can used to staff resources at their stores given the expected store traffic due to certain weather conditions.

Perhaps the most significant new source of third-party information are mobile data. As each one of us are now tethered to our mobile phones, these devices now act like pseudo loyalty cards. The obvious concern is around the issue of data governance which in this case is about transparency to the consumer when giving consent. This implies that the consumer be informed on how the data is being collected, how it is being used, as well as ensuring that it is only used for that purpose. In dealing with this concern, many organizations will obtain consumer consent directly through a marketing campaign where an app is downloaded that allows the capture of this information.

Among the many uses of data collected from smartphones is that it can be used to observe how a person is moving through a specific location, including how that person is moving through a store. Valuable visit behavior information related to recency, frequency, and duration within a given location can all be created as inputs to a predictive model to help drive more consumer engagement. At a more granular level, this information can help a retailer understand which products consumers walk by on a more regular basis. Marketing offers can then be offered in real-time based on the consumers' location within the store.

This open source philosophy has democratized the availability of third-party data. Yet, as analytics and data science practitioners, we also need to consider the data governance issues that arise such as respecting the data privacy concerns of the consumer. The key for organizations in using third-party data, particularly at an individual level, is to be proactive in obtaining consumer consent. The obvious need for consent just makes common business sense as organizations would want to target and engage with consumers who have proactively given their consent.

The rise of social media has accelerated the concern for data governance as individuals leave their digital footprint through their use of multiple social media platforms. This digital footprint can be used as third-party data by organizations assuming the right data governance protocols are in place. This just enhances the level of understanding and any resulting predictive analytics solutions about consumers, which a given organization would be unable to obtain from first-party data alone. But data security and privacy must be an intrinsic foundation in using any third-party data within our data science solutions. These solutions will only be accepted (or should be, at any rate) by practitioners who adhere to strong data governance protocols.

Finding that balance of data governance and data science is often not an easy process, but one that must be pursued as practitioners continue to push the data science envelope.