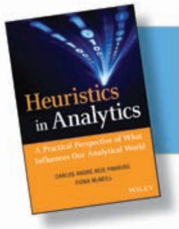
The Heuristics in Analytics

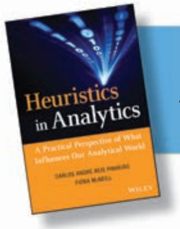
A Practical Perspective of What Influences Our Analytical World

CARLOS ANDRE REIS PINHEIRO

FIONA McNEILL



Contents


Heuristics in Analytics
A Practical Perspective of What Influences Our Analytical World
WILEY

CHAPTER 1
Introduction

Analytics is used to address many different types of business problems. It is used to understand customer behavior and how consumers may be adopting new products and services. It is used to describe different marketplace scenarios and their impacts. It is also used to decipher competitor's movements and patterns. And it is used for predicting potential future revenue, detecting risk, uncovering fraud, minimizing bad debt, optimizing operational processes, and more. Analytics is used in all of these business applications.

In most cases, and in particular in customer scenarios, there are many factors that cause misunderstanding of what is currently happing within a market, or even what is happing with a specific customer within a particular market.

It is always important to bear in mind that consumers present different types of behaviors and in accordance with the market they are interacting with. As a customer, I can be very aggressive in terms of purchasing high-tech products, often buying cutting-edge gadgets. However, I am quite conservative in terms of investing, putting my money into low-risk accounts. There is no one, overall general behavior for any customer. We each behave in different ways depending upon the situation in which we find ourselves. Essentially, we wear different hats, having distinct behaviors that are observable—each in relation to the distinct roles we play. And sometimes, even in similar scenarios, we may play different roles and exhibit different behaviors, depending on the other scenario actors that are involved.

All analytical models, whether they are supervised (classification), semi-supervised, or unsupervised (segmentation), take into consideration most of the structured information that companies currently hold in their databases. They include information about customer characteristics, the products and services that they offer, and how customers interact with them. They include financial inputs such as credit rating, payment history, late payment durations, and so forth. All of this information, however, describes only a limited part of the end-consumer's behavior. In other words, we really can't say too much about an individual customer's profile, but we can describe how they behave in a single scenario, like when using the company's products and services.

You could say that, based on my historical data, I am an aggressive buyer of high-tech gadgets. And it is just as possible to state, based on my buying history, that I work hard to purchase high-tech products in advance. But this behavior doesn't replicate to other situations, like my conservative financial investment behavior. I also might be very price sensitive regarding telecommunications services, but on the other hand, I may not be sensitive to aspects such as quality when it comes to hotel rooms. The important thing to keep in mind is that there isn't an overall understanding of behavior. Instead, behavior is always in relation to something, to hotels, financial investments, or telecommunication preferences.

Consider for a moment, understanding and even predicting behavior in a telecommunications scenario. Most analytical models consider call frequency and duration, demographic information about the consumer, billing and payment history, and—when available and in reasonable quality—historical contacts with customer care channels. Based on such data, companies are able to build the majority of the analytical models used to examine common business issues such as product/service cross-sell or up-sell, churn, revenue leakage, risk specification, or fraud detection. Furthermore, for classification problems (that is, the ones that use a target class for training), historical information is quite crucial in that it teaches the model which behavior is most highly relevant to that particular event. What are the main characteristics for all customers who bought some particular product? How do they behave before this purchasing event? Which variables were most relevant to describe the business event or triggered it? Historical data, when it is in relation to a particular business event, teaches the analytical model to foresee the likelihood for each customer to perform when exposed to a similar event.

However, this is a purely mathematical approach. Even more specifically, it is a purely statistical approach. The analytical model teaching, also called the *training process*, is based on the average customer's behavior. However all customers with similar past behavior will not proceed in the same way, will not purchase the same product, will not consume the same service in the same way, and so on.

For example, according to my past behavior, and as represented in my historical data, I might be about to purchase a particular bundle of telecommunications services because customers who have been behaving like me have bought this bundle in the past, after a similar sequence of events. So, it is quite reasonable for any company to think that now is my turn. Then, the week that I'm going to buy that bundle approaches. Most unfortunately, one special Sunday afternoon, my soccer team lost the derby. It was the final match of the championship, and we lost to our biggest rival. So instead, my forthcoming week is a sequence of five long days of frustration from the loss, and I'm certainly not in the mood to buy anything. Instead, I hide myself and simply wait for time to move on. This completely external event was not considered by the model and yet has changed everything to do with the accuracy of my predicted behavior. Statistically I should have purchased the bundle that week, and the likelihood of it would be around 87 percent. Unfortunately, the analytical model didn't take into account that possible result for the final match. And with great sadness, this particular variable—the result of the final match—was indeed most relevant in my actualized behavior. It is the single factor that made all difference in my buying something or not.

These external influences happen all the time in our lives. Very often they impact analytical models, especially those that are defined for business purposes. It is not possible to consider all variables, all attributes, all information required to create a particular inference. Everything in modeling is about an approximation. As my historical behavior was quite similar to other customers who did buy that particular bundle, my likelihood of purchasing the same bundle might also be very high. But it isn't a definite or a sure thing that I will buy it at all. It is just an approximation. It might be a highly accurate approximation, but in the end, it is just a simple approximation. The likelihood assigned to each customer is simply an approximation of how they might eventually behave in relation to a particular business event.

This fact shouldn't push us to give up on analytical endeavors. As a matter of fact, it should do just the opposite. It hopefully brings us even closer to understanding the true value of analytics. Unexpected events will always take place, and they will always affect predicted outcomes. Analytical models work well for the majority of cases, for most of the observations, and in most applications. Unexpected or uncontrolled events will always occur and typically affect a few observations within the entire modeling scenario. However, there are some events that will impact the entire analysis population, like a war, an earthquake, or a hurricane, and as such, a new historical behavior is built.

Analytical methods that understand the past and that are prepared to explain present circumstances do provide forecasts into the future that improve business decisions. Other books describe the value that analytics provides.[1] This book is different. It examines the unimagined and unforeseen events that impact analytic results, describing the art of analytics, which is founded in the science of mathematics and statistics.

The formula to predict a particular event works a lot like the standard conditions for temperature and pressure in chemistry. If everything is right, if the temperature is in the expected range, as well as the pressure, then the formula forecasts the outcome quite well. While we certainly can have exceptions, the formula is just a way to model a particular scenario and be aware of what is coming next, and what could be expected in certain conditions. Likewise in science, and several other disciplines, this approach is the closest we can get to being in touch with reality. It is much more enlightened than doing nothing. The key is to properly understand what is happening in order to dramatically increase your model's usefulness.

[1] See for example, Stubbs, E. *The Value of Business Analytics: Identifying the Path to Profitability*, (Hoboken, New Jersey) Wiley & Sons, 2011.

THE MONTY HALL PROBLEM

The Monty Hall Problem is a very good example of how important it is to be well aware of activities in the marketplace, the corporate environment, and other factors that can influence consumer behavior. And equally important, it illustrates how critical it is to understand the modeling scenario in order to predict activities and events. In order to increase your chances, particularly from a corporate perspective, it is important to understand the equations that, at least with reasonable accuracy, explain the scenario under investigation—even if that scenario is largely composed of a situation that is dependent on chance. It is then possible to at least create an expected outcome of a particular scenario, whether it be probabilistic (based on historical information of past events) or stochastic (as would be with a sequence of random activities). Breaking this down, if you are going to flip a coin, and you are about to bet based on it, you should know your chances of winning the toss are about 50-50, no matter whether you choose heads or tails. Although this sounds quite simple and straightforward, companies don't do this very often. Companies typically do not prepare themselves for upcoming events—gambling even more than they should. The Monty Hall Problem is a case that illustrates this notion, that the knowledge about the scenario and the chances involved make all the difference between winning and losing.

Let's Make a Deal is a television game show originally produced in the United States and thereafter popularized throughout the world. The show's premise is to have members of the audience accept or reject deals offered to them by the host of the show. The members of the audience who participate in the game usually have to analyze and weigh all the possibilities assigned to a particular offer described by the host. This offer could be a valuable prize or a totally useless item. Monty Hall was the famous actor and producer who served as the host for this game show for many years.

The Monty Hall Problem is in fact a probability puzzle, considered to be a paradox because although the result seems implausible it is statistically observed to be true. This problem was first proposed by Steve Selvin, in a letter to the *American Statistician* in 1975.[2] This

[2] *American Statistician*, Vol. 29, No. 1 – 1975 "Letters to the Editor": http://www.jstor.org/stable/2683689

problem was published again in *Parade Magazine* in September 9th, 1990, within the Sunday "Ask Marilyn" column, on page 16.

> "Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say no. 1, and the host, who knows what's behind the doors, opens another door, say no. 3, which has a goat. He then says to you, "Do you want to pick door no. 2?" The question is: Is it to your advantage to switch your choice?

In her response, Marilyn vos Savant said that the contestant should always switch to the other door and by doing so, she explained, that the contestant was going to double his or her chances. She was flooded with more than ten thousand letters from all over the country, including letters from five hundred PhDs, mathematicians, and statisticians associated with revered and prestigious universities and institutions. The majority of the letters expressed concerned with Savant's mistake, and some of them even asked her to confess her error in the name of math. Intuitively, the game concept leads us to believe that each door has a one-in-three chance to have the car behind of it, and the fact that one of the doors has been opened does change the probability of the other two. Once one door is opened, the chances of having a car behind one of the other two doors are one in two for each one left. When one door is opened it is like we have essentially changed the original scenario, from three doors to just two. But in reality, there is no changing the probabilities for each door, either before or after the first door has been opened. All doors keep the same one-in-three chance of having a car behind of it, both in the occasion of the first question (three doors closed) and on the occasion of the second one (two doors closed and on opened).

Marilyn vos Savant stated that if the car is equally likely to be behind each door, a player who picked door number one and doesn't switch, has a one-in-three chance of winning the car, while a player who picks door number one and does switch has a two-out of-three chance. In this way, the players must switch the door to double their chances. It is quite simple, isn't it?

The Monty Hall Problem demands some assumptions. The first one is about fairness. The car should have an equally likely possibility of being behind any of those three doors, and the contestant can initially choose any door. Monty will always open another door, giving the contestant the opportunity to switch from their initial door choice. And finally, and mostly important, Monty knows what is behind the doors—and will always open a door that reveals a goat. If Monty revealed the car the game would otherwise be over. By revealing the goat behind the door, Monty doesn't alter the chances of the remaining doors to 50-50. Actually, and from the beginning, each one of the other two doors not chosen has a one-in-three chance and, therefore, the sum of the remaining doors is two out of three. The fact that Monty opens a door he knows has a goat behind it puts all two thirds of the chances to win in the door not yet picked by the contestant.

You can also understand this problem by calculating the probability of losing. Switching the door makes you lose only if you have previously picked the door with the car behind. And initially, you have one in three chances to pick the car. Hence, if you switch the door, you have a two in three chance to not lose (1/3 for each remaining door).

To finish this up, let's think about the details of this problem illustrated in a table showing all possible arrangements of one car and two goats behind a total of three doors, as shown in Table 1.1. Suppose you pick door no. 1. In the first row of the table, you can see that if you switch you lose and if you stay you win. In the second row, if you switch you win and if you stay with your original choice, you lose. Finally, in the third arrangement, if you switch you win and if you stay you lose. By choosing door no. 1, if you keep the door and stay with that choice, you win one time. But if you switch you have a chance to win two times. Analogous calculus can be made for doors 2 and 3. At the end, by switching the door you should double your chances to win. Table 1.1 shows the options and the chances in keeping or switching the doors.

Table 1.1 Itemized Choices in the Monty Hall Problem

Door 1	Door 2	Door 3	Switch	Stay
Car	Goat	Goat	Goat	Car
Goat	Car	Goat	Car	Goat
Goat	Goat	Car	Car	Goat

The point is that in spite of this particular problem being about gambling, by knowing the proper theory and by understanding the correct way to calculate and model in the scenario, you can double your chances of winning. The same thing happens with analytics. In spite of the heuristic influence over real-world events, such as the loss of my soccer team on that awful Sunday, by knowing the proper way to model the business scenario you can substantially increase the chances of applied model success. In theory, if a company offers a particular product to a customer, the chances that he or she will get that product should be 50-50. But if the company knows about this particular customer, their past behavior, the current behavior for similar profiles, products, and services held by similar customers, and so on, then the company could substantially increase the chance of succeeding in this selling process.

EVOLVING ANALYTICS

Over the past few years, analytics has evolved into a major discipline for many corporations, particularly those exposed to a highly competitive marketplace. Analytics is composed of a set of distinctive approaches that use technologies and methods to describe business processes, to support decision making, to solve issues, and to help with improving performance.

Analytics can involve simple business intelligence applications, such as predefined multidimensional databases that are queried, to more advanced interactive visualizations of data across different dimensions, allowing analysts to glean insight from large volumes of data. These types of business analytics tools can be applied to single, built-to-purpose applications all the way up to large data marts that are used to examine questions in particular department(s). They can also be associated with a corporate data warehouse—as either a collection of multidimensional databases or as a single, high-performance business intelligence platform from which multipurpose tools are used to examine business problems across the whole company. Typically, these types of applications are associated with raw data collected from transactional systems, aggregated based on a particular business perspective, and examined through a user friendly front-end—allowing analysts to execute several types of queries in a short period of time. Business insights very often come from such multidimensional examinations of data as deep inquiries of operational inputs—formatted in a digestible, managed layout.

Analytics also goes a step further to include statistical and mathematical analyses. These types of analytic endeavors are often completed on a more ad hoc basis, demanding timely examination associated with a particular business purpose. And usually, this type of analysis is not periodic nor does it demand frequent reprocessing of the task. It oftentimes does, however, require output monitoring to foresee if changes are occurring that require adjustments to the model over time. Such a model could be developed to examine the overall marketplace, a snapshot inquiry about customer behavior, or a forecast study in relation to costs, sales, or market growth, for example. This sort of analysis is often performed by using statistical and/or mathematical procedures and is commonly done in a stand-alone environment. As an ad hoc process it usually does not require the same infrastructure of a production environment. It is a set of tasks and procedures that are used to understand a particular business question and, once deployed, raise useful knowledge that can change operations and activities.

Finally, there is a third layer of analytics that is composed of data mining models. The data mining discipline includes artificial intelligence algorithms such as neural networks, decision trees, association rules, and genetic algorithms, among others. These artificial intelligence methods often use a more mathematical approach and are well suited to specific business issues, like predicting a particular business event from large volumes of inputs. Association rules, also known as induction rules, are very good at describing particular subject relationships, such as in retail. They are used to understand how products are selling together, highlighting the correlations between products to identify grouping, or sell with relationships. Consumers who buy bread, cheese, and ham usually buy butter as well. Consumers who buy red wine and Grana Padano cheese commonly also buy honey and cinnamon. These rules highlight product correlations, some of which might be quite useful, others not. Two metrics affiliated with association rules help us understand how relevant and strong the rule is—namely support and confidence. However, if you run an association algorithm over a hospital database, particularly using data from the obstetric department, a rule would probably emerge that women have babies, with both a high level of support and confidence. Of course, that rule has no business value—and although true, is not useful to any particular business decision.

This brings to light a very common phenomenon in analytics—you have to build a bridge between the technical procedure and the business results. Outcomes from analytics should raise the knowledge about the market, the customers, the products, and so forth, enabling companies to deploy practical, well-informed actions that improve their businesses. The results are beneficial associations and rules are that are not only informative but can also be applied to business actions. Validation of the value to the business derived from analytics-based knowledge should always be performed and encompasses the measured effect that analytics has on activities, events, and processes.

Artificial intelligence (AI) techniques are often assigned to methods that focus on classification or prediction. Commonly, the predictive AI methods are referred to as supervised learning models, given that they demand historical data and a target variable. The target variable is what the model is trying to predict and the historical information is used to train the model in order to predict the target variable. This historic information drives the model's learning process, correlating the past behavior to the target. This learning process allows the model to foresee possible values for the target variable in the future. Artificial neural networks, decision trees, and regression equations are the most commonly used data mining techniques. These types of models usually require less business knowledge validation because they are trained in recognizing patterns in the data. So, once this type of model has its premise, which is the target variable, the outcome results are based on the pattern recognized in the historical information, rather than by some particular type of business knowledge itself. AI tasks are more about the training process and the pattern recognition rather than an overall analysis in traditional statistics (the latter of which doesn't have previous history or a historical premise).

AI classification methods are also pattern recognition procedures—that, instead of predicting a target variable, focus on delineating patterns that describe the data. This type of modeling includes clustering, k-means and self-organizing map methods—all of which require business validation because there is no premise in relation to the model training; that is, no target variable. Therefore, business expertise validation determines if the results are sound—demonstrating that bridge between the technical process and the business knowledge. A classification model essentially looks at the past, learns from it (creating a pattern), and uses it to apply that pattern to data not involved in the model training process. A clustering model usually requires a calculation that identifies the correlation between the main characteristics of the clusters found and the business threats and opportunities. Each cluster holds some set of characteristics which describes the cluster and which lead to business actions, like protecting the company from threats or exploring some hidden opportunities. Customers can also be stimulated to migrate from one cluster to another, say from a medium-value cluster to a high-value cluster, or from a disloyal cluster to a loyal one.

Many business actions can benefit from the insight gleaned from clustering techniques, particularly those that involve understanding groups of customers, products, and services. Distinct marketing campaigns can be based on the characteristics associated with individual clusters giving different incentives and alternate messages to each. A cluster, in and of itself, highlights some trend, type of behavior, or a particular need—some specific grouping driven by the data itself. Identified trends can be used to create a new product, the type of behavior can be used to establish an incentive, and the need can be used to adjust a service. There are many interactions companies can deploy and benefit from by using the subjective and descriptive information emanating from the clustering process.

The Business Relevance of Analytics

Analytical modeling plays a key role in distinguishing companies within any marketplace by driving opportunities for competitive advancement. This competitive advantage enables companies to be more innovative, helping them stand above the others in their market. Analytics helps identify when and what products to launch, informs which services will maintain customer loyalty, and optimizes product and service price, quality, and turnover. But perhaps even more important than being innovative is to maintain innovation—and where analytics becomes crucial. Innovation means solving problems with simple solutions rather than complex ones, in an appropriate amount of time, and using a feasible replication process. In order to sustain innovation companies need to create a suitable environment to identify problems, threats, and opportunities; to create quick solutions that address the issues; and to deploy these solutions into the production chain. If the solution is too complex to be readily understood, too expensive to be deployed into a production process, or if it takes too long to be developed, then this solution certainly isn't innovative. It is just a solution, not feasible, not touchable, and not applicable to the company.

Be simple. Be fast. Be innovative.

Although easy to say, it isn't as easy to do. Companies around the world are trying to do exactly this, although many aren't succeeding. What's the secret? Unfortunately, there is no secret. It's like they say in the cartoon movie *Kung Fu Panda*. The secret of the amazing noodles was . . . there is no secret at all. It was all about the passion, love, and care put into making the noodles. The secret is in each one of us, who work with passion, love, respect, and who put all effort and energy into doing our best. The secret is to look at the past, to observe the present, and to try to foresee the future. So the secret is all about us, is it? Is it just a matter of looking at the past, observing the present, and foreseeing the future? Then why doesn't this approach always work? It doesn't always work because of the heuristic factor. Even though we may proceed on a well-trodden path and follow a typical pattern, there are so many unpredictable variables in our world, each with so many attributes to be considered that even a small change in the overall environment might alter everything.

You can do everything exactly the right way, but perhaps not in the right time. You could do it in exactly the right time and the right way, but perhaps not considering all the variables involved. You can even do everything in the proper way, in the right time, but by not taking into account some small external factor (which is unknowingly completely relevant to the entire model) that may be unmapped, unpredictable, and untracked. This one factor may be something you would never consider being important to your model—it could be a natural event, a political fact, a social change, or an economic disruption. The same approach, by taking the same steps, performing exactly the same routines, would thrive perfectly if done a bit earlier, or even a bit later, but not right now. This is the imponderable! This is something you cannot predict. So what do you do? How can you manage this heuristic world? We have no other tip—but to keep trying to model the scenario at hand.

Organizations need to use the best tools they have available in order to adapt themselves to the scenarios, to both current and future business environments and to both current and pending changes. Analytics will help you to understand all these changes, all these business scenarios, for all corporate environments. And, even when an analytical model fails, the analysis of that failure will help you understand why it happened, why your prediction didn't materialize as expected, and, hopefully, what you would need to do to different next time. Even the failures help you reroute your strategy and drive your company toward the proper trail.

Being innovative is not the destination; it is the journey. In order to be innovative, companies need to steady themselves into this analytical path by monitoring modeled outcomes and improving models over time. Increasing the usage of these models and by converting this whole analytic environment into an operational process that exists across the enterprise drives access to innovation. The organization's strategy should totally direct the analytical environment, and in turn, the analytical environment should totally support the company's strategy.

This relationship between organizational strategy and the analytical environment can be envisioned as three distinct steps. These steps may be performed at different stages of analytics maturity—or may even occur simultaneously. Irrespective of timing, each of these steps resonates with different stages of applied analytic procedures, each of which are aimed at addressing specific business issues, and with particular goals.

1. **Stage One Analytics**. This first layer of analytics provides long-term informational insight, helping organizations analyze trends and forecast business scenarios. Data warehouse,

data marts, multidimensional applications (OLAP—On Line Analytical Processing), and interactive visual analysis usually support this stage one purpose. These inputs support analyses geared toward identifying trends, historical event patterns, and business scenarios. This analysis is concerned with presenting information about past sales by region, branches, products, and, of course, changes that have occurred over time. You can easily replace sales by any other business issue such as churn, subscription, claims, investments, accounts, and so on. Also, you can replace the dimensions region, branch, and product by any business factor you may be interest in analyzing. However, you can't replace the dimension time, which should be always a consideration in this exploratory analytical approach. Often there is a production environment available that readily provides this kind of analytical information in a pre-defined environment, usually through a web portal.

2. **Stage Two Analytics**. A second layer of analytics maps out the internal and external environments that impact the question at hand. This can include market considerations, the customers' behavior and the competitor's actions, as well as details about the products and services that the organization offers. Questions that are explored in this stage include: How profitable are my products/services? How well have they been adopted by the target audience? How well do they suit the customer's need? Statistical analyses support these tasks, with correlations, topic identification, and association statistics methods. Usually in these cases there is an analytical environment available to perform such queries and analyses. However, further distinguishing it from the first stage, there is typically no production environment that provides real-time answers, nor a predefined web portal or any other interface for rapid response to such questions. This stage of analysis is performed on demand, when business departments request deep information about a particular business issue.
3. **Stage Three Analytics**. Finally, the third layer of analytics is driven by to the company's strategy. Model development

> is directed by core business issues such as cross sell, up sell, churn, fraud, and risk, and models are also deployed and used once the results are derived. Data mining models that use artificial intelligence or statistics commonly support these types of endeavors, deploying supervised and unsupervised models to classify and predict some particular event and to recognize groups of similar behavior within the customer base for subscribed business change.

For example, let's consider what analyses are required when a company decides to launch a new product. Before establishing the proper packaging or price, the company may decide to run a deep study about the marketplace, the competitors, and the consumers. This study should take into consideration current customer needs: Are customers willing to adopt the product? What price might they be willing to pay for it? Do competitors have similar products in the market? If so, how much do they charge? Does the company have pre-existing products that compete with this new one? All these questions might be addressed by using the second layer of analytics. This task is completely on demand, and it would be required to support the product launch.

A more in-depth analysis regarding how customers consume similar products, taking into account historical information about sales, might lead to a predictive model that establishes the likelihood of customers purchasing this new product. This predictive model would support sales campaigns, by defining target customers who have higher likelihoods of purchase, for example. This type of procedure would be associated with the third layer of analytics.

And finally, once that product has been launched, the company could monitor the sales success over the time. Business analysts might have a clear view about how well sales for the product occur in relation to different customer segments, different types of promotions, how profitable the product is in different branches, regions, sales channel, and so forth. This type of task is associated with the first layer of analytics, by delivering readily available insights through a web portal, published reports, interactive visualizations, and other ad hoc queries about that product across different business dimensions.

The entire analytical environment, including applications that support the three stages of analytics, should all relate to the organization's strategy, cover all business issues, and be aligned with the company's priorities. For a new company just starting out, a key objective might be to acquire as many customers as possible. In this scenario a customer acquisition dashboard should be deployed as part of the first stage of analytics, in order to monitor the changing size of the customer base. A market analysis that describes customers' needs should be performed in the second stage, to understand which products and services must be launched. And a predictive model that targets acquisition strategies to the most appropriate prospects should be developed in the third stage of analytics.

On the other hand, if this company is well established and there are several other players emerging in their market, similar applications should be put in place but would be focused on monitoring different activities and events. In this situation the organization would want to monitor, understand, and predict churn (i.e., the rate at which customers leave the company), as well as develop predictive models that target cross and up-sell marketing activities. Nevertheless, the same three layers of analytics are used to adequately cover all the relevant business issues and organizational priorities.

The Role of Analytics in Innovation

Analytics plays a crucial role in modern corporate innovation. The outcomes from analytical models are used to drive new sales processes, to change customer experiences in order to avoid churn, and to identify triggers detecting fraud, risk, or any sort of corporate threat, as well as many other business issues. The knowledge from analytical models is commonly assigned to recognize customer behavior, to predict an event, or to assess the relationship between events, impacting company actions and activities.

The three layers of analytics provide a foundation for data-driven innovation, both creating and delivering new knowledge and accessible information. Each one of these layers might support insights and decisions for different members of the organization, based on their role and responsibility, however, in innovative organizations, access to analytically based answers is fundamental throughout the company. Data is seen as a corporate asset and analytical methods become intellectual property.

Innovation is a wonderful process. It continually evolves, allowing companies to remain ahead of competitors, ahead in the market, and ahead of its time. However, innovation has a price. Perhaps it is an intangible price—and maybe even a higher price than we could imagine. Innovation demands companies stay at the pinnacle of available technology and be on the leading edge of new business actions. But even more than this, innovation requires people to change their minds.

We've all heard that innovation is, at least in part, about thinking outside the box. To ourselves, we often wonder: Which box is that? The proverbial box is really the virtual hedge that each one of us creates and that confines our own ideas. This is a box that we all have to some degree, and many fear it will be criticized, it needs to be argued, and may be interrogated by something new. It is a box composed of our shield against the new and all associated consequences that new can bring forth. By instinct we are afraid of the unknown, the new, and the changing. We can be quite confident in our comfort zone or existing box, but quite unsteady outside of it.

Innovation demands change. It means we must take a chance and address a particular situation and put into place something that may never have been tried before. It means to pursue an idea, check if it is true, and examine if it is valid, feasible, consistent, understandable, comprehensive, and applicable. And in this respect, innovation means to try, and sometimes get it right and make things better, and sometimes not. Therefore, innovation is a trial-and-error process, and as such it is also a heuristic process.

Innovation in a Changing World

Everything changes. The market changes, the consumer changes, the technology changes, and thus products and services must change as well. Analytical models raise the business knowledge regarding what has changed and what needs to change. The new knowledge delivered by analytics is about the company itself, the competitive environment, and the market, but mostly it is about the consumers/constituents that the company serves.

Change is also dynamic. Consumer behaviors change, sometimes for no clear reason and sometimes in response to other events and stimuli. We, as consumers, leave behind a trail of behavior, the way we buy, use, complain, inquire, and so on. By our very behavior we create history for all companies that care about acquiring or retaining our share of wallet to use and to foresee what we might do in the future. However, keep in mind that the unforeseen exists, and someone's team loses the final match of the national championship, or someone's girlfriend decides to split up, someone's son gets ill, or someone is fired. From this day to the next week, month, or perhaps even quarter, consumers can completely change their behavior, putting themselves out of the confidence intervals of the expected curve of behavior, thereby increasing false-positive occurrences.

Analytics is geared toward understanding the average, to accurately forecast for the majority, to target most of the population at hand. What companies, analysts, and data miners need to bear in mind is how heuristic this process can be and, as a result, how they need to monitor and maintain all analytical models to reflect changing conditions.

In the end, because of the dynamic condition, it becomes even more important to accurately frame business questions, and in order to do that well, analytics becomes even more important and essential, helping identify both the known and the unknown. And in many forward-thinking, highly competitive marketplaces, such as banking, telecommunications, and online retail, the use of analytics has become a mandatory corporate strategy.

SUMMARY

This chapter introduced the concepts and foundation of the analytics process, presenting its main purposes: how to apply analytics to solve real-world business problems, how the production and operational environments need to support it, and the heuristic characteristics assigned to the development of most analytical models.

The Monty Hall problem was used to exemplify the importance of theoretical grounding and fundamentals of mathematical science, including statistics and probability, and how such fundamentals are critical to appropriately framing an analytic problem, and hence a solution. The foundation of any analytical model is the rigorousness of the underlying theory assigned to the technique chosen to address the business issue at hand.

The evolution of the analytical processes was also covered in this chapter, describing the different stages involved in business analytics, what the goals are for each stage, and the most common types of insight derived from each one. And while these stages can be chronological it is really a cycle, which improves upon itself with iterations in the process. The distinct stages are typically geared toward a particular purpose and audience, composed of a different set of tools and techniques for each stage in the process.

The relevance of analytics to innovation was also presented, making quite clear how crucial the knowledge stemming from analytical models is to companies who wish to thrive in their marketplaces.

The rest of the book examines different aspects of analytical endeavors, the lifecycle of analytic processes, and the many factors, both internal and external, that influence the definition of the problem, the description of the scenario, the choice of the attributes, the technique, and ultimately the outcome.

Chapter 2 describes how randomness impacts both model development and the results that can be expected. Regardless of how precisely we try to define a particular problem, or how well we describe the problem situation and the variables used to describe it—it all becomes an approximation. The randomness associated with most business problems that organizations are trying to solve influences the results and can make model training more difficult, increase the model error, and affect the overall accuracy and value. In particular, when we are talking about human behavior, we are depicting all scenarios as an approximation of their reality. We are prescribing the general or average behavior to an individual and due to this approach, the expected results might change in the end.

Chapter 3 details the heuristic approach, how we use it (consciously or not) during the model definition, the development phase, and particularly in the adjustment steps, done to ensure the outcomes achieved satisfy the objectives established. This chapter describes how heuristics are present in most of the analytical tasks we perform—in the definition/selection of attributes used to explain a particular problem, in the approximation we chose to describe the boundaries of the problem scenario, and in the parameters used to develop the analytical model (including the techniques used). Many aspects of this analytical cycle are actually a set of choices we have to make as analysts, some based on previous experience, others based on trial and error. These choices define the rationale as to why we chose an artificial neural network instead of a decision tree or regression and answer questions regarding why, by selecting an artificial neural networks, we chose the particular topology, error propagation technique, and particular type of activation function.

Chapter 4 presents the analytical approach to develop and deploy analytical models, considering some of the more common techniques applied to solve real-world business problems. Most of the statistical approaches in and of themselves are not that difficult to implement. The challenge of making them successful lies in the reasonableness of the interpretation. This translation of model outcomes to business rules is one of the most important aspects governing the success of statistical models in both operational and production environments. We focus in this chapter on the analytical approach based on data mining, mostly artificial intelligence and mathematical models. Techniques such as artificial neural networks, decision trees, and genetic algorithms, as well as optimization models such as social network analysis, are described. The competitive cycle is used to describe practical application steps for analytical modeling. We show how and where each type of model might be used in order to improve the business performance and competitiveness. This cycle covers steps from data analysis to persistence, including data mining and outcomes application, and the chapter is generally focused on data mining.

Chapter 5 complements Chapter 4 with a description of practical case studies. Real-world examples that illustrate data mining methods in knowledge discovery are described. In practice, such analysis is used to understand customer behavior—highlighting the most likely customer to purchase (acquisition), to consume (cross and up selling), and to churn (retention). The analytical models described include clustering techniques, predictive models for collection—including insolvency segmentation and bad debt prediction—and fraud detection.

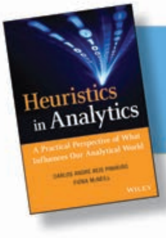
Chapter 6 describes a very special type of optimization models, known as graph analysis (aka network analysis). We discuss the fundamental concepts, types of the graph structures, the network metrics, and some analyses approaches.

Chapter 7 illustrates the use of graph analysis by way of three real-world business cases. Social networks are a particular case of graph analysis being used by some organizations driven by the desire to understand big data relationships. Two of the case studies are associated with the telecommunications industry: one associated with churn and sales diffusion, the other describing a fraud scenario. Exaggeration of claims in insurance, and how to identify potential suspicious actors, is also discussed.

Chapter 8 explores a newer technique in the analytical world, that of transforming unstructured text data into one that is structured. Unstructured text data is, by definition, descriptive and includes commentaries contained in call centers memos, log notes from servicing events, and reports from claims or interviews. Such text data can be verbose, such as in research papers, patent applications, and even books, or terse, like that from social media tweets. Regardless of the source, analysis of text data can provide deep insight as to why customers behave as they do, what they say they will do, and how they perceive what they are doing. Given this method is based on the meaning of words, the analysis of text requires interpretation, thereby making it heuristic by its very definition.

This book illustrates the overall customer lifecycle and how analytics can be successfully deployed in order to improve each and every step of this cycle. Regardless of the lifecycle phase, analytics can be used to solve a wide range of business issues. To the uninitiated, analytics might look like very complex, mystical science, heavily based in mathematics, statistics, or linguistics, composed of formulas and countless algorithms. And perhaps, behind the scenes, it is indeed like that. But when analytical tools hold the promise of making our life easier and our organizations more successful, it is better to dive in and begin to derive value from analytic methods. Most of the good tools, such as SAS, contain a substantial range of algorithms and techniques that are user ready—and oftentimes require minimal configuration effort. This book aims to describe how analytics can solve business problems in practice, and perhaps even more important, how these outcomes might be translated into successful business actions. A marketing perspective is used to depict a variety of possible analytics usage in companies, considering distinct industries, including telecommunications, banking, insurance, retail, entertainment, and others. Common to all industries, this overview considers practical examples, covering analytical models to acquire customers, to improve sales, avoid turnover (aka churn), as well as to detect fraud, bad debt, payment risk, and collection. By using a suite of analytical models, organizations are better prepared to appropriately serve customers and more accurately comprehend the marketplaces they are involved with, the competitors they face, and how they can boost their business strength.

Throughout the book, many examples are presented to illustrate analytical processes for all three stages described in this chapter. All of these examples were developed with SAS software, including products such as Base SAS®, SAS® Enterprise Miner™, SAS/STAT®, SAS/GRAPH®, SAS/OR®, and SAS® Customer Link Analytics. Most of these examples of analytical implementations were really developed in practice, and mostly in larger organizations.


Heuristics in Analytics
A Practical Perspective of What Influences Our Analytical World
CARLOS ANDRE REIS PINHEIRO
FIONA McNEILL
WILEY