Predicting the future, Part 1: What is predictive analytics?

Alex Guazzelli May 29, 2012

You can use predictive analytics to solve your most challenging problems. It helps you discover patterns in the past, which can signal what is ahead. This first article of a four part series focuses on predictive analytics. It starts by looking at analytics in general, and then positions data-driven analytics against business rules and expert knowledge. Both types of knowledge can enhance your decision-making ability. Predictive analytics is able to discover hidden patterns in data that the human expert may not see. It is in fact the result of mathematics applied to data. As such, it benefits from clever mathematical techniques as well as good data. Given that we can apply predictive analytics to a myriad of datasets in different industries and verticals, this article helps you identify a few applications of predictive analytics on your own.

View more content in this series

About this series

This is Part 1 of a four part series on predictive analytics. Part 1 gives a general overview of predictive analytics. Part 2 will focus on predictive modeling techniques, which are the mathematical algorithms that make up the core of predictive analytics. Part 3 will then put these techniques to use and describe the making of a predictive solution. Finally, Part 4 will focus on the deployment of predictive analytics, that is, the process of putting predictive solutions to work.

Introduction to analytics

Today we live with an ever-expanding sea of data. To navigate it safely, we use analytics. Without analytics, we would simply drown, not really knowing what happened or what will happen next. In this article, we focus on the knowledge obtained from analytics, which we may classify as descriptive or predictive. While descriptive analytics lets us know what happened in the past, predictive analytics focuses on what will happen next.

Our need to understand past events has led to a discipline that we now call *business intelligence*. It allows us to make decisions based on statistics obtained from historical data. For example:

1. How many customers have churned or defected due to attrition in the last six weeks?

© Copyright IBM Corporation 2012

2. How much money was lost due to fraud in the last three months?
3. How often are support tickets being created?
4. Where are customers located (maybe shown using Google maps)?

Descriptive analytics goes a long way in allowing for sound business decisions based on facts, not feelings. However, descriptive analytics is simply not enough. In the society we live in today, it is imperative that decisions be highly accurate and repeatable. For this, companies are using predictive analytics to literally tap into the future and, in doing so, define sound business decisions and processes.

As a discipline, *Predictive Analytics* has been around for many decades. A hot topic in academia for many years, its relevance in industry increased together with the amount of data being captured from people (for example, from on-line transactions and social networks) and sensors (for example, from GPS mobile devices) as well as the availability of cost-effective processing power, be it Cloud or Hadoop-based.

Data driven versus expert knowledge

It is fascinating to think of knowledge and how we transfer and use it. Traditionally, we counted on domain experts to help us get the most out of a particular process. Expert knowledge is based on experience and is used everyday by all companies to influence day-today operations. Given how we can translate expert knowledge into a set of business rules, we've built decision-based systems to automatically apply the knowledge elicited from human experts. IBM ILOG is a prime example of a system that translates expert knowledge into a set of IF-THEN statements that we can put to work right away.

On the other hand, data-driven knowledge, as its name suggests, is based upon data—usually, lots of it. A few decades ago, a series of statistical techniques emerged with the intent of uncovering data patterns typically hidden to the human eye. Given that we capture data in an ever-increasing volume today, these techniques are proving indispensable to extracting value from data, making processes repeatable and accurate.

The movie *Moneyball* exemplifies that really well. In the movie, a group of experienced recruiting agents offer their first-hand knowledge and hunches on which players should be pursued to be part of the team. That is contrasted with a data-driven approach in which knowledge is extracted from the data already available for each player, and a team assembled from that. Although *Moneyball* chooses one type of knowledge over another, in most cases, we should and do use expert knowledge and data-driven knowledge together.

Analytics is able to produce sound statistics, predictions, and scores. It is up to a rules-based system, however, to decide on what to do with all that data-driven knowledge. For example, we may use a series of rules to trigger business decisions depending upon the output obtained by a predictive model. For example, if a model exists to predict the risk of customer churn or defection, we may put rules known to mitigate churn in place to define specific business decisions according to different risk levels. Therefore, if risk is high, we may give a customer a 20% discount on his or her next purchase, but if risk is very high, we may give a 50% discount instead.

What is a predictive model?

A predictive model is simply a mathematical function that is able to learn the mapping between a set of input data variables, usually bundled into a record, and a response or target variable.

We refer to this learning as *supervised* because, during training, data is presented to a predictive model with the input data and the desired output or outcome. Training is repeated until the model learns the mapping function between the given inputs and desired output. Examples of predictive models using supervised learning include back-propagation neural networks, support vector machines, and decision trees. A predictive model may also use unsupervised learning. In this case, it is only presented with the input data. Its task is then to figure out how different input data records relate to each other. Clustering is the most commonly used type of predictive models, which uses unsupervised learning.

So, as an example, imagine that you want to create a predictive model that will be able to tell who among your customers is most likely to churn (20 or 50 percent discounts anyone?). You first go back to your historical data in search of features that you could use to build a model to do so. By looking at your database, you are able to compile a list of attrition-related features for both existing and past customers that churned. It may include the number of complaints in the last 6 months, the number of support tickets opened in the last 4 weeks, how often and how much money the customer spent buying merchandise or services (on-line or in-store), and generic information such as age, gender, and demographics. Figure 1 shows two such customers together with the features obtained for each of them. Customer 1 is an existing customer and seems to be satisfied. Customer 2, however, has churned.

Figure 1. Two customers and their input features.

Customer 1
No complaints in last 6 months
Opened 1 support tickets in the last 4 weeks
Spent a total of $9,876 buying merchandise
Spent a total of $987 in services
Purchased 12 items in last 4 weeks
Is 54 years old
Is a male
Lives in Chicago
...

Customer 2
3 complaints in last 6 months
Opened 2 support tickets in the last 4 weeks
Spent a total of $1,234 buying merchandise
Spent a total of $123 in services
Purchased 2 items in last 4 weeks
Is 34 years old
Is a male
Lives in Los Angeles
...

In a supervised learning type of scenario, as shown in Figure 2 you present all your customer data to a predictive technique during training. In this case, the input is comprised of all the features you came up with (satisfaction-related, demographic, and so on) for each customer as well as the associated outcome. The outcome tells the predictive model if the data record represents a customer who did or did not churn. The rationale here is that the model is able to learn the

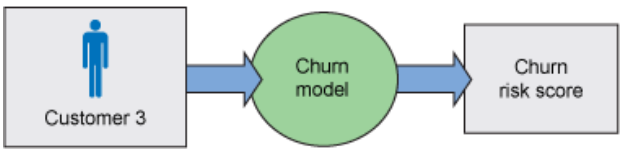
differences, or patterns, between the two groups: existing satisfied customers and customers that defected.

Figure 2. Customer data comprising of input features and outcome is presented to a predictive model during training

After we build a predictive model, we need to validate it. The question validation tries to answer is two fold: "Does it work?" and if so, "How accurate is it?" If the answer to the first question is a resounding *yes* and the answer to the second question is *highly accurate*, you know your model works and that it is able to generalize well. All you need now is to make it available for execution. For that, it needs to be operationally deployed. Luckily, a standard called PMML (Predictive Model Markup Language) exists that allows predictive models to easily move between different systems. With PMML, we can use an application such as IBM SPSS Statistics to build and validate a predictive model that we can then save as a PMML file. As such, we can directly upload it into a scoring engine such as the Zementis ADAPA engine where it is ready to use in real-time. For more information about PMML and the technologies describe here, see Related topics.

After deployment, we can use the churn model to monitor all existing customer activity. A good predictive model is able to generalize its knowledge to compute the churn risk even for customers it has never encountered before. Figure 3 shows the data for one such customer, identified as customer 3, being presented to our churn model. If the model perceives that a pattern of churn is emerging for a particular customer, it will increase its risk or output accordingly until there is a business decision on your part to address it. When that particular customer is again satisfied with your products and services, the risk will diminish, since the churn pattern is no longer detected.

Figure 3. When operationally deployed, the newly created churn model is used to score new and existing customers to compute a churn risk score for each customer. If a high churn risk is detected, procedures may be put in place to mitigate it.



The importance of good data

My first impulse in naming this section was to call it "The importance of data" because without data, there is no analytics and for that matter, predictive analytics. During my career as a data-mining scientist, I have seen many well-intentioned business folks in search of a predictive analytics solution for their company. While they know that predictive analytics can help their bottom line, they have insufficient data. That is, not enough data for a data scientist to actually train a

model that makes sense. In an era of Big Data, you may be surprised how often that actually happens. For certain predictive models to be able to learn and generalize, it takes thousands and thousands of records. In line with our example above, a hundred or so records containing data for customers that churned in the past may not be enough. If not enough data is used for training, a model may not be able learn or worse, it may over fit. That means that it learns everything about the given data during training, but it is incapable of generalizing that knowledge when presented with new data. It is simply unable to predict.

If enough data is available, it is then a question of how good it is. That's because, the quality of the data will directly reflect the quality of the model. To put it mildly: garbage in, garbage out!

Years ago, my team and I were tasked to build a model for predicting whether a certain manufacturing process was going to result in failure or success. The manufacturing took approximately eight hours to complete and consumed a great deal of resources. Only after completion were the company's quality assurance engineers able to determine if something went wrong during manufacturing. If so, the entire batch had to be scrapped, and a new one started again from scratch. The idea here was that we could look at data obtained in the past for all the stages of the process for batches that turned good and bad. Then, we could train a model to detect when things start to look bad early on in the process. As exciting as it was, we never actually built any models. During data analysis, we found out that the data was corrupted or worse, missing for some of the key manufacturing phases. More importantly, we could not find the outcome, or the information that would allow us to separate good from bad batches. That means we would not be able to use a supervised learning technique. And, missing important parts of the input data jeopardized the use of an unsupervised technique.

Data by itself does not translate to predictive value. Good data does.

Applications of predictive analytics

When first learning about predictive analytics I received a book by Duda, Hart, and Stock entitled *Pattern Classification* (see Related topics). This book is now considered a classic reference in the field. In it, the authors build a pattern classification system around a fish processing plant. In such a plant, they use a predictive solution to classify incoming fish into salmon or sea bass depending on input features such as length and intensity of scales. In 2010, I gave a presentation at the Rules Fest Conference in San Jose on predictive analytics. In the presentation, entitled "Follow your Rules, but listen to your Data" (see Related topics), I used the same example to show the rules-focused audience not only how we can solve a problem like this using predictive analytics, but also how predictive analytics can work together with business rules to improve decision making. The idea obviously was to use the example in a similar way as Duda, Hart, and Stock. That is, as a generic example of how to build and apply a predictive solution and let the audiences generalize it to other applications. In this article, I used customer churn instead. In any case, so that you do generalize the knowledge you obtained so far to a host of new applications, I will describe below more ways in which predictive analytics is transforming our world into a smarter place.

An application of predictive analytics that has been extremely successful for many years now is fraud detection. Every time you swipe your credit card or use it on-line, chances are your

transaction is being analyzed in real-time for its likelihood of being fraud. Depending on the perceived risk, most institutions implement a set of business rules that may even decline a high-risk transaction. This is the ultimate goal of predictive analytics in the fight against crime, that is, preventing it from happening in the first place.

In an article written earlier for developerWorks, I list a few important applications of predictive analytics in healthcare. Medicare fraud is most definitely on top of the list, but so is the use of predictive analytics to implement effective preventive care. By knowing which patients are at a higher risk of developing a certain disease, we may put preventive measures in place to mitigate risk and ultimately save lives. Lately, predictive analytics has been the center of attention on a highly publicized contest in which historical claims data is used to reduce the number of hospital readmissions (see Related topics).

Additionally, companies use predictive analytics to recommend products and services. Nowadays, we have already grown to expect good recommendations for movies, books, and songs from our favorite stores and merchants. By the same token, we are also experiencing marketing campaigns that are tailored more and more to our tastes and preferences based, for example, on the content of our emails, on-line postings and searches.

Other applications focus on data obtained from sensors. For example, we can use GPS mobile device data to predict traffic. As these systems become increasingly precise, we will be able to use them to alter our own transportation choices. For example, we might take the train one day if the road is predicted to be completely clogged with cars.

Furthermore, the availability of small and cost-efficient sensors that report on the current status of structures such as bridges and buildings as well as machinery such as energy transformers, water and air pumps, gates, and valves has enabled the use of predictive analytics to maintain or make changes to materials or processes before faults and accidents happen. By enabling the building of predictive maintenance models, the use of data from sensors is a clear way towards helping to ensure safety. The oil-spill disaster in the Gulf of Mexico in 2010 and the collapse of the I-35W Mississippi River bridge in 2007 are only two examples of major accidents that could be prevented if sensors and predictive maintenance models had been in place.

Conclusion

In an ever-expanding sea of data, collected from people and sensors, predictive analytics provides essential navigational tools for companies and individuals to successfully reach their destination. It does that by forecasting what is about to happen so that one can respond appropriately to stay on the most accurate, safe, repeatable, profitable, and efficient course.

The use of predictive analytics is already revolutionizing the way we interact with our environment. As the quantity of data increases and the quality improves, aided by the availability of cost-efficient processing power, predictive analytics is bound to be even more pervasive than it is today. If you already identified a few problems you plan to address with predictive analytics, you will agree that this was not a difficult prediction to make.

Related topics

- Follow your Rules, but listen to your Data: Watch Alex Guazzelli's presentation at the Rules Fest 2010 Conference which focuses on the differences between data-driven and expert knowledge as well as the benefits of bringing the two together.
- Predictive analytics in healthcare (Alex Guazzelli, developerWorks, November 2011): Read this article on the challenges and applications of predictive analytics in healthcare.
- The Heritage Heath Prize: Find out more about the highly publicized contest that aims to identify who will be admitted to a hospital within the next year, using historical claims data.
- What is PMML? (Alex Guazzelli, developerWorks, September 2010): Read this article on the PMML standard used by analytics companies to represent and move predictive solutions between systems.
- Predictive Analytics: Read the Wikipedia page on predictive analytics for an overview of common applications and techniques used to make predictions about the future.
- PMML in Action (2nd Edition): Unleashing the Power of Open Standards for Data Mining and Predictive Analytics (Alex Guazzelli, Wen-Ching Lin, Tridivesh Jena; CreateSpace, Jan 2012): Learn to represent your predictive models as you take a practical look at PMML.
- The Data Mining Group (DMG) is an independent, vendor led consortium that develops data mining standards, such as the Predictive Model Markup Language (PMML).
- Zementis PMML Resources page: Explore complete PMML examples.
- Data Mining: Find more about this topic in Wikipedia.
- PMML discussion group: Join this LinkedIn group.
- IBM ILOG: Learn more about this recognized industry leader in Business Rule Management Systems (BRMS), visualization components, optimization and supply chain solutions that enriches the IBM software portfolio and fortifies the IBM Smarter Planet initiative.
- IBM SPSS Statistics 20 puts the power of advanced statistical analysis in your hands. Whether you are a beginner or an experienced statistician, its comprehensive set of tools will meet your needs.
- Try the IBM ILOG CPLEX Optimization Studio 90-day trial: Rapidly develop optimization-based decision support applications.
- Evaluate IBM WebSphere Application Server: Build, deploy, and manage robust, agile and reusable SOA business applications and services of all types while reducing application infrastructure costs with IBM WebSphere Application Server.
- developerWorks podcasts: Listen to interesting interviews and discussions for software developers.

© Copyright IBM Corporation 2012
(www.ibm.com/legal/copytrade.shtml)
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)