2

Analytics: A Comprehensive Study

Analytics is the understanding and communication of significant patterns of data. Analytics is applied in businesses to improve their performances. Some of the aspects explained in this text are software analytics, embedded analytics, learning analytics and social media analytics. The section on analytics offers an insightful focus, keeping in mind the complex subject matter.

Business Analytics

Business analytics (BA) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. In contrast, business intelligence traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning, which is also based on data and statistical methods.

Business analytics makes extensive use of statistical analysis, including explanatory and predictive modeling, and fact-based management to drive decision making. It is therefore closely related to management science. Analytics may be used as input for human decisions or may drive fully automated decisions. Business intelligence is querying, reporting, online analytical processing (OLAP), and "alerts."

In other words, querying, reporting, OLAP, and alert tools can answer questions such as what happened, how many, how often, where the problem is, and what actions are needed. Business analytics can answer questions like why is this happening, what if these trends continue, what will happen next (that is, predict), what is the best that can happen (that is, optimize).

Examples of Application

Banks, such as Capital One, use data analysis (or *analytics*, as it is also called in the business setting), to differentiate among customers based on credit risk, usage and other characteristics and then to match customer characteristics with appropriate product offerings. Harrah's, the gaming firm, uses analytics in its customer loyalty programs. E & J Gallo Winery quantitatively analyzes and predicts the appeal of its wines. Between 2002 and 2005, Deere & Company saved more than $1 billion by employing a new analytical tool to better optimize inventory. A telecoms company that pursues efficient call centre usage over customer service may save money.

Types of Analytics

- Decision analytics: supports human decisions with visual analytics the user models to reflect reasoning.

- Descriptive analytics: gains insight from historical data with reporting, scorecards, clustering etc.
- Predictive analytics: employs predictive modeling using statistical and machine learning techniques
- Prescriptive analytics: recommends decisions using optimization, simulation, etc.

Basic Domains within Analytics

- Behavioral analytics
- Cohort Analysis
- Collections analytics
- Contextual data modeling - supports the human reasoning that occurs after viewing "executive dashboards" or any other visual analytics
- Cyber analytics
- Enterprise Optimization
- Financial services analytics
- Fraud analytics
- Marketing analytics
- Pricing analytics
- Retail sales analytics
- Risk & Credit analytics
- Supply Chain analytics
- Talent analytics
- Telecommunications
- Transportation analytics

History

Analytics have been used in business since the management exercises were put into place by Frederick Winslow Taylor in the late 19th century. Henry Ford measured the time of each component in his newly established assembly line. But analytics began to command more attention in the late 1960s when computers were used in decision support systems. Since then, analytics have changed and formed with the development of enterprise resource planning (ERP) systems, data warehouses, and a large number of other software tools and processes.

In later years the business analytics have exploded with the introduction to computers. This change has brought analytics to a whole new level and has made the possibilities endless. As far as analyt-

ics has come in history, and what the current field of analytics is today many people would never think that analytics started in the early 1900s with Mr. Ford himself.

Challenges

Business analytics depends on sufficient volumes of high quality data. The difficulty in ensuring data quality is integrating and reconciling data across different systems, and then deciding what subsets of data to make available.

Previously, analytics was considered a type of after-the-fact method of forecasting consumer behavior by examining the number of units sold in the last quarter or the last year. This type of data warehousing required a lot more storage space than it did speed. Now business analytics is becoming a tool that can influence the outcome of customer interactions. When a specific customer type is considering a purchase, an analytics-enabled enterprise can modify the sales pitch to appeal to that consumer. This means the storage space for all that data must react extremely fast to provide the necessary data in real-time.

Competing on Analytics

Thomas Davenport, professor of information technology and management at Babson College argues that businesses can optimize a distinct business capability via analytics and thus better compete. He identifies these characteristics of an organization that are apt to compete on analytics:

- One or more senior executives who strongly advocate fact-based decision making and, specifically, analytics
- Widespread use of not only descriptive statistics, but also predictive modeling and complex optimization techniques
- Substantial use of analytics across multiple business functions or processes
- Movement toward an enterprise level approach to managing analytical tools, data, and organizational skills and capabilities

Analytics

Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight.

Organizations may apply analytics to business data to describe, predict, and improve business performance. Specifically, areas within analytics include predictive analytics, prescriptive analytics, enterprise decision management, retail analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modeling, web analytics, sales force sizing and optimization, price and promotion modeling, predictive science, credit risk analysis, and fraud analytics. Since analytics can require extensive computation, the algorithms

and software used for analytics harness the most current methods in computer science, statistics, and mathematics.

Analytics vs. Analysis

Analytics is multidisciplinary. There is extensive use of mathematics and statistics, the use of descriptive techniques and predictive models to gain valuable knowledge from data—data analysis. The insights from data are used to recommend action or to guide decision making rooted in business context. Thus, analytics is not so much concerned with individual analyses or analysis steps, but with the entire methodology. There is a pronounced tendency to use the term *analytics* in business settings e.g. text analytics vs. the more generic text mining to emphasize this broader perspective.. There is an increasing use of the term *advanced analytics*, typically used to describe the technical aspects of analytics, especially in the emerging fields such as the use of machine learning techniques like neural networks to do predictive modeling.

Examples

Marketing Optimization

Marketing has evolved from a creative process into a highly data-driven process. Marketing organizations use analytics to determine the outcomes of campaigns or efforts and to guide decisions for investment and consumer targeting. Demographic studies, customer segmentation, conjoint analysis and other techniques allow marketers to use large amounts of consumer purchase, survey and panel data to understand and communicate marketing strategy.

Web analytics allows marketers to collect session-level information about interactions on a website using an operation called sessionization. Google Analytics is an example of a popular free analytics tool that marketers use for this purpose. Those interactions provide web analytics information systems with the information necessary to track the referrer, search keywords, identify IP address, and track activities of the visitor. With this information, a marketer can improve marketing campaigns, website creative content, and information architecture.

Analysis techniques frequently used in marketing include marketing mix modeling, pricing and promotion analyses, sales force optimization and customer analytics e.g.: segmentation. Web analytics and optimization of web sites and online campaigns now frequently work hand in hand with the more traditional marketing analysis techniques. A focus on digital media has slightly changed the vocabulary so that *marketing mix modeling* is commonly referred to as *attribution modeling* in the digital or marketing mix modeling context.

These tools and techniques support both strategic marketing decisions (such as how much overall to spend on marketing, how to allocate budgets across a portfolio of brands and the marketing mix) and more tactical campaign support, in terms of targeting the best potential customer with the optimal message in the most cost effective medium at the ideal time.

Portfolio Analytics

A common application of business analytics is portfolio analysis. In this, a bank or lending agency has a collection of accounts of varying value and risk. The accounts may differ by the social status

(wealthy, middle-class, poor, etc.) of the holder, the geographical location, its net value, and many other factors. The lender must balance the return on the loan with the risk of default for each loan. The question is then how to evaluate the portfolio as a whole.

The least risk loan may be to the very wealthy, but there are a very limited number of wealthy people. On the other hand, there are many poor that can be lent to, but at greater risk. Some balance must be struck that maximizes return and minimizes risk. The analytics solution may combine time series analysis with many other issues in order to make decisions on when to lend money to these different borrower segments, or decisions on the interest rate charged to members of a portfolio segment to cover any losses among members in that segment.

Risk Analytics

Predictive models in the banking industry are developed to bring certainty across the risk scores for individual customers. Credit scores are built to predict individual's delinquency behavior and widely used to evaluate the credit worthiness of each applicant. Furthermore, risk analyses are carried out in the scientific world and the insurance industry. It is also extensively used in financial institutions like Online Payment Gateway companies to analyse if a transaction was genuine or fraud. For this purpose they use the transaction history of the customer. This is more commonly used in Credit Card purchase, when there is a sudden spike in the customer transaction volume the customer gets a call of confirmation if the transaction was initiated by him/her. This helps in reducing loss due to such circumstances.

Digital Analytics

Digital analytics is a set of business and technical activities that define, create, collect, verify or transform digital data into reporting, research, analyses, recommendations, optimizations, predictions, and automations. This also includes the SEO (Search Engine Optimization) where the keyword search is tracked and that data is used for marketing purposes. Even banner ads and clicks come under digital analytics. All marketing firms rely on digital analytics for their digital marketing assignments, where MROI (Marketing Return on Investment) is important.

Security Analytics

Security analytics refers to information technology (IT) solutions that gather and analyze security events to bring situational awareness and enable IT staff to understand and analyze events that pose the greatest risk. Solutions in this area include security information and event management solutions and user behavior analytics solutions.

Software Analytics

Software analytics is the process of collecting information about the way a piece of software is used and produced.

Challenges

In the industry of commercial analytics software, an emphasis has emerged on solving the chal-

lenges of analyzing massive, complex data sets, often when such data is in a constant state of change. Such data sets are commonly referred to as big data. Whereas once the problems posed by big data were only found in the scientific community, today big data is a problem for many businesses that operate transactional systems online and, as a result, amass large volumes of data quickly.

The analysis of unstructured data types is another challenge getting attention in the industry. Unstructured data differs from structured data in that its format varies widely and cannot be stored in traditional relational databases without significant effort at data transformation. Sources of unstructured data, such as email, the contents of word processor documents, PDFs, geospatial data, etc., are rapidly becoming a relevant source of business intelligence for businesses, governments and universities. For example, in Britain the discovery that one company was illegally selling fraudulent doctor's notes in order to assist people in defrauding employers and insurance companies, is an opportunity for insurance firms to increase the vigilance of their unstructured data analysis. The McKinsey Global Institute estimates that big data analysis could save the American health care system $300 billion per year and the European public sector €250 billion.

These challenges are the current inspiration for much of the innovation in modern analytics information systems, giving birth to relatively new machine analysis concepts such as complex event processing, full text search and analysis, and even new ideas in presentation. One such innovation is the introduction of grid-like architecture in machine analysis, allowing increases in the speed of massively parallel processing by distributing the workload to many computers all with equal access to the complete data set.

Analytics is increasingly used in education, particularly at the district and government office levels. However, the complexity of student performance measures presents challenges when educators try to understand and use analytics to discern patterns in student performance, predict graduation likelihood, improve chances of student success, etc. For example, in a study involving districts known for strong data use, 48% of teachers had difficulty posing questions prompted by data, 36% did not comprehend given data, and 52% incorrectly interpreted data. To combat this, some analytics tools for educators adhere to an over-the-counter data format (embedding labels, supplemental documentation, and a help system, and making key package/display and content decisions) to improve educators' understanding and use of the analytics being displayed.

One more emerging challenge is dynamic regulatory needs. For example, in the banking industry, Basel III and future capital adequacy needs are likely to make even smaller banks adopt internal risk models. In such cases, cloud computing and open source R (programming language) can help smaller banks to adopt risk analytics and support branch level monitoring by applying predictive analytics.

Risks

The main risk for the people is discrimination like price discrimination or statistical discrimination. Analytical processes can also result in discriminatory outcomes that may violate anti-discrimination and civil rights laws. There is also the risk that a developer could profit from the ideas or work done by users, like this example: Users could write new ideas in a note taking app, which could then be sent as a custom event, and the developers could profit from those ideas. This can happen because the ownership of content is usually unclear in the law.

If a user's identity is not protected, there are more risks; for example, the risk that private information about users is made public on the internet.

In the extreme, there is the risk that governments could gather too much private information, now that the governments are giving themselves more powers to access citizens' information.

Software Analytics

Software Analytics refers to analytics specific to software systems and related software development processes. It aims at describing, predicting, and improving development, maintenance, and management of complex software systems. Methods and techniques of software analytics typically rely on gathering, analyzing, and visualizing information found in the manifold data sources in the scope of software systems and their software development processes---software analytics "turns it into actionable insight to inform better decisions related to software".

Software analytics represents a base component of software diagnosis that generally aims at generating findings, conclusions, and evaluations about software systems and their implementation, composition, behavior, and evolution. Software analytics frequently uses and combines approaches and techniques from statistics, prediction analysis, data mining, and scientific visualization. For example, software analytics can map data by means of software maps that allow for interactive exploration.

Data under exploration and analysis by Software Analytics exists in software lifecycle, including source code, software requirement specifications, bug reports, test cases, execution traces/logs, and real-world user feedback, etc. Data plays a critical role in modern software development, because hidden in the data is the information and insight about the quality of software and services, the experience that software users receive, as well as the dynamics of software development.

Insightful information obtained by Software Analytics is information that conveys meaningful and useful understanding or knowledge towards performing the target task. Typically insightful information cannot be easily obtained by direct investigation on the raw data without the aid of analytic technologies.

Actionable information obtained by Software Analytics is information upon which software practitioners can come up with concrete solutions (better than existing solutions if any) towards completing the target task.

Software Analytics focuses on trinity of software systems, software users, and software development process:

Software Systems. Depending on scale and complexity, the spectrum of software systems can span from operating systems for devices to large networked systems that consist of thousands of servers. System quality such as reliability, performance and security, etc., is the key to success of modern software systems. As the system scale and complexity greatly increase, larger amount of data, e.g., run-time traces and logs, is generated; and data becomes a critical means to monitor, analyze, understand and improve system quality.

Software Users. Users are (almost) always right because ultimately they will use the software and services in various ways. Therefore, it is important to continuously provide the best experience to users. Usage data collected from the real world reveals how users interact with software and services. The data is incredibly valuable for software practitioners to better understand their customers and gain insights on how to improve user experience accordingly.

Software Development Process. Software development has evolved from its traditional form to exhibiting different characteristics. The process is more agile and engineers are more collaborative than that in the past. Analytics on software development data provides a powerful mechanism that software practitioners can leverage to achieve higher development productivity.

In general, the primary technologies employed by Software Analytics include analytical technologies such as machine learning, data mining and pattern recognition, information visualization, as well as large-scale data computing & processing.

History

In May 2009, Software Analytics was first coined and proposed when Dr. Dongmei Zhang founded the Software Analytics Group (SA) at Microsoft Research Asia (MSRA). The term has become well known in the software engineering research community after a series of tutorials and talks on software analytics were given by Dr. Dongmei Zhang and her colleagues, in collaboration with Professor Tao Xie from North Carolina State University, at software engineering conferences including a tutorial at the IEEE/ACM International Conference on Automated Software Engineering (ASE 2011), a talk at the International Workshop on Machine Learning Technologies in Software Engineering (MALETS 2011), a tutorial and a keynote talk given by Dr. Dongmei Zhang at the IEEE-CS Conference on Software Engineering Education and Training (CSEE&T 2012), a tutorial at the International Conference on Software Engineering (ICSE 2012) - Software Engineering in Practice Track, and a keynote talk given by Dr. Dongmei Zhang at the Working Conference on Mining Software Repositories (MSR 2012).

In November 2010, Software Development Analytics (Software Analytics with focus on Software Development) was proposed by Thomas Zimmermann and his colleagues at the Empirical Software Engineering Group (ESE) at Microsoft Research Redmond in their FoSER 2010 paper. A goldfish bowl panel on software development analytics was organized by Thomas Zimmermann and Professor Tim Menzies from West Virginia University at the International Conference on Software Engineering (ICSE 2012), Software Engineering in Practice track.

Software Analytics Providers

- CAST Software
- IBM Cognos Business Intelligence
- Kiuwan
- Microsoft Azure Application Insights
- Nalpeiron Software Analytics
- New Relic

- Squore
- Tableau Software
- Trackerbird Software Analytics

Embedded Analytics

Embedded analytics is the technology designed to make data analysis and business intelligence more accessible by all kind of application or user.

Definition

According to Gartner analysts Kurt Schlegel, traditional business intelligence were suffering in 2008 a lack of integration between the data and the business users. This technology intention is to be more pervasive by real-time autonomy and self-service of data visualization or customization, meanwhile decision makers, business users or even customers are doing their own daily workflow and tasks.

History

First mentions of the concept were made by Howard Dresner, consultant, author, former Gartner analyst and inventor of the term "business intelligence". Consolidation of business intelligence "doesn't mean the BI market has reached maturity" said Howard Dresner while he was working for Hyperion Solutions, a company that Oracle bought in 2007. Oracle started then to use the term "embedded analytics" at their press release for Oracle® Rapid Planning on 2009. Gartner Group, a company for which Howard Dresner has been working, finally added the term to their IT Glossary on November 5, 2012. . It was clear this was a mainstream technology when Dresner Advisory Services published the 2014 Embedded Business Intelligence Market Study as part of the Wisdom of Crowds® Series of Research, including 24 vendors.

Tools

- Actuate
- Dundas Data Visualization
- GoodData
- IBM
- icCube
- Logi Analytics
- Pentaho
- Qlik
- SAP

- SAS
- Tableau
- TIBCO
- Sisense

Learning Analytics

Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. A related field is educational data mining. For general audience introductions, see:

- The Educause Learning Initiative Briefing
- The Educause Review on Learning analytics
- And the UNESCO "Learning Analytics Policy Brief" (2012)

What is Learning Analytics?

The definition and aims of Learning Analytics are contested. One earlier definition discussed by the community suggested that "Learning analytics is the use of intelligent data, learner-produced data, and analysis models to discover information and social connections for predicting and advising people's learning."

But this definition has been criticised:

1. *"I somewhat disagree with this definition - it serves well as an introductory concept if we use analytics as a support structure for existing education models. I think learning analytics - at an advanced and integrated implementation - can do away with pre-fab curriculum models"*. George Siemens, 2010.
2. *"In the descriptions of learning analytics we talk about using data to "predict success". I've struggled with that as I pore over our databases. I've come to realize there are different views/levels of success."* Mike Sharkey 2010.

A more holistic view than a mere definition is provided by the framework of learning analytics by Greller and Drachsler (2012). It uses a general morphological analysis (GMA) to divide the domain into six "critical dimensions".

A systematic overview on learning analytics and its key concepts is provided by Chatti et al. (2012) and Chatti et al. (2014) through a reference model for learning analytics based on four dimensions, namely data, environments, context (what?), stakeholders (who?), objectives (why?), and methods (how?).

It has been pointed out that there is a broad awareness of analytics across educational institutions for various stakeholders, but that the way 'learning analytics' is defined and implemented may vary, including:

1. for individual learners to reflect on their achievements and patterns of behaviour in relation to others;
2. as predictors of students requiring extra support and attention;
3. to help teachers and support staff plan supporting interventions with individuals and groups;
4. for functional groups such as course team seeking to improve current courses or develop new curriculum offerings; and
5. for institutional administrators taking decisions on matters such as marketing and recruitment or efficiency and effectiveness measures."

In that briefing paper, Powell and MacNeill go on to point out that some motivations and implementations of analytics may come into conflict with others, for example highlighting potential conflict between analytics for individual learners and organisational stakeholders.

Gašević, Dawson, and Siemens argue that computational aspects of learning analytics need to be linked with the existing educational research if the field of learning analytics is to deliver to its promise to understand and optimize learning.

Differentiating Learning Analytics and Educational Data Mining

Differentiating the fields of educational data mining (EDM) and learning analytics (LA) has been a concern of several researchers. George Siemens takes the position that educational data mining encompasses both learning analytics and academic analytics, the former of which is aimed at governments, funding agencies, and administrators instead of learners and faculty. Baepler and Murdoch define academic analytics as an area that *"...combines select institutional data, statistical analysis, and predictive modeling to create intelligence upon which learners, instructors, or administrators can change academic behavior"*. They go on to attempt to disambiguate educational data mining from academic analytics based on whether the process is hypothesis driven or not, though Brooks questions whether this distinction exists in the literature. Brooks instead proposes that a better distinction between the EDM and LA communities is in the roots of where each community originated, with authorship at the EDM community being dominated by researchers coming from intelligent tutoring paradigms, and learning anaytics researchers being more focused on enterprise learning systems (e.g. learning content management systems).

Regardless of the differences between the LA and EDM communities, the two areas have significant overlap both in the objectives of investigators as well as in the methods and techniques that are used in the investigation. In the MS program offering in Learning Analytics at Teachers College, Columbia University, students are taught both EDM and LA methods.

History

The Context of Learning Analytics

In "The State of Learning Analytics in 2012: A Review and Future Challenges" Rebecca Ferguson tracks the progress of analytics for learning as a development through:

1. The increasing interest in 'big data' for business intelligence

2. The rise of online education focussed around Virtual Learning Environments (VLEs), Content Management Systems (CMSs), and Management Information Systems (MIS) for education, which saw an increase in digital data regarding student background (often held in the MIS) and learning log data (from VLEs). This development afforded the opportunity to apply 'business intelligence' techniques to educational data
3. Questions regarding the optimisation of systems to support learning particularly given the question regarding how we can know whether a student is engaged/understanding if we can't see them?
4. Increasing focus on evidencing progress and professional standards for accountability systems
5. This focus led to a teacher stakehold in the analytics - given that they are associated with accountability systems
6. Thus an increasing emphasis was placed on the pedagogic affordances of learning analytics
7. This pressure is increased by the economic desire to improve engagement in online education for the deliverance of high quality - affordable - education

History of The Techniques and Methods of Learning Analytics

In a discussion of the history of analytics, Cooper highlights a number of communities from which learning analytics draws techniques, including:

1. Statistics - which are a well established means to address hypothesis testing
2. Business Intelligence - which has similarities with learning analytics, although it has historically been targeted at making the production of reports more efficient through enabling data access and summarising performance indicators.
3. Web analytics - tools such as Google analytics report on web page visits and references to websites, brands and other keyterms across the internet. The more 'fine grain' of these techniques can be adopted in learning analytics for the exploration of student trajectories through learning resources (courses, materials, etc.).
4. Operational research - aims at highlighting design optimisation for maximising objectives through the use of mathematical models and statistical methods. Such techniques are implicated in learning analytics which seek to create models of real world behaviour for practical application.
5. Artificial intelligence and Data mining - Machine learning techniques built on data mining and AI methods are capable of detecting patterns in data. In learning analytics such techniques can be used for intelligent tutoring systems, classification of students in more dynamic ways than simple demographic factors, and resources such as 'suggested course' systems modelled on collaborative filtering techniques.
6. Social Network Analysis - SNA analyses relationships between people by exploring implicit (e.g. interactions on forums) and explicit (e.g. 'friends' or 'followers') ties online and offline. SNA developed from the work of sociologists like Wellman and Watts, and mathe-

maticians like Barabasi and Strogatz. The work of these individuals has provided us with a good sense of the patterns that networks exhibit (small world, power laws), the attributes of connections (in early 70's, Granovetter explored connections from a perspective of tie strength and impact on new information), and the social dimensions of networks (for example, geography still matters in a digital networked world). It is particularly used to explore clusters of networks, influence networks, engagement and disengagement, and has been deployed for these purposes in learning analytic contexts.

7. Information visualization - visualisation is an important step in many analytics for sensemaking around the data provided - it is thus used across most techniques (including those above).

History of Learning Analytics in Higher Education

The first graduate program focused specifically on learning analytics was created by Dr. Ryan Baker and launched in the Fall 2015 semester at Teachers College - Columbia University. The program description states that "data about learning and learners are being generated today on an unprecedented scale. The fields of learning analytics (LA) and educational data mining (EDM) have emerged with the aim of transforming this data into new insights that can benefit students, teachers, and administrators. As one of world's leading teaching and research institutions in education, psychology, and health, we are proud to offer an innovative graduate curriculum dedicated to improving education through technology and data analysis."

Analytic Methods

Methods for learning analytics include:

- Content analysis - particularly of resources which students create (such as essays)
- Discourse Analytics Discourse analytics aims to capture meaningful data on student interactions which (unlike 'social network analytics') aims to explore the properties of the language used, as opposed to just the network of interactions, or forum-post counts, etc.
- Social Learning Analytics which is aimed at exploring the role of social interaction in learning, the importance of learning networks, discourse used to sensemake, etc.
- Disposition Analytics which seeks to capture data regarding student's dispositions to their own learning, and the relationship of these to their learning. For example, "curious" learners may be more inclined to ask questions - and this data can be captured and analysed for learning analytics.

Analytic Outcomes

Analytics have been used for:

- Prediction purposes, for example to identify 'at risk' students in terms of drop out or course failure
- Personalization & Adaptation, to provide students with tailored learning pathways, or assessment materials

- Intervention purposes, providing educators with information to intervene to support students
- Information visualization, typically in the form of so-called learning dashboards which provide overview learning data through data visualisation tools

Software

Much of the software that is currently used for learning analytics duplicates functionality of web analytics software, but applies it to learner interactions with content. Social network analysis tools are commonly used to map social connections and discussions. Some examples of learning analytics software tools:

- Student Success System - a predictive learning analytics tool that predicts student performance and plots learners into risk quadrants based upon engagement and performance predictions, and provides indicators to develop understanding as to why a learner is not on track through visualizations such as the network of interactions resulting from social engagement (e.g. discussion posts and replies), performance on assessments, engagement with content, and other indicators
- SNAPP - a learning analytics tool that visualizes the network of interactions resulting from discussion forum posts and replies.
- LOCO-Analyst - a context-aware learning tool for analytics of learning processes taking place in a web-based learning environment
- SAM - a Student Activity Monitor intended for Personal Learning Environments
- BEESTAR INSIGHT - a real-time system that automatically collects student engagement and attendance & provides analytics tools and dashboards for students, teachers & management
- Solutionpath StREAM- A leading UK based real-time system that leverage predictive models to determine all facets of student engagement using structured and unstructured sources for all institutional roles

Ethics & Privacy

The ethics of data collection, analytics, reporting and accountability has been raised as a potential concern for Learning Analytics (e.g.,), with concerns raised regarding:

- Data ownership
- Communications around the scope and role of Learning Analytics
- The necessary role of human feedback and error-correction in Learning Analytics systems
- Data sharing between systems, organisations, and stakeholders
- Trust in data clients

As Kay, Kom and Oppenheim point out, the range of data is wide, potentially derived from: "*Recorded activity; student records, attendance, assignments, researcher information (CRIS).

- Systems interactions; VLE, library / repository search, card transactions.
- Feedback mechanisms; surveys, customer care.
- External systems that offer reliable identification such as sector and shared services and social networks."

Thus the legal and ethical situation is challenging and different from country to country, raising implications for: "*Variety of data - principles for collection, retention and exploitation.

- Education mission - underlying issues of learning management, including social and performance engineering.
- Motivation for development of analytics – mutuality, a combination of corporate, individual and general good.
- Customer expectation – effective business practice, social data expectations, cultural considerations of a global customer base. *Obligation to act – duty of care arising from knowledge and the consequent challenges of student and employee performance management."

In some prominent cases like the inBloom disaster even full functional systems have been shut down due to lack of trust in the data collection by governments, stakeholders and civil rights groups. Since then, the Learning Analytics community has extensively studied legal conditions in a series of experts workshops on 'Ethics & Privacy 4 Learning Analytics' that constitute the use of trusted Learning Analytics. Drachsler & Greller released a 8-point checklist named DELICATE that is based on the intensive studies in this area to demystify the ethics and privacy discussions around Learning Analytics.

1. D-etermination: Decide on the purpose of learning analytics for your institution.
2. E-xplain: Define the scope of data collection and usage.
3. L-egitimate: Explain how you operate within the legal frameworks, refer to the essential legislation.
4. I-nvolve: Talk to stakeholders and give assurances about the data distribution and use.
5. C-onsent: Seek consent through clear consent questions.
6. A-nonymise: De-identify individuals as much as possible
7. T-echnical aspects: Monitor who has access to data, especially in areas with high staff turnover.
8. E-xternal partners: Make sure externals provide highest data security standards

It shows ways to design and provide privacy conform Learning Analytics that can benefit all stakeholders. The full DELICATE checklist is publicly available here.

Open Learning Analytics

Chatti, Muslim and Schroeder note that the aim of Open Learning Analytics (OLA) is to improve learning effectiveness in lifelong learning environments. The authors refer to OLA as an ongoing analytics process that encompasses diversity at all four dimensions of the learning analytics reference model.

Predictive Analytics

Predictive analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

The defining functional effect of these technical approaches is that predictive analytics provides a predictive score (probability) for each individual (customer, employee, healthcare patient, product SKU, vehicle, component, machine, or other organizational unit) in order to determine, inform, or influence organizational processes that pertain across large numbers of individuals, such as in marketing, credit risk assessment, fraud detection, manufacturing, healthcare, and government operations including law enforcement.

Predictive analytics is used in actuarial science, marketing, financial services, insurance, telecommunications, retail, travel, healthcare, child protection, pharmaceuticals, capacity planning and other fields.

One of the most well known applications is credit scoring, which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time.

Definition

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Predictive analytics is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element. This distinguish-

es it from forecasting. For example, "Predictive analytics—Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions." In future industrial systems, the value of predictive analytics will be to predict and prevent potential issues to achieve near-zero break-down and further be integrated into prescriptive analytics for decision optimization. Furthermore, the converted data can be used for closed-loop product life cycle improvement which is the vision of Industrial Internet Consortium.

Types

Generally, the term predictive analytics is used to mean predictive modeling, "scoring" data with predictive models, and forecasting. However, people are increasingly using the term to refer to related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making, but have different purposes and the statistical techniques underlying them vary.

Predictive Models

Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models in many areas, such as marketing, where they seek out subtle data patterns to answer questions about customer performance, or fraud detection models. Predictive models often perform calculations during live transactions, for example, to evaluate the risk or opportunity of a given customer or transaction, in order to guide a decision. With advancements in computing speed, individual agent modeling systems have become capable of simulating human behaviour or reactions to given stimuli or scenarios.

The available sample units with known attributes and known performances is referred to as the "training sample". The units in other samples, with known attributes but unknown performances, are referred to as "out of [training] sample" units. The out of sample bear no chronological relation to the training sample units. For example, the training sample may consists of literary attributes of writings by Victorian authors, with known attribution, and the out-of sample unit may be newly found writing with unknown authorship; a predictive model may aid in attributing a work to a known author. Another example is given by analysis of blood splatter in simulated crime scenes in which the out of sample unit is the actual blood splatter pattern from a crime scene. The out of sample unit may be from the same time as the training units, from a previous time, or from a future time.

Descriptive Models

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do. Instead, descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling

tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

Decision Models

Decision models describe the relationship between all the elements of a decision—the known data (including results of predictive models), the decision, and the forecast results of the decision—in order to predict the results of decisions involving many variables. These models can be used in optimization, maximizing certain outcomes while minimizing others. Decision models are generally used to develop decision logic or a set of business rules that will produce the desired action for every customer or circumstance.

Applications

Although predictive analytics can be put to use in many applications, we outline a few examples where predictive analytics has shown positive impact in recent years.

Analytical Customer Relationship Management (CRM)

Analytical customer relationship management (CRM) is a frequent commercial application of predictive analysis. Methods of predictive analysis are applied to customer data to pursue CRM objectives, which involve constructing a holistic view of the customer no matter where their information resides in the company or the department involved. CRM uses predictive analysis in applications for marketing campaigns, sales, and customer services to name a few. These tools are required in order for a company to posture and focus their efforts effectively across the breadth of their customer base. They must analyze and understand the products in demand or have the potential for high demand, predict customers' buying habits in order to promote relevant products at multiple touch points, and proactively identify and mitigate issues that have the potential to lose customers or reduce their ability to gain new ones. Analytical customer relationship management can be applied throughout the customers lifecycle (acquisition, relationship growth, retention, and win-back). Several of the application areas described below (direct marketing, cross-sell, customer retention) are part of customer relationship management.

Child Protection

Over the last 5 years, some child welfare agencies have started using predictive analytics to flag high risk cases. The approach has been called "innovative" by the Commission to Eliminate Child Abuse and Neglect Fatalities (CECANF), and in Hillsborough County, Florida, where the lead child welfare agency uses a predictive modeling tool, there have been no abuse-related child deaths in the target population as of this writing.

Clinical Decision Support Systems

Experts use predictive analysis in health care primarily to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease, and other lifetime illnesses. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care. A working definition has been proposed by

Jerome A. Osheroff and colleagues: *Clinical decision support (**CDS**) provides clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care. It encompasses a variety of tools and interventions such as computerized alerts and reminders, clinical guidelines, order sets, patient data reports and dashboards, documentation templates, diagnostic support, and clinical workflow tools.*

A 2016 study of neurodegenerative disorders provides a powerful example of a CDS platform to diagnose, track, predict and monitor the progression of Parkinson's disease. Using large and multi-source imaging, genetics, clinical and demographic data, these investigators developed a decision support system that can predict the state of the disease with high accuracy, consistency and precision. They employed classical model-based and machine learning model-free methods to discriminate between different patient and control groups. Similar approaches may be used for predictive diagnosis and disease progression forecasting in many neurodegenerative disorders like Alzheimer's, Huntington's, Amyotrophic Lateral Sclerosis, as well as for other clinical and biomedical applications where Big Data is available.

Collection Analytics

Many portfolios have a set of delinquent customers who do not make their payments on time. The financial institution has to undertake collection activities on these customers to recover the amounts due. A lot of collection resources are wasted on customers who are difficult or impossible to recover. Predictive analytics can help optimize the allocation of collection resources by identifying the most effective collection agencies, contact strategies, legal actions and other strategies to each customer, thus significantly increasing recovery at the same time reducing collection costs.

Cross-sell

Often corporate organizations collect and maintain abundant data (e.g. customer records, sale transactions) as exploiting hidden relationships in the data can provide a competitive advantage. For an organization that offers multiple products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to efficient cross sales, or selling additional products to current customers. This directly leads to higher profitability per customer and stronger customer relationships.

Customer Retention

With the number of competing services available, businesses need to focus efforts on maintaining continuous customer satisfaction, rewarding consumer loyalty and minimizing customer attrition. In addition, small increases in customer retention have been shown to increase profits disproportionately. One study concluded that a 5% increase in customer retention rates will increase profits by 25% to 95%. Businesses tend to respond to customer attrition on a reactive basis, acting only after the customer has initiated the process to terminate service. At this stage, the chance of changing the customer's decision is almost zero. Proper application of predictive analytics can lead to a more proactive retention strategy. By a frequent examination of a customer's past service usage, service performance, spending and other behavior patterns, predictive models can determine the likelihood of a customer terminating service sometime soon. An intervention with lucrative offers

can increase the chance of retaining the customer. Silent attrition, the behavior of a customer to slowly but steadily reduce usage, is another problem that many companies face. Predictive analytics can also predict this behavior, so that the company can take proper actions to increase customer activity.

Direct Marketing

When marketing consumer products and services, there is the challenge of keeping up with competing products and consumer behavior. Apart from identifying prospects, predictive analytics can also help to identify the most effective combination of product versions, marketing material, communication channels and timing that should be used to target a given consumer. The goal of predictive analytics is typically to lower the cost per order or cost per action.

Fraud Detection

Fraud is a big problem for many businesses and can be of various types: inaccurate credit applications, fraudulent transactions (both offline and online), identity thefts and false insurance claims. These problems plague firms of all sizes in many industries. Some examples of likely victims are credit card issuers, insurance companies, retail merchants, manufacturers, business-to-business suppliers and even services providers. A predictive model can help weed out the "bads" and reduce a business's exposure to fraud.

Predictive modeling can also be used to identify high-risk fraud candidates in business or the public sector. Mark Nigrini developed a risk-scoring method to identify audit targets. He describes the use of this approach to detect fraud in the franchisee sales reports of an international fast-food chain. Each location is scored using 10 predictors. The 10 scores are then weighted to give one final overall risk score for each location. The same scoring approach was also used to identify high-risk check kiting accounts, potentially fraudulent travel agents, and questionable vendors. A reasonably complex model was used to identify fraudulent monthly reports submitted by divisional controllers.

The Internal Revenue Service (IRS) of the United States also uses predictive analytics to mine tax returns and identify tax fraud.

Recent advancements in technology have also introduced predictive behavior analysis for web fraud detection. This type of solution utilizes heuristics in order to study normal web user behavior and detect anomalies indicating fraud attempts.

Portfolio, Product or Economy-level Prediction

Often the focus of analysis is not the consumer but the product, portfolio, firm, industry or even the economy. For example, a retailer might be interested in predicting store-level demand for inventory management purposes. Or the Federal Reserve Board might be interested in predicting the unemployment rate for the next year. These types of problems can be addressed by predictive analytics using time series techniques. They can also be addressed via machine learning approaches which transform the original time series into a feature vector space, where the learning algorithm finds patterns that have predictive power.

Project Risk Management

When employing risk management techniques, the results are always to predict and benefit from a future scenario. The capital asset pricing model (CAP-M) "predicts" the best portfolio to maximize return. Probabilistic risk assessment (PRA) when combined with mini-Delphi techniques and statistical approaches yields accurate forecasts. These are examples of approaches that can extend from project to market, and from near to long term. Underwriting and other business approaches identify risk management as a predictive method.

Underwriting

Many businesses have to account for risk exposure due to their different services and determine the cost needed to cover the risk. For example, auto insurance providers need to accurately determine the amount of premium to charge to cover each automobile and driver. A financial company needs to assess a borrower's potential and ability to pay before granting a loan. For a health insurance provider, predictive analytics can analyze a few years of past medical claims data, as well as lab, pharmacy and other records where available, to predict how expensive an enrollee is likely to be in the future. Predictive analytics can help underwrite these quantities by predicting the chances of illness, default, bankruptcy, etc. Predictive analytics can streamline the process of customer acquisition by predicting the future risk behavior of a customer using application level data. Predictive analytics in the form of credit scores have reduced the amount of time it takes for loan approvals, especially in the mortgage market where lending decisions are now made in a matter of hours rather than days or even weeks. Proper predictive analytics can lead to proper pricing decisions, which can help mitigate future risk of default.

Technology and Big Data Influences

Big data is a collection of data sets that are so large and complex that they become awkward to work with using traditional database management tools. The volume, variety and velocity of big data have introduced challenges across the board for capture, storage, search, sharing, analysis, and visualization. Examples of big data sources include web logs, RFID, sensor data, social networks, Internet search indexing, call detail records, military surveillance, and complex data in astronomic, biogeochemical, genomics, and atmospheric sciences. Big Data is the core of most predictive analytic services offered by IT organizations. Thanks to technological advances in computer hardware—faster CPUs, cheaper memory, and MPP architectures—and new technologies such as Hadoop, MapReduce, and in-database and text analytics for processing big data, it is now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for new insights. It is also possible to run predictive algorithms on streaming data. Today, exploring big data and using predictive analytics is within reach of more organizations than ever before and new methods that are capable for handling such datasets are proposed

Analytical Techniques

The approaches and techniques used to conduct predictive analytics can broadly be grouped into regression techniques and machine learning techniques.

Regression Techniques

Regression models are the mainstay of predictive analytics. The focus lies on establishing a mathematical equation as a model to represent the interactions between the different variables in consideration. Depending on the situation, there are a wide variety of models that can be applied while performing predictive analytics. Some of them are briefly discussed below.

Linear Regression Model

The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. This relationship is expressed as an equation that predicts the response variable as a linear function of the parameters. These parameters are adjusted so that a measure of fit is optimized. Much of the effort in model fitting is focused on minimizing the size of the residual, as well as ensuring that it is randomly distributed with respect to the model predictions.

The goal of regression is to select the parameters of the model so as to minimize the sum of the squared residuals. This is referred to as ordinary least squares (OLS) estimation and results in best linear unbiased estimates (BLUE) of the parameters if and only if the Gauss-Markov assumptions are satisfied.

Once the model has been estimated we would be interested to know if the predictor variables belong in the model—i.e. is the estimate of each variable's contribution reliable? To do this we can check the statistical significance of the model's coefficients which can be measured using the t-statistic. This amounts to testing whether the coefficient is significantly different from zero. How well the model predicts the dependent variable based on the value of the independent variables can be assessed by using the R^2 statistic. It measures predictive power of the model i.e. the proportion of the total variation in the dependent variable that is "explained" (accounted for) by variation in the independent variables.

Discrete Choice Models

Multivariate regression (above) is generally used when the response variable is continuous and has an unbounded range. Often the response variable may not be continuous but rather discrete. While mathematically it is feasible to apply multivariate regression to discrete ordered dependent variables, some of the assumptions behind the theory of multivariate linear regression no longer hold, and there are other techniques such as discrete choice models which are better suited for this type of analysis. If the dependent variable is discrete, some of those superior methods are logistic regression, multinomial logit and probit models. Logistic regression and probit models are used when the dependent variable is binary.

Logistic Regression

In a classification setting, assigning outcome probabilities to observations can be achieved through the use of a logistic model, which is basically a method which transforms information about the binary dependent variable into an unbounded continuous variable and estimates a regular multivariate model.

The Wald and likelihood-ratio test are used to test the statistical significance of each coefficient *b* in the model. A test assessing the goodness-of-fit of a classification model is the "percentage correctly predicted".

Multinomial Logistic Regression

An extension of the binary logit model to cases where the dependent variable has more than 2 categories is the multinomial logit model. In such cases collapsing the data into two categories might not make good sense or may lead to loss in the richness of the data. The multinomial logit model is the appropriate technique in these cases, especially when the dependent variable categories are not ordered (for examples colors like red, blue, green). Some authors have extended multinomial regression to include feature selection/importance methods such as random multinomial logit.

Probit Regression

Probit models offer an alternative to logistic regression for modeling categorical dependent variables. Even though the outcomes tend to be similar, the underlying distributions are different. Probit models are popular in social sciences like economics.

A good way to understand the key difference between probit and logit models is to assume that the dependent variable is driven by a latent variable z, which is a sum of a linear combination of explanatory variables and a random noise term.

We do not observe z but instead observe y which takes the value 0 (when $z < 0$) or 1 (otherwise). In the logit model we assume that the random noise term follows a logistic distribution with mean zero. In the probit model we assume that it follows a normal distribution with mean zero. Note that in social sciences (e.g. economics), probit is often used to model situations where the observed variable y is continuous but takes values between 0 and 1.

Logit Versus Probit

The probit model has been around longer than the logit model. They behave similarly, except that the logistic distribution tends to be slightly flatter tailed. One of the reasons the logit model was formulated was that the probit model was computationally difficult due to the requirement of numerically calculating integrals. Modern computing however has made this computation fairly simple. The coefficients obtained from the logit and probit model are fairly close. However, the odds ratio is easier to interpret in the logit model.

Practical reasons for choosing the probit model over the logistic model would be:

- There is a strong belief that the underlying distribution is normal
- The actual event is not a binary outcome (*e.g.*, bankruptcy status) but a proportion (*e.g.*, proportion of population at different debt levels).

Time Series Models

Time series models are used for predicting or forecasting the future behavior of variables. These models account for the fact that data points taken over time may have an internal structure (such

as autocorrelation, trend or seasonal variation) that should be accounted for. As a result, standard regression techniques cannot be applied to time series data and methodology has been developed to decompose the trend, seasonal and cyclical component of the series. Modeling the dynamic path of a variable can improve forecasts since the predictable component of the series can be projected into the future.

Time series models estimate difference equations containing stochastic components. Two commonly used forms of these models are autoregressive models (AR) and moving-average (MA) models. The Box–Jenkins methodology (1976) developed by George Box and G.M. Jenkins combines the AR and MA models to produce the ARMA (autoregressive moving average) model which is the cornerstone of stationary time series analysis. ARIMA (autoregressive integrated moving average models) on the other hand are used to describe non-stationary time series. Box and Jenkins suggest differencing a non stationary time series to obtain a stationary series to which an ARMA model can be applied. Non stationary time series have a pronounced trend and do not have a constant long-run mean or variance.

Box and Jenkins proposed a three-stage methodology which includes: model identification, estimation and validation. The identification stage involves identifying if the series is stationary or not and the presence of seasonality by examining plots of the series, autocorrelation and partial autocorrelation functions. In the estimation stage, models are estimated using non-linear time series or maximum likelihood estimation procedures. Finally the validation stage involves diagnostic checking such as plotting the residuals to detect outliers and evidence of model fit.

In recent years time series models have become more sophisticated and attempt to model conditional heteroskedasticity with models such as ARCH (autoregressive conditional heteroskedasticity) and GARCH (generalized autoregressive conditional heteroskedasticity) models frequently used for financial time series. In addition time series models are also used to understand inter-relationships among economic variables represented by systems of equations using VAR (vector autoregression) and structural VAR models.

Survival or Duration Analysis

Survival analysis is another name for time to event analysis. These techniques were primarily developed in the medical and biological sciences, but they are also widely used in the social sciences like economics, as well as in engineering (reliability and failure time analysis).

Censoring and non-normality, which are characteristic of survival data, generate difficulty when trying to analyze the data using conventional statistical models such as multiple linear regression. The normal distribution, being a symmetric distribution, takes positive as well as negative values, but duration by its very nature cannot be negative and therefore normality cannot be assumed when dealing with duration/survival data. Hence the normality assumption of regression models is violated.

The assumption is that if the data were not censored it would be representative of the population of interest. In survival analysis, censored observations arise whenever the dependent variable of interest represents the time to a terminal event, and the duration of the study is limited in time.

An important concept in survival analysis is the hazard rate, defined as the probability that the

event will occur at time t conditional on surviving until time t. Another concept related to the hazard rate is the survival function which can be defined as the probability of surviving to time t.

Most models try to model the hazard rate by choosing the underlying distribution depending on the shape of the hazard function. A distribution whose hazard function slopes upward is said to have positive duration dependence, a decreasing hazard shows negative duration dependence whereas constant hazard is a process with no memory usually characterized by the exponential distribution. Some of the distributional choices in survival models are: F, gamma, Weibull, log normal, inverse normal, exponential etc. All these distributions are for a non-negative random variable.

Duration models can be parametric, non-parametric or semi-parametric. Some of the models commonly used are Kaplan-Meier and Cox proportional hazard model (non parametric).

Classification and Regression Trees (CART)

Globally-optimal classification tree analysis (GO-CTA) (also called hierarchical optimal discriminant analysis) is a generalization of optimal discriminant analysis that may be used to identify the statistical model that has maximum accuracy for predicting the value of a categorical dependent variable for a dataset consisting of categorical and continuous variables. The output of HODA is a non-orthogonal tree that combines categorical variables and cut points for continuous variables that yields maximum predictive accuracy, an assessment of the exact Type I error rate, and an evaluation of potential cross-generalizability of the statistical model. Hierarchical optimal discriminant analysis may be thought of as a generalization of Fisher's linear discriminant analysis. Optimal discriminant analysis is an alternative to ANOVA (analysis of variance) and regression analysis, which attempt to express one dependent variable as a linear combination of other features or measurements. However, ANOVA and regression analysis give a dependent variable that is a numerical variable, while hierarchical optimal discriminant analysis gives a dependent variable that is a class variable.

Classification and regression trees (CART) are a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively.

Decision trees are formed by a collection of rules based on variables in the modeling data set:

- Rules based on variables' values are selected to get the best split to differentiate observations based on the dependent variable
- Once a rule is selected and splits a node into two, the same process is applied to each "child" node (i.e. it is a recursive procedure)
- Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. (Alternatively, the data are split as much as possible and then the tree is later pruned.)

Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

A very popular method for predictive analytics is Leo Breiman's Random forests.

Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) is a non-parametric technique that builds flexible models by fitting piecewise linear regressions.

An important concept associated with regression splines is that of a knot. Knot is where one local regression model gives way to another and thus is the point of intersection between two splines.

In multivariate and adaptive regression splines, basis functions are the tool used for generalizing the search for knots. Basis functions are a set of functions used to represent the information contained in one or more variables. Multivariate and Adaptive Regression Splines model almost always creates the basis functions in pairs.

Multivariate and adaptive regression spline approach deliberately overfits the model and then prunes to get to the optimal model. The algorithm is computationally very intensive and in practice we are required to specify an upper limit on the number of basis functions.

Machine Learning Techniques

Machine learning, a branch of artificial intelligence, was originally employed to develop techniques to enable computers to learn. Today, since it includes a number of advanced statistical methods for regression and classification, it finds application in a wide variety of fields including medical diagnostics, credit card fraud detection, face and speech recognition and analysis of the stock market. In certain applications it is sufficient to directly predict the dependent variable without focusing on the underlying relationships between variables. In other cases, the underlying relationships can be very complex and the mathematical form of the dependencies unknown. For such cases, machine learning techniques emulate human cognition and learn from training examples to predict future events.

A brief discussion of some of these methods used commonly for predictive analytics is provided below. A detailed study of machine learning can be found in Mitchell (1997).

Neural Networks

Neural networks are nonlinear sophisticated modeling techniques that are able to model complex functions. They can be applied to problems of prediction, classification or control in a wide spectrum of fields such as finance, cognitive psychology/neuroscience, medicine, engineering, and physics.

Neural networks are used when the exact nature of the relationship between inputs and output is not known. A key feature of neural networks is that they learn the relationship between inputs and output through training. There are three types of training in neural networks used by different networks, supervised and unsupervised training, reinforcement learning, with supervised being the most common one.

Some examples of neural network training techniques are backpropagation, quick propagation, conjugate gradient descent, projection operator, Delta-Bar-Delta etc. Some unsupervised network architectures are multilayer perceptrons, Kohonen networks, Hopfield networks, etc.

Multilayer Perceptron (MLP)

The multilayer perceptron (MLP) consists of an input and an output layer with one or more hidden layers of nonlinearly-activating nodes or sigmoid nodes. This is determined by the weight vector and it is necessary to adjust the weights of the network. The backpropagation employs gradient fall to minimize the squared error between the network output values and desired values for those outputs. The weights adjusted by an iterative process of repetitive present of attributes. Small changes in the weight to get the desired values are done by the process called training the net and is done by the training set (learning rule).

Radial Basis Functions

A radial basis function (RBF) is a function which has built into it a distance criterion with respect to a center. Such functions can be used very efficiently for interpolation and for smoothing of data. Radial basis functions have been applied in the area of neural networks where they are used as a replacement for the sigmoidal transfer function. Such networks have 3 layers, the input layer, the hidden layer with the RBF non-linearity and a linear output layer. The most popular choice for the non-linearity is the Gaussian. RBF networks have the advantage of not being locked into local minima as do the feed-forward networks such as the multilayer perceptron.

Support Vector Machines

support vector machines (SVM) are used to detect and exploit complex patterns in data by clustering, classifying and ranking the data. They are learning machines that are used to perform binary classifications and regression estimations. They commonly use kernel based methods to apply linear classification techniques to non-linear classification problems. There are a number of types of SVM such as linear, polynomial, sigmoid etc.

Naïve Bayes

Naïve Bayes based on Bayes conditional probability rule is used for performing classification tasks. Naïve Bayes assumes the predictors are statistically independent which makes it an effective classification tool that is easy to interpret. It is best employed when faced with the problem of 'curse of dimensionality' i.e. when the number of predictors is very high.

k-nearest Neighbours

The nearest neighbour algorithm (KNN) belongs to the class of pattern recognition statistical methods. The method does not impose a priori any assumptions about the distribution from which the modeling sample is drawn. It involves a training set with both positive and negative values. A new sample is classified by calculating the distance to the nearest neighbouring training case. The sign of that point will determine the classification of the sample. In the k-nearest neighbour classifier, the k nearest points are considered and the sign of the majority is used to classify the sample. The performance of the kNN algorithm is influenced by three main factors: (1) the distance measure used to locate the nearest neighbours; (2) the decision rule used to derive a classification from the k-nearest neighbours; and (3) the number of neighbours used to classify the new sample. It can be proved that, unlike other methods, this method is universally asymptotically convergent, i.e.: as

the size of the training set increases, if the observations are independent and identically distributed (i.i.d.), regardless of the distribution from which the sample is drawn, the predicted class will converge to the class assignment that minimizes misclassification error.

Geospatial Predictive Modeling

Conceptually, geospatial predictive modeling is rooted in the principle that the occurrences of events being modeled are limited in distribution. Occurrences of events are neither uniform nor random in distribution—there are spatial environment factors (infrastructure, sociocultural, topographic, etc.) that constrain and influence where the locations of events occur. Geospatial predictive modeling attempts to describe those constraints and influences by spatially correlating occurrences of historical geospatial locations with environmental factors that represent those constraints and influences. Geospatial predictive modeling is a process for analyzing events through a geographic filter in order to make statements of likelihood for event occurrence or emergence.

Tools

Historically, using predictive analytics tools—as well as understanding the results they delivered—required advanced skills. However, modern predictive analytics tools are no longer restricted to IT specialists. As more organizations adopt predictive analytics into decision-making processes and integrate it into their operations, they are creating a shift in the market toward business users as the primary consumers of the information. Business users want tools they can use on their own. Vendors are responding by creating new software that removes the mathematical complexity, provides user-friendly graphic interfaces and/or builds in short cuts that can, for example, recognize the kind of data available and suggest an appropriate predictive model. Predictive analytics tools have become sophisticated enough to adequately present and dissect data problems, so that any data-savvy information worker can utilize them to analyze data and retrieve meaningful, useful results. For example, modern tools present findings using simple charts, graphs, and scores that indicate the likelihood of possible outcomes.

There are numerous tools available in the marketplace that help with the execution of predictive analytics. These range from those that need very little user sophistication to those that are designed for the expert practitioner. The difference between these tools is often in the level of customization and heavy data lifting allowed.

Notable open source predictive analytic tools include:

- Apache Mahout
- GNU Octave
- KNIME
- OpenNN
- Orange
- R
- scikit-learn

- Weka

Notable commercial predictive analytic tools include:

- Alpine Data Labs
- Alteryx
- Angoss KnowledgeSTUDIO
- BIRT Analytics
- IBM SPSS Statistics and IBM SPSS Modeler
- KXEN Modeler
- Mathematica
- MATLAB
- Minitab
- LabVIEW
- Neural Designer
- Oracle Advanced Analytics
- Pervasive
- Predixion Software
- RapidMiner
- RCASE
- Revolution Analytics
- SAP HANA and SAP BusinessObjects Predictive Analytics
- SAS and SAS Enterprise Miner
- STATA
- Statgraphics
- STATISTICA
- TeleRetail
- TIBCO

Beside these software packages, specific tools have also been developed for industrial applications. For example, Watchdog Agent Toolbox has been developed and optimized for predictive analysis in prognostics and health management applications and is available for MATLAB and LabVIEW.

The most popular commercial predictive analytics software packages according to the Rexer Analytics Survey for 2013 are IBM SPSS Modeler, SAS Enterprise Miner, and Dell Statistica.

PMML

In an attempt to provide a standard language for expressing predictive models, the Predictive Model Markup Language (PMML) has been proposed. Such an XML-based language provides a way for the different tools to define predictive models and to share these between PMML compliant applications. PMML 4.0 was released in June, 2009.

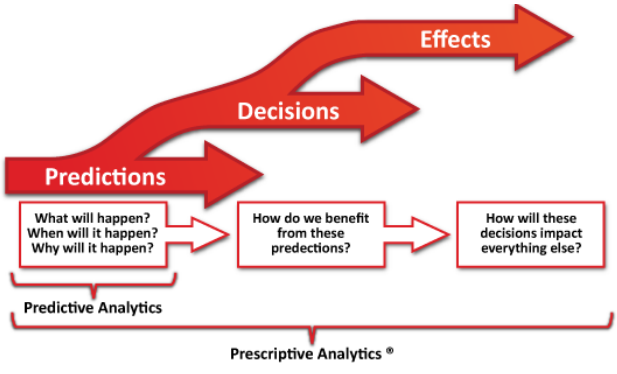
Criticism

There are plenty of skeptics when it comes to computers and algorithms abilities to predict the future, including Gary King, a professor from Harvard University and the director of the Institute for Quantitative Social Science. People are influenced by their environment in innumerable ways. Trying to understand what people will do next assumes that all the influential variables can be known and measured accurately. "People's environments change even more quickly than they themselves do. Everything from the weather to their relationship with their mother can change the way people think and act. All of those variables are unpredictable. How they will impact a person is even less predictable. If put in the exact same situation tomorrow, they may make a completely different decision. This means that a statistical prediction is only valid in sterile laboratory conditions, which suddenly isn't as useful as it seemed before."

Prescriptive Analytics

Prescriptive analytics is the third and final phase of analytics (BA) which also includes descriptive and predictive analytics.

Referred to as the "final frontier of analytic capabilities," prescriptive analytics entails the application of mathematical and computational sciences suggests decision options to take advantage of the results of descriptive and predictive analytics. The first stage of business analytics is descriptive analytics, which still accounts for the majority of all business analytics today. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Most management reporting - such as sales, marketing, operations, and finance - uses this type of post-mortem analysis.



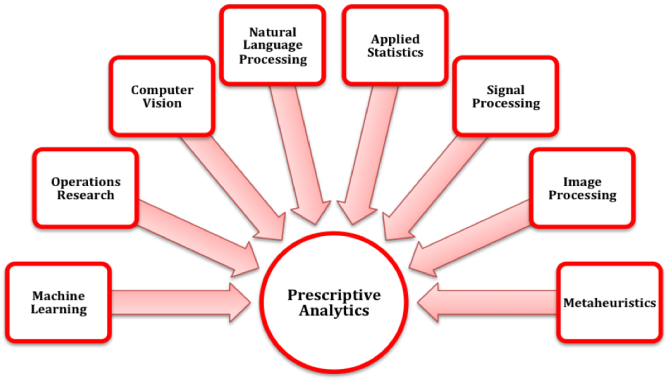
Prescriptive Analytics extends beyond predictive analytics by specifying both the actions necessary to achieve predicted outcomes, and the interrelated effects of each decision

The next phase is predictive analytics. Predictive analytics answers the question what is likely to happen. This is when historical data is combined with rules, algorithms, and occasionally external data to determine the probable future outcome of an event or the likelihood of a situation occurring. The final phase is prescriptive analytics, which goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the implications of each decision option.

Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen. Further, prescriptive analytics suggests decision options on how to take advantage of a future opportunity or mitigate a future risk and shows the implication of each decision option. Prescriptive analytics can continually take in new data to re-predict and re-prescribe, thus automatically improving prediction accuracy and prescribing better decision options. Prescriptive analytics ingests hybrid data, a combination of structured (numbers, categories) and unstructured data (videos, images, sounds, texts), and business rules to predict what lies ahead and to prescribe how to take advantage of this predicted future without compromising other priorities.

All three phases of analytics can be performed through professional services or technology or a combination. In order to scale, prescriptive analytics technologies need to be adaptive to take into account the growing volume, velocity, and variety of data that most mission critical processes and their environments may produce.

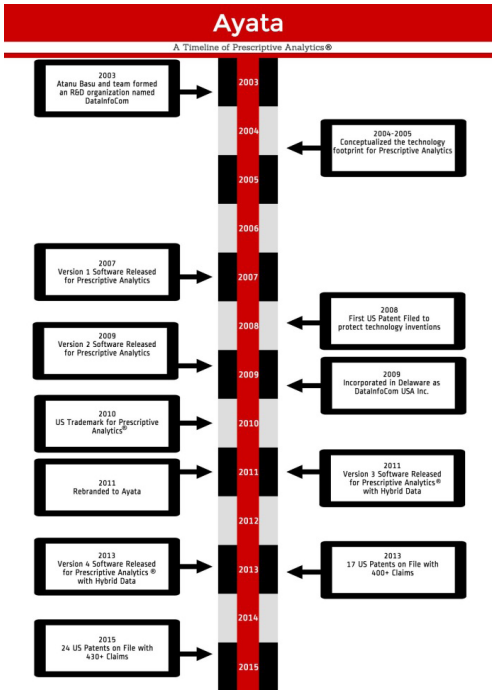
One criticism of prescriptive analytics is that its distinction from predictive analytics is ill-defined and therefore ill-conceived.



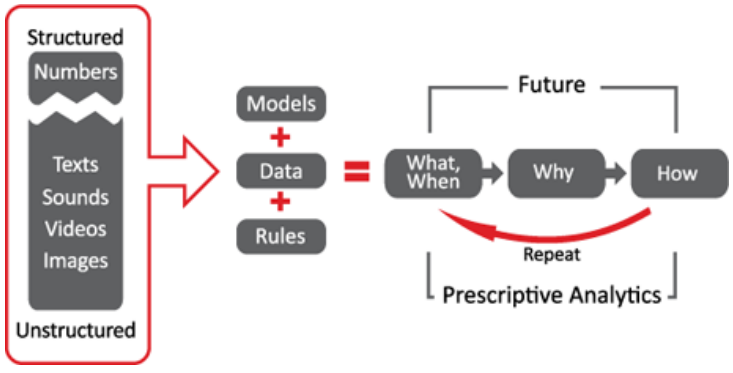
The scientific disciplines that comprise Prescriptive Analytics

History

While the term Prescriptive Analytics, first coined by IBM and later trademarked by Ayata, the underlying concepts have been around for hundreds of years. The technology behind prescriptive analytics synergistically combines hybrid data, business rules with mathematical models and computational models. The data inputs to prescriptive analytics may come from multiple sources: internal, such as inside a corporation; and external, also known as environmental data. The data may be structured, which includes numbers and categories, as well as unstructured data, such as texts, images, sounds, and videos. Unstructured data differs from structured data in that its format varies widely and cannot be stored in traditional relational databases without significant effort at data transformation. More than 80% of the world's data today is unstructured, according to IBM.



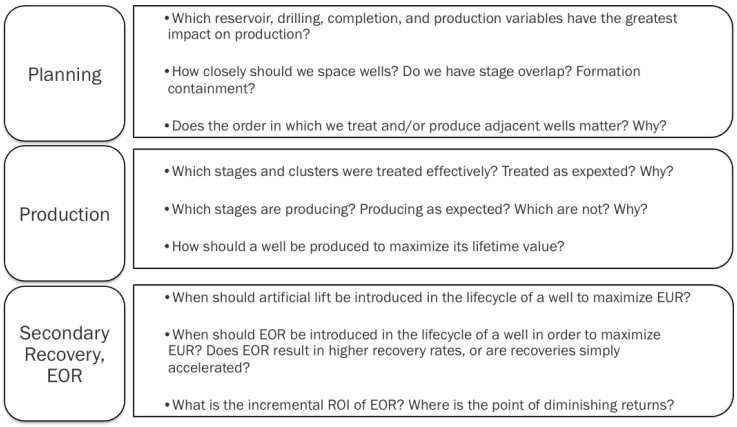
Timeline tracing evolution of Prescriptive Analytics capability and software



Prescriptive Analytics incorporates both structured and unstructured data, and uses a combination of advanced analytic techniques and disciplines to predict, prescribe, and adapt.

In addition to this variety of data types and growing data volume, incoming data can also evolve with respect to velocity, that is, more data being generated at a faster or a variable pace. Business rules define the business process and include objectives constraints, preferences, policies, best practices, and boundaries. Mathematical models and computational models are techniques derived from mathematical sciences, computer science and related disciplines such as applied statistics, machine learning, operations research, natural language processing, computer vision, pattern recognition, image processing, speech recognition, and signal processing. The correct application of all these methods and the verification of their results implies the need for resources on a massive scale including human, computational and temporal for every Prescriptive Analytic project. In order to spare the expense of dozens of people, high performance machines and weeks of work one must consider the reduction of resources and therefore a reduction in the accuracy or reliability of the outcome. The preferable route is a reduction that produces a probabilistic result within acceptable limits.

Applications in Oil and Gas

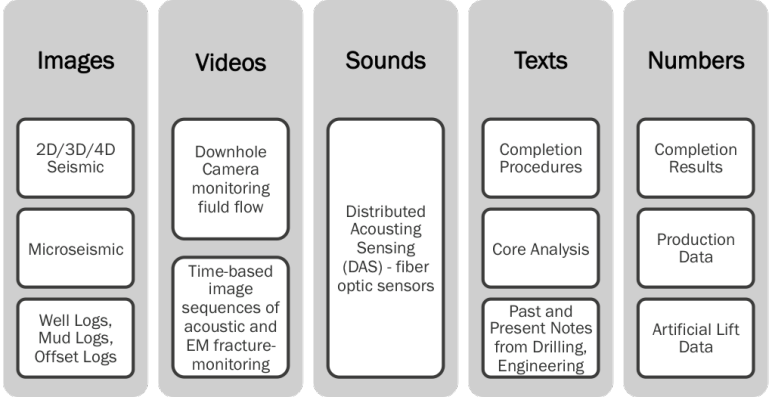


Key Questions Prescriptive Analytics software answers for oil and gas producers

Energy is the largest industry in the world ($6 trillion in size). The processes and decisions related to oil and natural gas exploration, development and production generate large amounts of data. Many types of captured data are used to create models and images of the Earth's structure and layers 5,000 - 35,000 feet below the surface and to describe activities around the wells themselves, such as depositional characteristics, machinery performance, oil flow rates, reservoir temperatures and pressures. Prescriptive analytics software can help with both locating and producing hydrocarbons

by taking in seismic data, well log data, production data, and other related data sets to prescribe specific recipes for how and where to drill, complete, and produce wells in order to optimize recovery, minimize cost, and reduce environmental footprint.

Unconventional Resource Development



Examples of structured and unstructured data sets generated and by the oil and gas companies and their ecosystem of service providers that can be analyzed together using Prescriptive Analytics software

With the value of the end product determined by global commodity economics, the basis of competition for operators in upstream E&P is the ability to effectively deploy capital to locate and extract resources more efficiently, effectively, predictably, and safely than their peers. In unconventional resource plays, operational efficiency and effectiveness is diminished by reservoir inconsistencies,

and decision-making impaired by high degrees of uncertainty. These challenges manifest themselves in the form of low recovery factors and wide performance variations.

Prescriptive Analytics software can accurately predict production and prescribe optimal configurations of controllable drilling, completion, and production variables by modeling numerous internal and external variables simultaneously, regardless of source, structure, size, or format. Prescriptive analytics software can also provide decision options and show the impact of each decision option so the operations managers can proactively take appropriate actions, on time, to guarantee future exploration and production performance, and maximize the economic value of assets at every point over the course of their serviceable lifetimes.

Oilfield Equipment Maintenance

In the realm of oilfield equipment maintenance, Prescriptive Analytics can optimize configuration, anticipate and prevent unplanned downtime, optimize field scheduling, and improve maintenance planning. According to General Electric, there are more than 130,000 electric submersible pumps (ESP's) installed globally, accounting for 60% of the world's oil production. Prescriptive Analytics has been deployed to predict when and why an ESP will fail, and recommend the necessary actions to prevent the failure.

In the area of Health, Safety, and Environment, prescriptive analytics can predict and preempt incidents that can lead to reputational and financial loss for oil and gas companies.

Pricing

Pricing is another area of focus. Natural gas prices fluctuate dramatically depending upon supply, demand, econometrics, geopolitics, and weather conditions. Gas producers, pipeline transmission companies and utility firms have a keen interest in more accurately predicting gas prices so that they can lock in favorable terms while hedging downside risk. Prescriptive analytics software can accurately predict prices by modeling internal and external variables simultaneously and also provide decision options and show the impact of each decision option.

Applications in Healthcare

Multiple factors are driving healthcare providers to dramatically improve business processes and operations as the United States healthcare industry embarks on the necessary migration from a largely fee-for service, volume-based system to a fee-for-performance, value-based system. Prescriptive analytics is playing a key role to help improve the performance in a number of areas involving various stakeholders: payers, providers and pharmaceutical companies.

Prescriptive analytics can help providers improve effectiveness of their clinical care delivery to the population they manage and in the process achieve better patient satisfaction and retention. Providers can do better population health management by identifying appropriate intervention models for risk stratified population combining data from the in-facility care episodes and home based telehealth.

Prescriptive analytics can also benefit healthcare providers in their capacity planning by using analytics to leverage operational and usage data combined with data of external factors such as

economic data, population demographic trends and population health trends, to more accurately plan for future capital investments such as new facilities and equipment utilization as well as understand the trade-offs between adding additional beds and expanding an existing facility versus building a new one.

Prescriptive analytics can help pharmaceutical companies to expedite their drug development by identifying patient cohorts that are most suitable for the clinical trials worldwide - patients who are expected to be compliant and will not drop out of the trial due to complications. Analytics can tell companies how much time and money they can save if they choose one patient cohort in a specific country vs. another.

In provider-payer negotiations, providers can improve their negotiating position with health insurers by developing a robust understanding of future service utilization. By accurately predicting utilization, providers can also better allocate personnel.

Social Media Analytics

Social Media Analytics as a part of social analytics is the process of gathering data from stakeholder conversations on digital media and processing into structured insights leading to more information-driven business decisions and increased customer centrality for brands and businesses.

Social media analytics can also be referred as social media listening, social media monitoring or social media intelligence.

Digital media sources for social media analytics include social media channels, blogs, forums, image sharing sites, video sharing sites, aggregators, classifieds, complaints, Q&A, reviews, Wikipedia and others.

Social media analytics is an industry agnostic practice and is commonly used in different approaches on business decisions, marketing, customer service, reputation management, sales and others. There is an array of tools that offers the social media analysis, varying from the level of business requirement. Logic behind algorithms that are designed for these tools is selection, data pre-processing, transformation, mining and hidden pattern evaluation.

In order to make the complete process of social media analysis a success it is important that key performance indicators (KPIs) for objectively evaluating the data is defined.

Social media analytics is important when one needs to understand the patterns that are hidden in large amount of social data related to particular brands.

Homophily is used as a apart of analytics, it is a tendency that a contact between similar people occurs at a higher rate than among dissimilar people. According to research, two users who follow reciprocally share topical interests by mining their thousands of links. All these are used for taking major business decision in social media sectors.

The success of social media monitoring (SMM) tools may vary from one company to another. According to Soleman and Cohard (2016), beyond technical factors related to social media moni-

toring (SMM) (quality of sources, functionalities, quality of the tool), organizations must also take into account the need for new capabilities, human, managerial and organizational skills to take advantage of their SMM tools.

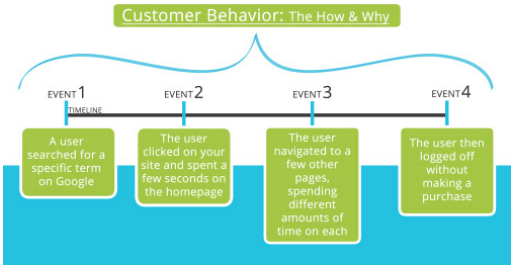
Behavioral Analytics

Behavioral analytics is a recent advancement in business analytics that reveals new insights into the behavior of consumers on eCommerce platforms, online games, web and mobile applications, and IoT. The rapid increase in the volume of raw event data generated by the digital world enables methods that go beyond typical analysis by demographics and other traditional metrics that tell us what kind of people took what actions in the past. Behavioral analysis focuses on understanding how consumers act and why, enabling accurate predictions about how they are likely to act in the future. It enables marketers to make the right offers to the right consumer segments at the right time.

Behavioral analytics utilizes the massive volumes of raw user event data captured during sessions in which consumers use application, game, or website, including traffic data like navigation path, clicks, social media interactions, purchasing decisions and marketing responsiveness. Also, the event-data can include advertising metrics like click-to-conversion time, as well as comparisons between other metrics like the monetary value of an order and the amount of time spent on the site. These data points are then compiled and analyzed, whether by looking at session progression from when a user first entered the platform until a sale was made, or what other products a user bought or looked at before this purchase. Behavioral analysis allows future actions and trends to be predicted based on the collection of such data.

While business analytics has a more broad focus on the who, what, where and when of business intelligence, behavioral analytics narrows that scope, allowing one to take seemingly unrelated data points in order to extrapolate, predict and determine errors and future trends. It takes a more holistic and human view of data, connecting individual data points to tell us not only what is happening, but also how and why it is happening.

Examples and Real World Applications



Visual Representation of Events that Make Up Behavioral Analysis

Data shows that a large percentage of users using a certain eCommerce platform found it by searching for “Thai food” on Google. After landing on the homepage, most people spent some time on the “Asian Food” page and then logged off without placing an order. Looking at each of these events

as separate data points does not represent what is really going on and why people did not make a purchase. However, viewing these data points as a representation of overall user behavior enables one to interpolate how and why users acted in this particular case.

Behavioral analytics looks at all site traffic and page views as a timeline of connected events that did not lead to orders. Since most users left after viewing the "Asian Food" page, there could be a disconnect between what they are searching for on Google and what the "Asian Food" page displays. Knowing this, a quick look at the "Asian Food" page reveals that it does not display Thai food prominently and thus people do not think it is actually offered, even though it is.

Behavioral analytics is becoming increasingly popular in commercial environments. Amazon.com is a leader in using behavioral analytics to recommend additional products that customers are likely to buy based on their previous purchasing patterns on the site. Behavioral analytics is also used by Target to suggest products to customers in their retail stores, while political campaigns use it to determine how potential voters should be approached. In addition to retail and political applications, behavioral analytics is also used by banks and manufacturing firms to prioritize leads generated by their websites. Behavioral analytics also allow developers to manage users in online-gaming and web applications.

Types

- Ecommerce and retail – Product recommendations and predicting future sales trends
- Online gaming – Predicting usage trends, load, and user preferences in future releases
- Application development – Determining how users use an application to predict future usage and preferences.
- Cohort analysis – Breaking users down into similar groups to gain a more focused understanding of their behavior.
- Security – Detecting compromised credentials and insider threats by locating anomalous behavior.
- Suggestions – People who liked this also liked...
- Presentation of relevant content based on user behavior.

Components of Behavioral Analytics

An ideal behavioral analytics solution would include:

- Real-time capture of vast volumes of raw event data across all relevant digital devices and applications used during sessions
- Automatic aggregation of raw event data into relevant data sets for rapid access, filtering and analysis
- Ability to query data in an unlimited number of ways, enabling users to ask any business question
- Extensive library of built-in analysis functions such as cohort, path and funnel analysis

- A visualization component

Subsets of Behavioral Analytics

Path Analysis (Computing)

Path analysis, is the analysis of a path, which is a portrayal of a chain of consecutive events that a given user or cohort performs during a set period of time while using a website, online game, or eCommerce platform. As a subset of behavioral analytics, path analysis is a way to understand user behavior in order to gain actionable insights into the data. Path analysis provides a visual portrayal of every event a user or cohort performs as part of a path during a set period of time.

While it is possible to track a user's path through the site, and even show that path as a visual representation, the real question is how to gain these actionable insights. If path analysis simply outputs a "pretty" graph, while it may look nice, it does not provide anything concrete to act upon.

Examples

In order to get the most out of path analysis the first step would be to determine what needs to be analyzed and what are the goals of the analysis. A company might be trying to figure out why their site is running slow, are certain types of users interested in certain pages or products, or if their user interface is set up in a logical way.

Now that the goal has been set there are a few ways of performing the analysis. If a large percentage of a certain cohort, people between the ages of 18-25, logs into an online game, creates a profile and then spends the next 10 minutes wandering around the menu page, then it may be that the user interface is not logical. By seeing this group of users following the path that they did a developer will be able to analyze the data and realize that after creating a profile, the "play game" button does not appear. Thus, path analysis was able to provide actionable data for the company to act on and fix an error.

In eCommerce, path analysis can help customize a shopping experience to each user. By looking at what products other customers in a certain cohort looked at before buying one, a company can suggest "items you may also like" to the next customer and increase the chances of them making a purchase. Also, path analysis can help solve performance issues on a platform. For example, a company looks at a path and realizes that their site freezes up after a certain combinations of events. By analyzing the path and the progression of events that led to the error, the company can pinpoint the error and fix it.

Evolution

Historically path analysis fell under the broad category of website analytics, and related only to the analysis of paths through websites. Path analysis in website analytics is a process of determining a sequence of pages visited in a visitor session prior to some desired event, such as the visitor purchasing an item or requesting a newsletter. The precise order of pages visited may or may not be important and may or may not be specified. In practice, this analysis is done in aggregate, ranking the paths (sequences of pages) visited prior to the desired event, by descending frequency of use. The idea is to determine what features of the website encourage the desired result. "Fallout anal-

ysis," a subset of path analysis, looks at "black holes" on the site, or paths that lead to a dead end most frequently, paths or features that confuse or lose potential customers.

With the advent of big data along with web based applications, online games, and eCommerce platforms, path analysis has come to include much more than just web path analysis. Understanding how users move through an app, game, or other web platform are all part of modern-day path analysis.

Understanding Visitors

In the real world when you visit a shop the shelves and products are not placed in a random order. The shop owner carefully analyzes the visitors and path they walk through the shop, especially when they are selecting or buying products. Next the shop owner will reorder the shelves and products to optimize sales by putting everything in the most logical order for the visitors. In a supermarket this will typically result in the wine shelf next to a variety of cookies, chips, nuts, etc. Simply because people drink wine and eat nuts with it.

In most web sites there is a same logic that can be applied. Visitors who have questions about a product will go to the product information or support section of a web site. From there they make a logical step to the frequently asked questions page if they have a specific question. A web site owner also wants to analyze visitor behavior. For example, if a web site offers products for sale, the owner wants to convert as many visitors to a completed purchase. If there is a sign-up form with multiple pages, web site owners want to guide visitors to the final sign-up page.

Path analysis answers typical questions like:
Where do most visitors go after they enter my home page?
Is there a strong visitor relation between product A and product B on my web site?.
Questions that can't be answered by page hits and unique visitors statistics.

Funnels and Goals

Google Analytics provides a path function with funnels and goals. A predetermined path of web site pages is specified and every visitor walking the path is a goal. This approach is very helpful when analyzing how many visitors reach a certain destination page, called an end point analysis.

Using Maps

The paths visitors walk in a web site can lead to an endless number of unique paths. As a result, there is no point in analyzing each path, but to look for the strongest paths. These strongest paths are typically shown in a graphical map or in text like: Page A --> Page B --> Page D --> Exit.

Cohort Analysis

Cohort analysis is a subset of behavioral analytics that takes the data from a given dataset (e.g. an eCommerce platform, web application, or online game) and rather than looking at all users as one unit, it breaks them into related groups for analysis. These related groups, or cohorts, usually share common characteristics or experiences within a defined time-span. Cohort analysis allows a company to "see patterns clearly across the life-cycle of a customer (or user), rather than slicing across

all customers blindly without accounting for the natural cycle that a customer undergoes." By seeing these patterns of time, a company can adapt and tailor its service to those specific cohorts. While cohort analysis is sometimes associated with a cohort study, they are different and should not be viewed as one and the same. Cohort analysis has come to describe specifically the analysis of cohorts in regards to big data and business analytics, while a cohort study is a more general umbrella term that describes a type of study in which data is broken down into similar groups.

Examples

The goal of a business analytic tool is to analyze and present actionable information. In order for a company to act on such info it must be relevant to the situation at hand. A database full of thousands or even millions of entries of all user data makes it tough to gain actionable data, as those data span many different categories and time periods. Actionable cohort analysis allows for the ability to drill down to the users of each specific cohort to gain a better understanding of their behaviors, such as if users checked out, and how much did they pay. In cohort analysis "each new group [cohort] provides the opportunity to start with a fresh set of users," allowing the company to look at only the data that is relevant to the current query and act on it.

In e-Commerce, a firm may only be interested in customers who signed up in the last two weeks and who made a purchase, which is an example of a specific cohort. A software developer may only care about the data from users who sign up after a certain upgrade, or who use certain features of the platform.

An example of cohort analysis of gamers on a certain platform: Expert gamers, cohort 1, will care

more about advanced features and lag time compared to new sign-ups, cohort 2. With these two cohorts determined, and the analysis run, the gaming company would be presented with a visual representation of the data specific to the two cohorts. It could then see that a slight lag in load times has been translating into a significant loss of revenue from advanced gamers, while new sign-ups have not even noticed the lag. Had the company simply looked at its overall revenue reports for all customers, it would not have been able to see the differences between these two cohorts. Cohort analysis allows a company to pick up on patterns and trends and make the changes necessary to keep both advanced and new gamers happy.

Deep Actionable Cohort Analytics

"An actionable metric is one that ties specific and repeatable actions to observed results [like user registration, or checkout]. The opposite of actionable metrics are vanity metrics (like web hits or number of downloads) which only serve to document the current state of the product but offer no insight into how we got here or what to do next." Without actionable analytics the information that is being presented may not have any practical application, as the only data points represent vanity metrics that do not translate into any specific outcome. While it is useful for a company to know how many people are on their site, that metric is useless on its own. For it to be actionable it needs to relate a "repeatable action to [an] observed result".

Performing Cohort Analysis

In order to perform a proper cohort analysis, there are four main stages:

- Determine what question you want to answer. The point of the analysis is to come up with actionable information on which to act in order to improve business, product, user experience, turnover, etc. To ensure that happens, it is important that the right question is asked. In the gaming example above, the company was unsure why they were losing revenue as lag time increased, despite the fact that users were still signing up and playing games.
- Define the metrics that will be able to help you answer the question. A proper cohort analysis requires the identification of an event, such as a user checking out, and specific properties, like how much the user paid. The gaming example measured a customer's willingness to buy gaming credits based on how much lag time there was on the site.
- Define the specific cohorts that are relevant. In creating a cohort, one must either analyze all the users and target them or perform attribute contribution in order to find the relevant differences between each of them, ultimately to discover and explain their behavior as a specific cohort. The above example splits users into "basic" and "advanced" users as each group differs in actions, pricing structure sensitivities, and usage levels.
- Perform the cohort analysis. The analysis above was done using data visualization which allowed the gaming company to realize that their revenues were falling because their higher-paying advanced users were not using the system as the lag time increased. Since the advanced users were such a large portion of the company's revenue, the additional basic user signups were not covering the financial losses from losing the advanced users. In order to fix this, the company improved their lag times and began catering more to their advanced users.

References

- Coker, Frank (2014). Pulse: Understanding the Vital Signs of Your Business (1st ed.). Bellevue, WA: Ambient Light Publishing. pp. 30, 39, 42,more. ISBN 978-0-9893086-0-1.
- Finlay, Steven (2014). Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods (1st ed.). Basingstoke: Palgrave Macmillan. p. 237. ISBN 1137379278.
- Siegel, Eric (2013). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (1st ed.). Wiley. ISBN 978-1-1183-5685-2.
- Nigrini, Mark (June 2011). "Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations". Hoboken, NJ: John Wiley & Sons Inc. ISBN 978-0-470-89046-2.
- Lindert, Bryan (October 2014). "Eckerd Rapid Safety Feedback Bringing Business Intelligence to Child Welfare" (PDF). Policy & Practice. Retrieved March 3, 2016.
- "New Strategies Long Overdue on Measuring Child Welfare Risk - The Chronicle of Social Change". The Chronicle of Social Change. Retrieved 2016-04-04.
- "Eckerd Rapid Safety Feedback® Highlighted in National Report of Commission to Eliminate Child Abuse and Neglect Fatalities". Eckerd Kids. Retrieved 2016-04-04.
- "A National Strategy to Eliminate Child Abuse and Neglect Fatalities" (PDF). Commission to Eliminate Child Abuse and Neglect Fatalities. (2016). Retrieved April 4, 2016.
- "Predictive Big Data Analytics: A Study of Parkinson's Disease using Large, Complex, Heterogeneous, Incongruent, Multi-source and Incomplete Observations". PLoS ONE. Retrieved 2016-08-10.
- "2014 Embedded Business Intelligence Market Study Now Available From Dresner Advisory Services". Market Wired. Retrieved August 2015.