PREFACE

WHY IS ANALYTICS THE CAREER FOR THE FUTURE ?

We live in a digital world inundated by endless amount of data. Facebook, Twitter, YouTube, LinkedIn, and other internet based businesses have led to an explosion of data in our world. Retail, telecom, healthcare, airlines, Hospitality and even the sports industry are collecting and analysing massive amounts of data.

Analytics can help companies synthesize data into insights that can increase business revenues and efficiencies. The biggest challenge has been making optimum use of this data.

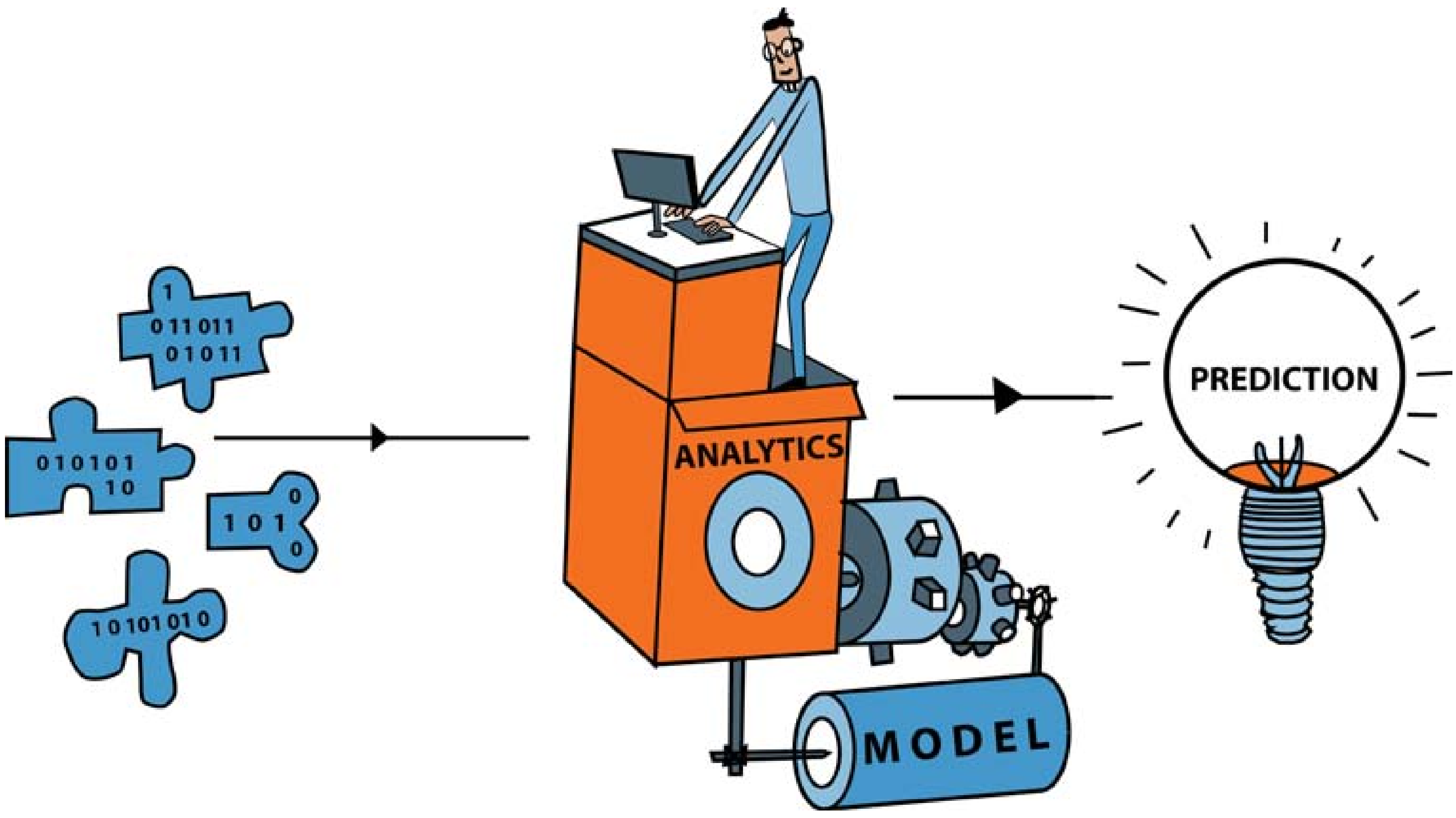
Professionals with the right analytic skills are in high demand.

McKinsey did a report on “Big data analytics” in 2011 and estimated a shortfall of “140,000 to 190,000 deep analytical talent” resources as well as “1.5 million data-savvy managers” in the US alone.

The possibilities for analytics continue to grow and evolve. Innovations in underlying technologies, platforms and analytic tools, drive this use. The need to evolve with their industries and skills in business analysis is a key part of this evolution. Analytics ensure that professionals keep pace with a fast evolving market.

In fact, we, the authors, believe that the “Masters in Business Administration” degree will also evolve to a “Masters in Business Analysis” degree.

So if you are excited about analytics and want to learn more about this field, you can begin your journey with this book. We hope you have a great career!



INTRO TO BEGINNER’S GUIDE

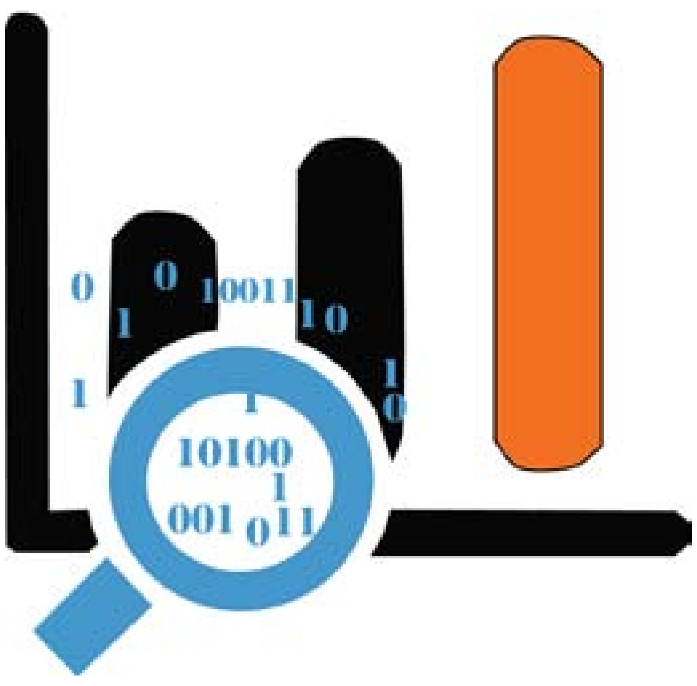
Are you looking for an exciting and rewarding career? Heard about analytics but don’t know if it is the right choice for you? The Beginner’s Guide to Analytics has been designed by industry experts and provides comprehensive information you need to start off in analytics. Over 25000 readers have already benefited from it and you could be next.

Let’s get you started in the exciting world of analytics!

The Beginner's Guide to Analytics is an in depth tutorial on everything you need to know about analytics as a career. You can download your copy of the world's most read guide on analytics right here.

WHAT IS ANALYTICS?

Analytics can be defined as “the analysis of data to draw hidden insights to aid decision making”. Businesses now a days generate huge amounts of data. This data contains many patterns and trends that, if identified, can act as valuable and powerful guides for business decisions and strategy. For example, an e-retailer like Amazon gets hundreds of thousands of buyers on its website every day. By analysing the behaviour of the past customers, Amazon is able to predict what new customers may be interested in.



WHY DO I NEED TO LEARN ANALYTICS?

According to IBM, 90% of the data in this world has been produced in the last 2 years. This is an indicator of the torrent of data that has engulfed all facets of our life. The human mind is not capable of dealing with such large amounts of data. This is especially true in business. Large businesses have thousands of gigabytes of data containing billions of pieces of information. We need specialized tools and techniques to make sense of so much information. Hence, analytics has become a crucial part of managing any business.

Currently, there is a global shortage of trained analytic professionals. Acquiring analytic skills will secure your career for the future.

HOW SHOULD I USE THIS GUIDE?

This guide is meant to serve as a starting point for those who have little or no prior exposure to analytics. If you are seriously interested in this field, we recommend reading this guide from front to back. It's short and easy to understand. There's a printable PDF version for those who'd prefer, and scores of links to resources like blogs, forums, e-groups, websites, videos, white papers and books.

We wish you all the best for your journey to becoming an analytics wiz. We hope you find this guide a useful first step.

CHAPTER 1 WHAT IS ANALYTICS?

Analytics can range from a simple exploration into how many sales of a particular product were made last year to a complex neural network model predicting which customers to target for next year's marketing campaign.

In 'Competing on analytics', Thomas Davenport defines analytics as *"the extensive use of data, statistical and quantitative analysis, exploratory , predictive models, and fact-based management to drive decisions and actions."*

In layman's terms it can be defined as *"the analysis of data to draw hidden insights to aid decision making".*

There is a little bit of analyst in everyone. Analytics is an integral part of most businesses. You do not need to be an "analyst" to do analysis. Analytics is an essential skill for running any kind of business successfully. Common applications of analytics include the study of business data using statistical analysis in order to discover and understand historical patterns with an eye to predicting and improving business performance in the future.

A call centre manager analyses his team's performance data for resource optimization. HR managers use employee data to predict attrition. Marketers use sales and marketing ROI data to decide the optimal allocation of the marketing budget.

In today's workplace, every manager and most contributors are leveraging analytics in one way or another.

DIFFERENT KINDS OF ANALYTICS

One of the ways to classify various kinds of analytics is by the domain it is applied in.

MARKETING ANALYTICS

Marketing analytics is the practice of measuring, analyzing and managing marketing performance to maximize its effectiveness and optimize return on investment (ROI). It allows marketers to be more efficient at their jobs and minimize wasting marketing budgets.

Some of the common types of analyses in marketing analytics include:

Marketing mix optimization – Marketing mix analysis uses techniques such as multivariate regressions to analyse sales and marketing time series data. It determines the effects of various marketing tactics in order to find out the optimal mix of various marketing activities to maximize revenue and/or profitability.

Marketing mix modelling - very common in the CPG/FMCG industry. Companies such as P&G, HLL, Coke and Pepsi swear by this technique to analyse and optimize their marketing spend.

Price analysis – This analysis is done to determine the price elasticity of a product using the historical price and sales data. The results are used to provide insights into expected volume at new prices, key price points, changing price sensitivity and competitive price matching.

Promotional analysis – Promotions data is analysed to understand the sales lift and ROI from various promotional activities such as In-Store Displays, Newspaper & Pre-Print Features, Coupons, In-Store, Mail/Online Offers, Special Packs, Special Events, and Discounts etc.

CUSTOMER ANALYTICS

Customer analytics, essentially a sub-set of marketing analytics, can be classified separately because of its importance to various businesses. It is the analysing of customer behaviour and demographics to develop a better understanding of the consumer leading to better business decisions. Customer analytics is very close to loyalty analytics and is often used interchangeably. Some of the common analyses include:

Customer segmentation –This is the classifying of a large customer population into smaller homogeneous groups. Modelling techniques such as regression modelling, clustering and decision trees are used.

Life time value analysis -This analysis helps quantify the life time value of a customer to the business. This helps identify the most profitable (and hence the most important) customers.
This type of analysis is very popularly in subscription businesses such as telecom and e-tailing.

Attrition analysis -helps identify customers who are likely to attrite in a short time. This knowledge can help devise strategies to for attrition prevention.

> Analytics is defined as the extensive use of data, statistical and quantitative analysis, exploratory and predictive models, and fact-based management to drive decisions and actions.

REGRESSION EXAMPLE

ANALYTICS IN HORSE RACING

Analytics is a high growth field and analysts are constantly finding new ways of applying analytics to business problems. Here is an interesting application of analytics in the leisure domain – specifically Horse racing.

Business Problem: So let us start with the business problem. Our client owns one of the most popular horse racing tracks in the world. They organize over 1000 races every year. A lot of money is bet on these races and one of the big revenue streams for our client is the commission on the bets that are placed. The client receives a fixed percentage of the total money that is bet on each race. This is called the "total wager" in horse racing terminology. The bigger the total wager, the more money our client stands to make.

It is, therefore, natural that the client is interested in understanding what factors affect the total wager amount of each race. For example, the bigger the race, the more interest it generates, and so the bigger the amount that is bet on it. Similarly races that happen on weekends or national holidays generate more wager than those that happen on work days.

Modeling Approach: This problem involves predicting the total amount bet on a race (or the total wager for a race). These kinds of problems can be solved effectively using a regression modelling approach. In this approach, we attempt to find correlations between various factors such as the number of horses running in a race and the dependent variable or the variable that we are trying to predict, in this case the total wager amount.

A Regression model lets us build an equation that captures all these correlations between the variable to be predicted and the predictor variables.

The Regression Equation

Purse amount = 5000 + (.6 * Prize money) + (370 * No. of horses) + (2400 * Holiday indicator)

Where,

Purse = the total prize money to be given away in that race

No. of horses = number of horses participating in the race

Holiday indicator = 1 if the race day is a national holiday; else 0

Model Interpretation

Now, how do we interpret this equation?

Let us take a couple of examples to understand this equation better.

RISK ANALYTICS

Risk analytics refers to analyses that are performed to understand, quantify and manage the risk associated with an activity. Common use of risk analytics includes identifying high-risk customers for a credit card or a loan company. Some of the popular analyses include:

Acquisition modelling – This is usually done on applications data (i.e. data collected at the time of the application) and predicts the likelihood of future default. This helps identify un

Behavioural scoring – help predict the risk and profitability of existing customers using their transaction and credit history. It also helps classify customers based on their risk profile.

Basel II analytics - The International Committee on Banking Supervision (Basel Committee) issued the Basel II Accord to improve the risk management practices of the world's banks. This has resulted in the availability of an extensive amount of data. Analysis of this data can provide more accurate estimations of risk exposures and the capital set aside to guard against the financial and operational risk. Base II analytics includes:

- Computation of Probability of Default (PD)
- Computation of Loss Given Default (LGD)
- Computation of Exposure at Default (EAD)
- Collection Scorecard development

WEB ANALYTICS

Web analytics is the analysis of internet data to understanding and optimizing web usage.It provides information about the number of visitors, hits and page views to a website. It helps gauge traffic and popularity trends which is useful for market research.

The use of Web analytics enables a business to retain and attract customer and increase their profitability. There are two categories of web analytics: off-site and on-site web analytics.

Off-site web analytics - refers to web measurement and analysis regardless of whether you own or maintain a website. It includes the measurement of a website's potential audience (opportunity), share of voice (visibility), and buzz (comments) on the internet.

On-site web analytics - measure a visitor's journey once on your website. This includes its drivers and conversions. For example, it tracks pages that encourage people to make a purchase and measures the commercial performance of the website. This data is typically compared against key performance indicators, and is used to improve a web site or marketing campaign's audience response.

HUMAN RESOURCE ANALYTICS

Most large organizations have some form of a human resource information system that captures employee information such as hiring date, compensation and its growth over time, promotions, roles and performance ratings. Some companies will also capture additional attributes such as skill ratings, previous experience, trainings attended etc.

This kind of information can be the source for important insights. Organizations can predict which of their employees are likely to leave them and use measures to retain the desirable ones and manage the attrition more effectively. Analytics can identify training needs and processes effectively. Data from the vendor management systems can be used by companies to track efficiency and effectiveness of recruitment processes.

Some sports clubs are great examples of successful HR analytics. Players are the most important asset and they invest a lot of money in them. It is understandable that the clubs are anxious to get the right personnel in their teams. Clubs like AC Milan
(football) and the Patriots (American football) have employed quantitative HR techniques to compete successfully in the recent past.

Analytics – Analytics can simply be defined as the process of breaking a problem into simpler parts and using inferences based on data to drive decisions. Analytics is not a tool or a technology; rather it is a way of thinking and acting.At one end, Analytics overlaps with statistics and higher mathematics. At the other, it merges seamlessly with programming and software development.Analytics has widespread applications in spheres as diverse as science, astronomy, genetics, financial services, telecom, retail, marketing, sports, gaming and health care.

> Clubs like AC Milan (football) and the Patriots (American football) have employed quantitative HR techniques to compete very successfully in the recent past.

> The use of Web analytics enables a business to attract more new visitors, retain existing customers and increase the profitability from each customer.

Race 1

Purse = $20000

No. of horses = 10

Holiday indicator = 0 i.e. race day is not a holiday

Race 2

Purse = $7500

No. of horses = 6

Holiday indicator = 1 i.e. race day is a holiday

Predicted purse

*Race 1 = 5000 + (.6 * 20000) + (370 * 10) + (2400 * 0) = 5000 + 12000 + 3700 + 0 = $20700*

*Race 2 = 5000 + (.6 * 7500) + (370 * 6) + (2400 * 1) = 5000 + 4500 + 2220 + 2400 = $14120*

Key takeaways

The most obvious takeaway from this is that the model predicts that about $20k will be bet on the first race and $14k on the second one.

But there are other learnings as well.

For every increase of $100 in the purse, the client can expect an increase of $60 in the total wager

If the number of horses increases by 1, the total wager is expected to increase by $370.

If the race happens on a holiday, one can expect an increase in the total wager by about $2400 compared to other days.

Creating business strategy from analytical insights

How does the client benefit from this analysis? This is of course the most pertinent question. How does a business use these insights for their benefit?

In this case, let us examine the 3 factors that have come out significant in the model.

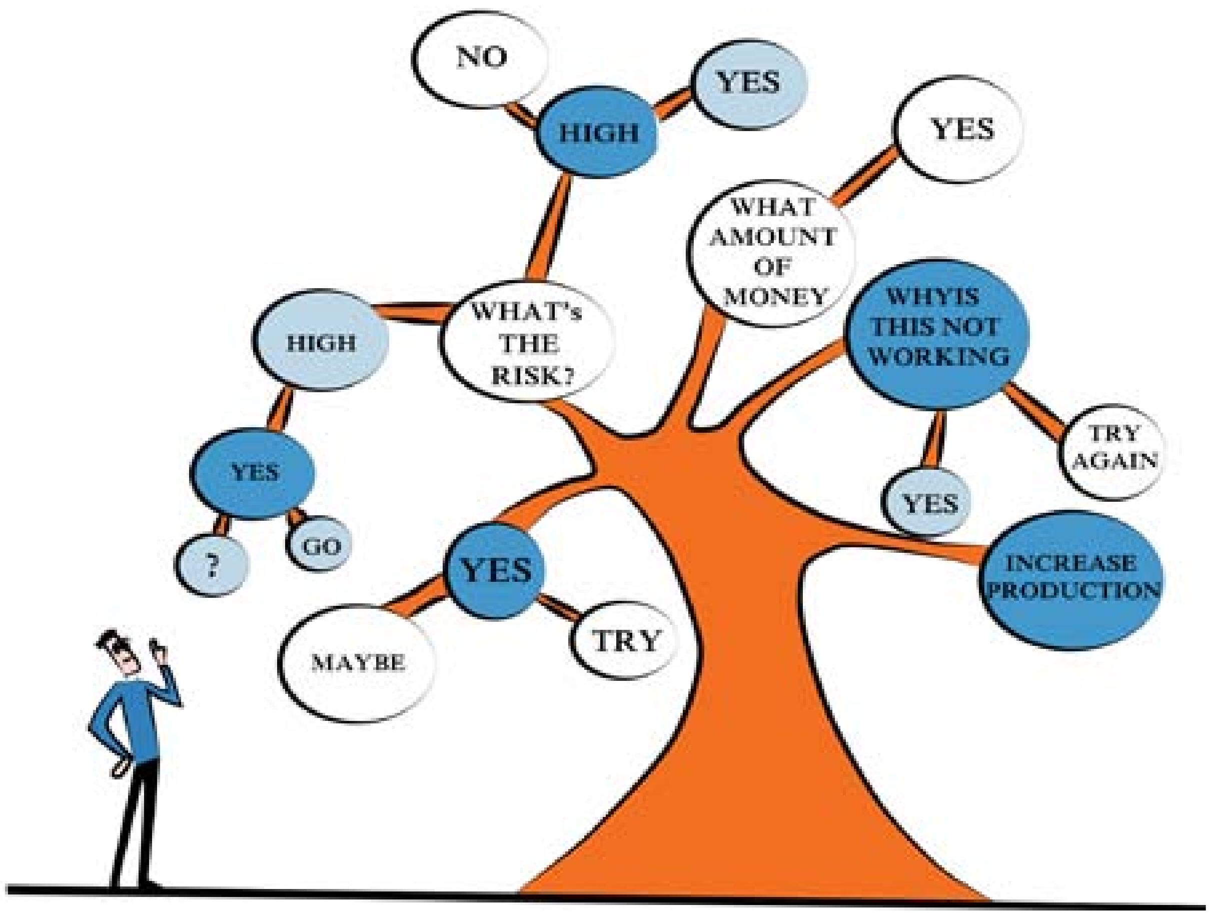
Prize money – We now understand that prize money has a direct correlation with the total wager amount. If you increase the prize money, the total wager amount will also increase. This is perhaps fairly intuitive, especially to those in this business. But our model has not only verified this hunch, but has also helped us quantify the nature of this relationship. For a $100 increase in the prize money, the total wager increases by $60. This information cannot be derived from someone's experience or intuition. Only real number crunching on data of thousands of races can help us get to this. Now, it is up to the client to decide if it makes business sense to increase the prize money of a race or not.

No. of horses in the race – Again, it is fairly obvious to people with experience in this field that the more the number of horses in a race, the more the money bet on it. But our model goes one step further and tells us that we can expect an increase of $370 in the total wager amount if we get an additional horse in the race.

Holiday indicator – The third factor is also not a surprise but more of a validation of intuition. Additionally, we now know that holidays add about $2400 per race to the total wager amount.

Analytics has helped us validate hypotheses about various affecting factors formed through years of experience. Additionally, it has helped us quantify the nature of the relationships. A smart business manager will use this information to make more informed decisions and prioritise more effectively.

In this case, our client deduced that his efforts are best spent on ensuring that he maximises the number of horse running in each race.



FRAUD ANALYTICS

Fraud detection is applicable to many industries- retail, banking, insurance, government agencies, law enforcement, and more. In banking, fraud can involve using stolen credit cards, forging checks, etc. In insurance, a quarter of the claims contain some form of fraud, resulting in approximately 10% of insurance pay-out dollars. Fraud can range from exaggerated losses to deliberately causing an accident for the pay-out. Since relatively few cases show fraud in a large population, finding these can be tricky.

Fraud analytics involves analysing millions of transactions and/or applications to spot patterns and detect fraud.

Logistic regression, neural networks and decision trees are some modelling techniques used in fraud detection.

> While analytics can be applied to many domains in different ways, the underlying analytics methodology as well as the modeling techniques remain the same regardless of the application. For example, logistic regression can be applied in marketing analytics as well as risk analytics. It can also be applied in web analytics or sports analytics.

HEALTHCARE ANALYTICS

Healthcare analytics is a fast growing field that focuses on application of analytics in the health care domain. Some of the common analyses include:

Clinical research analytics – This involves analysis of clinical trials data to see whether a drug has a beneficial effect.

Other analyses- include market forecast, marketing effectiveness and sales resource optimization

These are some of the most common domain-specific applications of analytics. Other applications include R&D analytics, optimization analytics, collections analytics, market research and many more.

For someone who is looking to enter into the field of analytics, it is easy to get overwhelmed by all the different domains and terminologies. However, it's important to remember that while analytics can be applied to many domains, the underlying methodology remains the same regardless of the application. For example, logistic regression can be applied in marketing, risk, sports and web analytics.

Today, we have covered the definition and application of analytics in different domains. We have also learnt that what changes across domains is the data you work with and the business problem you are trying to solve – and not the statistical technique or the basic analytic approach to follow. It is important to have a firm grasp on the different steps of the analytic methodology and underlying statistical techniques.

ANALYTICS TERMINOLOGY

There are many terms that are used in the analytics industry. To someone looking to enter into this field, dealing with such varied terminology is intimidating. Here, we will explain some of the commonly used and misused terms in Analytics.

Business Analytics – This term refers to the application of analytics specifically in business. It includes subsets like –

- Marketing analytics
- Risk analytics
- Fraud analytics
- CRM analytics
- Loyalty analytics
- Operations analytics
- HR analytics

Industries which rely extensively on analytics include –

- Retail
- Telecom
- Health care
- Airlines
- Consumer goods
- Manufacturing
- Sports
- Hotels
- Financial Services (Banks, Credit Cards, Loans, Insurance etc.)
- Any industry where large amounts of data is generated

> Data mining, machine learning, artificial intelligence and knowledge discovery are all terms that are often used interchangeably with analytics.

Predictive Analytics – The term emphasizes the predictive nature of analytics (as opposed to, say the retrospective nature of tools like OLAP). This is a term that is designed by sales people and marketers to add glamour to any business. "Predictive analytics" sounds fancier than just plain "analytics". In practise, predictive analytics is rarely used in isolation from descriptive analytics.
Descriptive analytics – Descriptive analytics refers to a set of techniques used to describe, explore or profile any kind of data. Any kind of reporting usually involves descriptive analytics. Data exploration and preparation are essential ingredients for

predictive modelling and rely heavily on descriptive analytics.

Advanced analytics –Advanced analytics is also a marketing driven terminology. “Advanced” adds a little more punch and glamour to “Analytics” and is preferred by marketers.

Big data analytics – This term has gained popularity very recently. According to IBM, the amount of data being produced in the world is increasing so fast that 90% of the data existing today was created in the last 2 years alone. Increasingly sophisticated tools are required to deal with such vast quantities of data. Hence, the term “big data analytics”

Data Mining – Data mining is the term that is most interchangeably used with “Analytics”. Data Mining is an older term used in the nineties and the early 2000s. However, data mining began to be confused with OLAP and led to a drive to use descriptive terms like “Predictive analytics”.
According to Google trends, “Analytics” overtook “Data mining” in 2005 and is about 5 times more popular now.

Knowledge Discovery – Knowledge discovery also means analytics, predictive analytics, advanced analytics, big data analytics and data mining.

Artificial Intelligence – The term “Artificial intelligence” was popular in the early stages of computing and analytics (in the 70s and 80s) but is now almost obsolete. There were a lot of comparisons between computing and human learning process and this is reflected in the terminology.

Machine Learning – Similar to “Artificial intelligence” this term too has lost its popularity to words like “Analytics” and its derivatives.

Business Intelligence (BI) – The phrase had promise when it stormed to popularity in the late 90s. It started off as a broad phase that encompassed descriptive and predictive analytics. However, it soon got mixed up with OLAP and reporting .Now its usage is largely in that context.

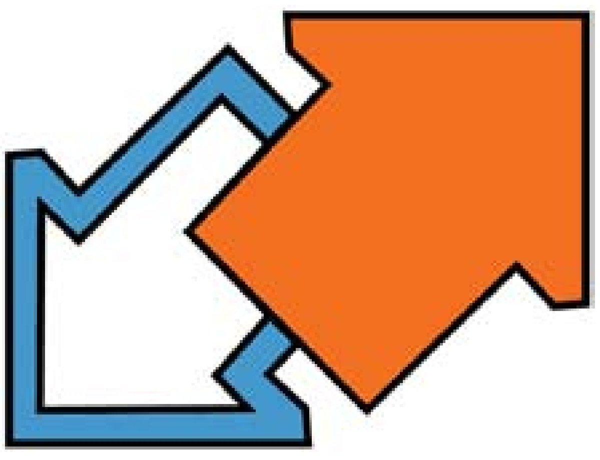
OLAP – Online analytical processing refers to descriptive analytic techniques of slicing and dicing data to understand and discover patterns and insights. The term is derived from another term “OLTP” – online transaction processing which comes from the data warehousing world.

Reporting – “Reporting” is perhaps the most unglamorous term in the world of analytics. Yet it is also one of the most widely used practices within the field. All businesses use reporting to aid decision making. While it is not “Advanced analytics” or even “Predictive analytics”, effective reporting requires a lot of skill and a good understanding of the data and domain.

Data warehousing – Ok, this may actually be considered more unglamorous than even “Reporting”. Data warehousing is the process of managing a database and involves extraction, transformation and loading (ETL) of data. Data warehousing precedes analytics. The data managed in a data warehouse is usually taken out and used for business analytics.

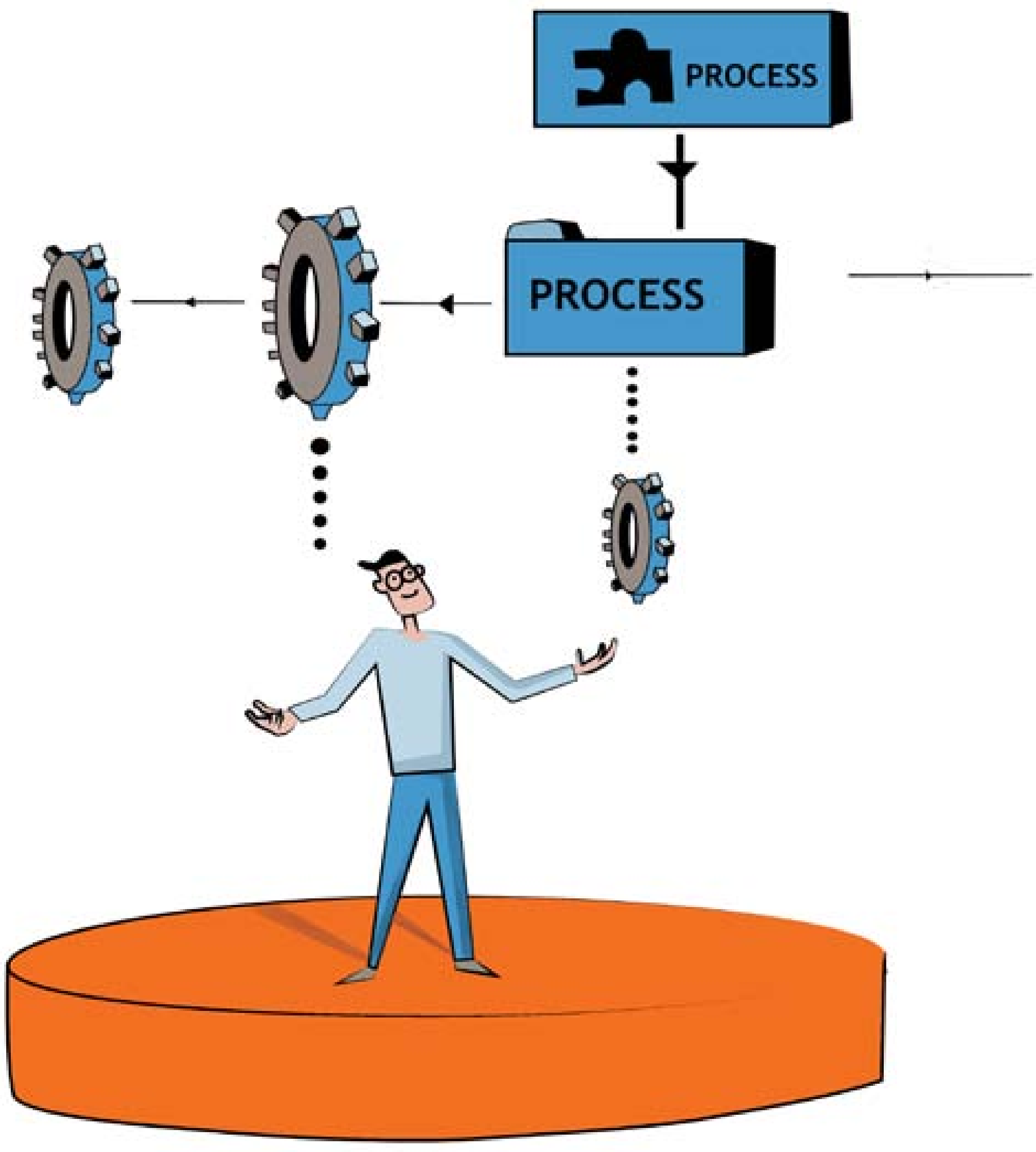
Statistics - Statistics is the study of the collection, organization, and interpretation of data. Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods the result of a major change in the statistics community. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solutions.

> A strong quantitative aptitude as well as a high degree of comfort in dealing with numbers and data are two of the most important attributes required to succeed in analytics.



LINKS TO WIKIPEDIA

- *http://en.wikipedia.org/wiki/Analytics*
- *http://en.wikipedia.org/wiki/Business_analytics*
- *http://en.wikipedia.org/wiki/Predictive_analytics*
- *http://en.wikipedia.org/wiki/Big_data*
- *http://en.wikipedia.org/wiki/Data_mining*
- *http://en.wikipedia.org/wiki/Knowledge_discovery*
- *http://en.wikipedia.org/wiki/Artificial_intelligence*
- *http://en.wikipedia.org/wiki/Machine_learning*
- *http://en.wikipedia.org/wiki/Business_intelligence*
- *http://en.wikipedia.org/wiki/Olap*
- *http://en.wikipedia.org/wiki/Data_warehouse*
- *http://en.wikipedia.org/wiki/Statistics*



CHAPTER 2 IS ANALYTICS RIGHT FOR ME?

Are you unsure about analytics being the right career choice for you? In this article we will discuss what it takes to succeed in analytics.

The purpose of this section is to identify the qualities desired in analyst as well as list the skills and knowledge required to succeed.

Analytics plays a strategic role in running any business and is thus becoming a career of choice for MBAs. Engineers and MCAs have also made a mark in this fast growing field.

Keen sense of intellectual curiosity: People who do well in analytics typically have a keen sense of inquisitiveness. They want to know the whys and hows of any situation. There has to be an interest in understanding the business issue and working out the specifics of the solution.

Mathematically oriented: To do well in analytics, you need to be comfortable with mathematical concepts, and not be afraid to use mathematical tools. This is not the career for you if the word Mathematics strikes fear in your heart!

Big picture vision: It is important to always remember the larger business issue that is being addressed through the process of working with data and dealing with minutiae.

Detail oriented: While it is important to remember the big picture, it is critical to pay attention to the details. While working with large volumes of data it is easy to lose sight of the specifics that add insight and understanding to solving business issues.

Interpretation skills: An analyst has to have the ability to interpret and solve problems. Numbers by themselves mean nothing. Experience and domain understanding give one the ability to interpret the results in the business context. The results aid in building strategies for business development.

There are other skills that one can acquire through education, training and experience.

- Knowledge of Statistics
- Knowledge of analytic tools like SAS, R, Excel, SQL, SPSS, Knowledgestudio etc.
- Knowledge of the analytics methodology – the different steps to be followed for any analytics project
- Predictive modelling techniques such as regression, decision trees, clustering, market basket analysis etc.
- Experience with large data sets
- Project work such as segmentation, attrition modelling etc.

WHAT DOES IT TAKE TO SUCCEED IN ANALYTICS?

Academic Qualifications for Analytics

Analytics is attracting talent from many different fields. Previously, candidates with a masters or doctoral degree in Statistics or Economics were preferred.Today graduates, postgraduates and PhDs from all quantitative such as math, science and commerce are part of the industry:

- BBAs and MBAs
- MCA- their expertise in programming languages and database management systems
- B.E.s and B.Techs- A combination of engineering and MBA is an asset in this field.
- Qualified people from Pharmaceutical and Health sciences.
- Professors

ANALYTICS CAREER IN INDIA

There are certain factors to be kept in mind :

- Indian Analytics companies have international clients giving the opportunity to work and travel globally.
- Working non-standard hours is typical but night shifts are not common. Shift timings are usually either 9am to 5pm or 12pm to 8pm.
- Since the basic statistical techniques remain constant irrespective of the domain, analysts can move from one domain to the other with relative ease. It provides an opportunity to work across multiple industries.
- Growth in analytics is largely merit driven. A good analyst on a fast track could easily move up to middle management in less than 5 years. Similarly, you could be stuck at the same level for more than 2-3 years, if you do not pick up the skills required to move up to the next level.
- Analytics softwares have evolved a lot in the past 2 decades. Tools have become sophisticated, faster and user-friendly, with intuitive, easy to learn GUIs.
- The focus of a good analyst is to translate a business problem into an analytic problem and an analytic solution into business strategy.

CHAPTER 3 HOW DO I LEARN MORE ON ANALYTICS?

Analytics was initially applied on data stored as tables. The scope expanded to included text from blogs, forums and websites. Digital media are also used for insights such as shopper behaviour inside a store.

Analytic tools are becoming faster, more sophisticated and user-friendly at the same time. There have been massive strides in the open source category with R becoming the most popular analytic tool in the world.

How does one keep pace with so many changes and advancements in the field? How does one even start building knowledge on analytics?

In this section, we will cover three sources of information on analytics –

- Books
- Blogs
- Other internet resources

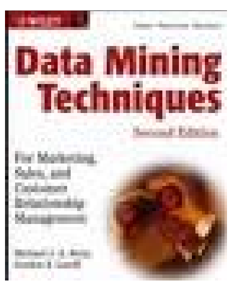
BOOKS ON ANALYTICS

There are a lot of good books written on analytics. Not all of them are easily available in the Indian market. Many of these titles can be ordered from online book stores like Flipkart. The books which have an Indian publisher are usually cheaper whereas foreign publications are more expensive.

We have divided our list of favourite books into two sections.

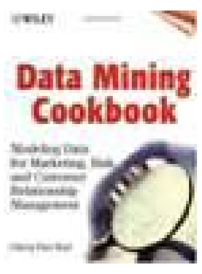
The first section covers books on the fundamentals of analytics and business statistics. We have also included books showcasing applications of analytics in various industries.

The second section consists of books around specific analytic tools or software. These books cover tools like SAS, SPSS, R, SQL and excel.



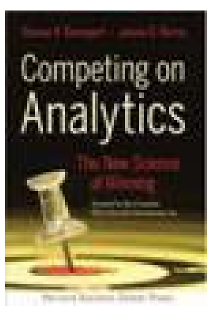
Data Mining Techniques by Michael Berry and Gordon Linoff

This is an excellent book on the most widely used analytic techniques. It starts off with defining data mining in the current business context and then summarizes some of the best practices in data mining. The book talks about some useful statistical concepts like p-value and chi-square as it takes the readers through the process of building a model. It explains analytic algorithms like Decision trees, market basket analysis, clustering, link analysis, clustering and survival analysis. The book is full of useful industry examples. This is the first book recommended for anyone with an interest in analytics



Data Mining Cookbook by Olivia Parr Rud

This book provides a detailed undestanding of the analytics methodology. It lists out several best analytic practices. The author primarily uses logistic regression as her technique and SAS and excels as the tools. The book has a very brief description of analytics so make sure you have some understanding of analytics before you get to this book.

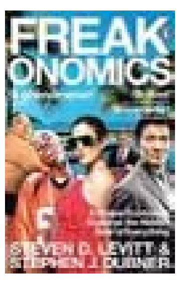


Competing on Analytics by Thomas Davenport

This is a much-needed addition to the literature of analytics. The book describes how some leading international companies are using analytics to out-smart their competition. The book is full of case studies as diverse as retail to pharmaceuticals to sports to telecom. This book is a valuable resource in understanding how companies like Amazon, Netflix and Capital One have developed analytics as their key differentiator.

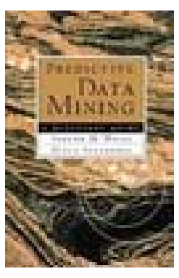
Statistics for Management by Richard Levin and David Rubin

This is a great guide for statistics used in analytics. Starting from simple central tendency measures to probability distributions to decision theory, this book covers the essentials of business statistics. It is used as course book in most management institutes in India. This book is recommended for anyone interested in Statistics.



Freakonomics by Steven Levitt

Freakonomics is an easy, interesting read, even for people who do not understand Statistics. The author uses the power of analytics to turn conventional wisdom on its head. With hypotheses like 'Legalization of abortions has led to crime reduction' the book throws up some very interesting questions for its readers. A must-read for everyone.



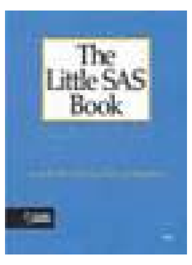
Predictive Data Mining by Sholom Weiss

This book has an easy style that will appeal to beginners. Some of the more complex topics may not be adequately addressed but a good book as an introduction.

Moneyball by Michael Lewis

This book studies the use of analytics in the sports industry. It case studies the Oakland Athletics, a US baseball team, with a payroll budget of less than a 3rd of some their rivals. Despite being small they have consistently been one of the best teams. The author shows how they leveraged analytics to get their advantage. OA analyzed metrics that were different from the ones traditionally looked at, but which they thought were more relevant to winning. A good example of analytics being used innovatively.

BOOKS ON ANALYTICS TOOLS

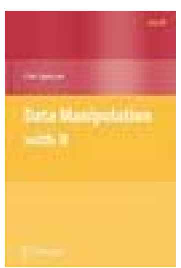


Little SAS Book by Lora Delwiche

This is a good book to learn SAS. It is readable as it is composed of two-page articles. Each one focuses on a specific task or function of SAS. The book is divided in 10 chapters that go through reading data sets, building reports, combining data sets, writing macros, using graphics, debugging SAS programs, etc. The book is a good reference for simple tasks. Any simple task you don't know? Just look in the index and you will find the corresponding function. However, for more advanced topics, the book is a bit light. Well, that's what the title also says.

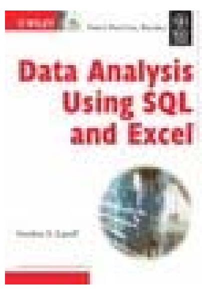
SAS Programming by Example by Ron Cody

It is one of the best books available for beginners in SAS. The book is simple and easy to read with many industry examples for better understanding. The book deals with the essentials of analytics and reporting using SAS.



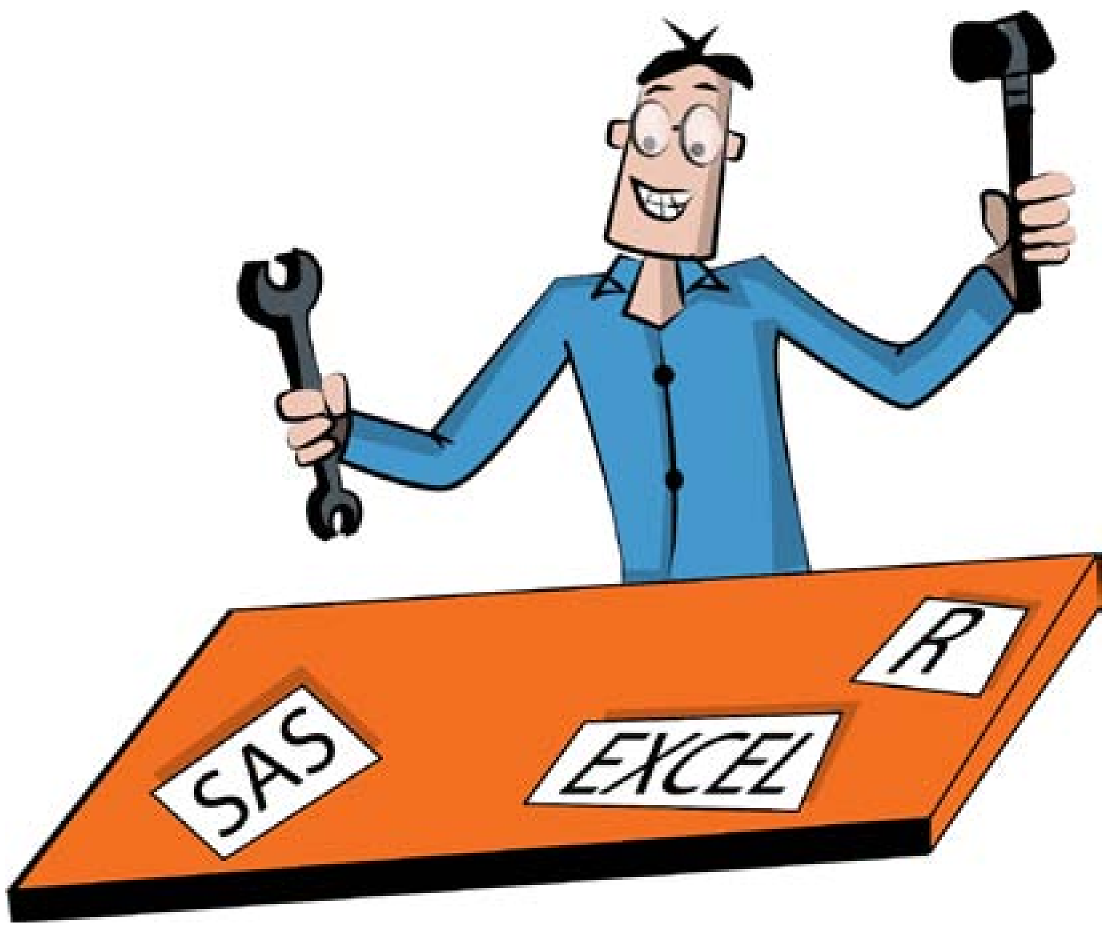
Data Manipulation in R by Phil Spector

This thin book provides a solid introduction to many of the functions and packages for importing, manipulating and processing data in R. Using a variety of examples based on data sets included with R, along with easily simulated data sets, the book is recommended to anyone who wants to learn R.



Data Analysis using SQL and Excel by Gordon Linoff

A good book on understanding how tools like SQL and Excel can be leveraged to extract useful business information from relational databases. The book is organized around chapters that become increasingly complex in the use of SQL, Excel, and data mining concepts.



BLOGS ON ANALYTICS

Blogs are a useful learning resource for any topic these days. Analytics has its fair share of bloggers. Here is a list of some useful blogs on analytics.

Decisionstats www.decisionstats.com - Ajay Ohri, an alumnus of IIM Lucknow, has one of the most popular blogs in the field of analytics. A salient feature is an Interviews section where leading analytics figures have been interviewed. He has been publishing articles almost at the rate of 1 a day for the last three and half years and has close to 1500 articles published on his blog.

Data Mining Blog - http://www.dataminingblog.com/ - This blog by Sandro Saitta, a Swiss national, covers research issues, recent applications, important events, interviews with leading actors, current trends and book reviews in the field of analytics. The blog has some interesting book reviews but what we like the most about this blog is the comprehensive list of analytics blogs.

This blog is written by Michael Berry and Gordon Linoff, two of the leading figures in the field of analytics. The two have written some of the most popular books in this field. Their books are used extensively in data mining courses by Universities across the globe. The blog is filled with excellent tips and knowledge nuggets from the writers' vast experience in analytics consulting and is a must-read for any business analyst.

Flowing Data - http://flowingdata.com/ - FlowingData is the visualization and statistics site that shows different applications of data analysis. The blog is written by Nathan Yau, a PhD candidate in UCLA. Nathan wants to make data useful for even those who are not very data savvy through effective use of visual techniques.

Abbott Analytics - http://abbottanalytics.blogspot.com/ - The Abbot analytics blog is about "Tips, tricks, and comments in data mining and predictive analytics, including data preprocessing, visualization, modeling, and model deployment"

The blog is hosted by Dean Abbott, president of Abbott Analytics in California, USA. Mr. Abbott has over 21 years of experience applying advanced data mining, data preparation, and data visualization methods in real-world data intensive problems, including fraud detection, response modeling and survey analysis.

SAS RESOURCES

Though the SAS language is almost 45 years old now, it remains one of the most widely used statistical programming languages, competing with RSPSS, Matlab and Stata. It is facing competition with the open source language R in academia, research and customized applications

Here is a list of the top learning resources for learning SAS language to work in this exciting field:

Books by SAS Press – The SAS Institute have a separate publishing division that focuses on the analytics ecosystem by generating and publishing a steady stream of books and literature on the SAS language. The SAS Bookstore is a great place to look for books at the beginner, intermediate and expert level at

SAS Global user's conference contains an archive of white papers presented at the annual conference for SAS users. SUGI is a useful resource for any SAS user.

The SAS-L Email Group- the SAS-L email group is a supportive and helpful resource for programmers at any level when stuck in a particular program. The archives are available at http://listserv.uga.edu/archives/sas-l.html . People posting on the list are expected to paste a sample of the dataset structure, and a clear explanation of what they are trying to achieve.

The SAS Community Website- SAS Users have an online wiki for collaboratively adding in code samples, programs and sharing tips. The site is available at http://www.sascommunity.org/wiki/Main_Page and it is quite useful in terms of daily tips to improve SAS language skills, blogs and papers.

Blogs at SAS.com – This is a collection of blogs that deal with business analytics, and technical application in various domains and SAS related happenings. With almost 29 blogs, http://blogs.sas.com is one of the better locations to build your perspective in business analytics and not just the technical aspects of writing code. SAS blogs are exclusively maintained by the SAS Institute team of communication people and feature experts from across multiple businesses.

SAS Online Document- SAS online document is the documentation available for everyone by web access. It is like a big book or HTML library and you can view the latest version at http://support.sas.com/documentation/onlinedoc/91pdf/index.html. It is very useful for fine-tuning your SAS language code and discovering extended functionality to your software as well as troubleshooting or debugging programs.

Papers from SAS Global Users Conference Proceedings (SUGI) – This is an archive of papers presented at the annual conference for SAS Users. The conferences have been taking place since 1976, and the papers represent the best innovative uses in the language. They can be accessed online at http://support.sas.com/events/sas-globalforum/previous/online.html

Thearling's blog - http://www.thearling.com/ - Kurt Thearling's **website** is an excellent place to start off if you want to learn about data mining. Kurt is a veteran in the analytics space with extensive experience in some of the leading analytic companies like Capital One, Dun & Bradstreet and Vertex. Kurt is one of the thought leaders in this space and has authored several books on analytics.

Though the site does not get updated frequently, it is a treasure trove of interesting information. Kurt has written a series of articles/papers that cover topics ranging from the basics of data mining to advanced domain-specific techniques. His articles are written in a clear and concise manner and are easy to understand for beginners as well.

There are many other useful blogs on analytics. If you do a google search for "analytics blogs" or "data mining blogs", you can find a lot of information. This link has a fairly comprehensive though dated list of data mining blogs: - http://www.kdnuggets.com/websites/blogs.html

www.thearling.com is an excellent resource for those starting off in analytics. www.dataminingblog.com is a great blog that covers news and developments in analytics.

RESOURCES ON R

R is the most popular open source analytic tool in the world. There are several resources for the R language.

There are numerous books on R. The good news is that like R, many of these books are also available for free.*

BOOKS ON R

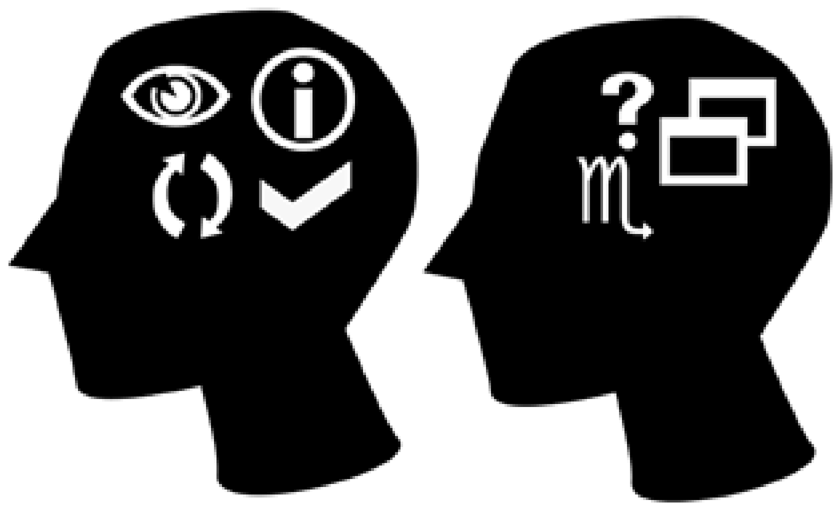
- An Introduction to R PDF HTML - W. N. Venables, D. M. Smith and the R Development Core Team
- R for Beginners PDF - Emmanuel Paradis
- Using R for Data Analysis and Graphics: Introduction, Code and Commentary PDF - J. H. Maindonald
- SimpleR: Using R for Introductory Statistics PDF - John Verzani
- The R Guide PDF - W. J. Owen
- Introduction to Probability and Statistics Using R URL - G. Jay Kerns
- The R Inferno PDF - Patrick Burns
- An Introduction to R PDF - Longhow Lam PDF

Analyticbridge is a community for analysts. This is a wonderful site to network within the analytics community.

Some of the books you have to pay for:

- Statistics: An Introduction using R - Michael J. Crawley
- The R Book - Michael J. Crawley
- A Beginner's Guide to R - Alain F. Zuur, Elena N. Leno, Erik H.W.G. Meesters
- A First Course in Statistical Programming with R - W. John Braun and Duncan J. Murdoch
- A Handbook of Statistical Analyses Using R - Brian S. Everitt and Ibrsten Hothorn
- A Modern Approach to Regression with R - Simon J. Sheather
- Data Manipulation with R - Phil Spector
- Introductory Statistics with R - Peter Dalgaard
- R for SAS and SPSS Users - Robert A. Muenchen
- R Graphics - Paul Murrell
- R in a Nutshell: A Desktop Quick Reference - Joseph Adler
- There is a lot more documentation on R available at http://cran.r-project.org/other-docs.html

Courtesy stackoverflow.com



OTHER ONLINE RESOURCES

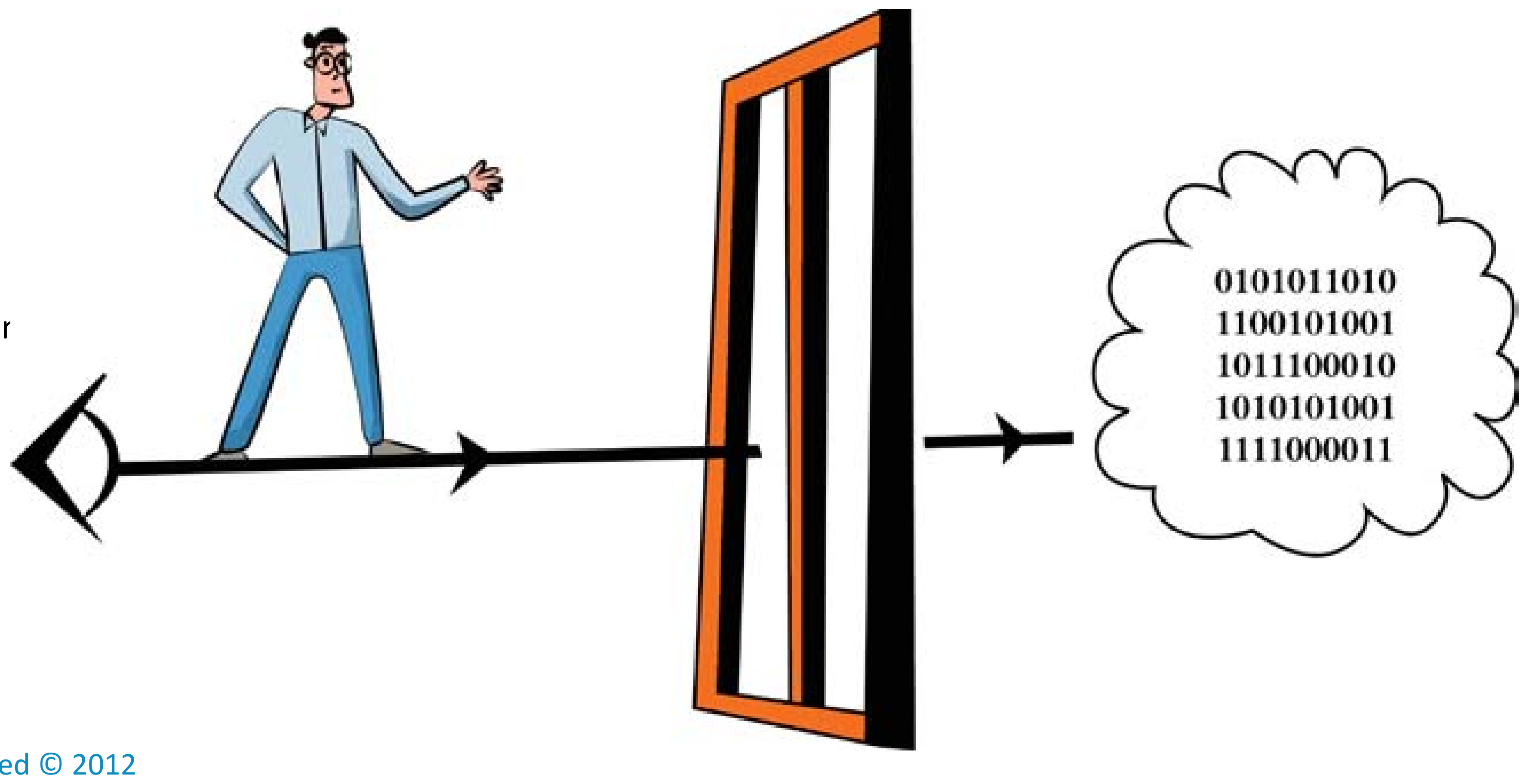
Kdnuggets – www.kdnuggets.com – Established in 1997, this website is a comprehensive resource for anything related to data mining and analytics. Visit this site to learn about news, events and jobs in analytics. They also have links to datasets and training material on analytics.

Analyticbridge – www.analyticbridge.com – Describes itself as the social network for analytics professionals. This is a wonderful site to connect and network within the analytics industry. It is also useful to stay updated on the latest news and events in the field.

LINKEDIN GROUPS

LinkedIn has thousands of groups that are like virtual communities where people of similar interests can share knowledge. Some of the popular analytics related groups are :

- Advanced analytics
- Business analytics
- Customer analytics group
- Global analytics network
- India analytics network
- Kdnuggets analytics and data mir
- SAS analytics and BI
- SAS and analytics users
- Your analytics career



YOUTUBE

These are some must-watch videos on Analytics

How it works: Analytics: http://youtu.be/_HbjsNaUJ2A

A brief history of intelligence: http://youtu.be/yVIcIRcAhxc

What can Business Analytics Do for You?: http://youtu.be/uP89kaDU40c

REvolutionAnalytics: This channel has many interesting videos on big data analytics using the open-source software R

SASsoftware: SAS's channel talks primarily about SAS but also has some good videos on analytics.

Googleanalytics: The official channel for all videos about and related to Google Analytics.

IBMbusinessanalytics: IBM's youtube channel focusing on analytics and its business application.

Jigsawacademy: Our own channel provides informative videos on analytics, interviews with industry leaders as well as snippets of our lectures.

CHAPTER 4 INDUSTRIES THAT USE ANALYTICS

Analytics can be applied to virtually any business or business process. Whether you are dealing in a niche product, commodity, monopoly or a competitive market, there is potential for analytical competition.

However, some industries are clearly more amenable to analytics than others. If a business generates a lot of transaction data, competing on analytics is a natural strategy.

Financial Services

Retail

FMCG

Pharmaceuticals

Travel

Hospitality and Entertainment

Logistics

Business to Business

Telecom

Sports

Value-add Services

INTERNATIONAL COMPANIES THAT USE ANALYTICS

INDIAN COMPANIES THAT USE ANALYTICS

IT Companies:

- Dell
- HP
- IBM
- Infosys
- Wipro
- Igate
- Tata Consultancy Services
- Cognizant
- HCL

Retail:

- Target
- Tesco
- Shopperstop
- Arvind Mills
- Pantaloons

Consulting Companies:

- Dun & Bradstreet
- McKinsey
- Accenture

Healthcare:

- Novartis
- Reddylabs
- IMS Health

Manufacturing:

- Caterpillar

Telecom:

- Nokia Networks
- Vodafone Telecommunications

E-commerce:

- Amazon
- Rediff
- Bharatmatrimony
- Times of India
- Ebay

FMCG:

- HLL
- P&G
- Pepsi
- Coke

Analytics Service Providers:

- Genpact
- Evalueserve
- Ugam Solutions
- WNS
- EXL Service
- Absolutdata
- CrossTab Decision-Craft
- Fractal Analytics
- Meritus
- IMRB
- Modelytics
- Denuosource
- Mu-Sigma
- Latent View Analytics
- Dexterity
- Pharmarc
- Manthan Systems
- Irevena
- Global Analytics Inc.
- Dunnhumbly

Financial Services:

- Amex
- Citibank
- HSBC
- Standard Chartered Bank
- Fair Isaac India
- ICICI Bank
- HDFC Bank
- Bajaj Allianz Insurance
- Bharati Axa Insurance
- Fidelity

CHAPTER 5 TOOLS FOR ANALYTICS

The range of analytical software goes from relatively simple statistical tools in spread sheets (ex-MS Excel) to statistical software packages (ex-KXEN, Statistica) to sophisticated business intelligence suites (ex-SAS, Oracle, SAP, IBM among the big players). Open source tools like R and Weka are also gaining popularity. Companies are also developing in-house tools designed for specific purposes.

Here is a list of 10 most popular analytic tools used in the business world.

SAS is currently amongst the most popular analytic tools in the business. IBM products such as SPSS and modeler are also fairly popular. R, the open source tool, is the tool for the future. WPS is a low-cost alternative for SAS users.

COMMERCIAL SOFTWARES

MS Excel: Almost every business has access to MS Office suite and Excel. Excel is an excellent reporting and dash boarding tool. For most business projects, even if you run the heavy statistical analysis on different software, you will still end up using Excel for the reporting and presentation.

Latest versions of Excel can handle tables with up to 1 million rows making it a powerful yet versatile tool.

SAS: SAS is the 5000 pound gorilla of the analytics world and claims to be the largest independent vendor in the business intelligence market. It is the most commonly used software in the Indian analytics market despite its monopolistic pricing. SAS software has wide ranging capabilities from data management to advanced analytics.

SPSS Modeler (Clementine): SPSS Modeler is a data mining software tool by SPSS Inc., an IBM company. It was originally named SPSS Clementine. This tool has an intuitive GUI and its point-and-click modelling capabilities are very comprehensive.

Statistica: is a statistics and analytics software package developed by StatSoft. It provides data analysis, data management, data mining, and data visualization procedures. Statistica supports a variety of analytic techniques and is capable of meeting most needs of the business. The GUI is not the most user-friendly and it may take a little more time to learn than some tools but it is a competitively priced product that is value for money.

Salford systems: provides a host of predictive analytics and data mining tools for businesses. The company specialises in classification and regression tree algorithms. Its MARS algorithm was originally developed by world-renowned Stanford statistician and physicist, Jerome Friedman. The software is easy to use and learn.

KXEN: is one of the few companies that are driving automated analytics. Their products, largely based on algorithms developed by the Russian mathematician Vladimir Vapnik, are easy to use, fast and can work with large amounts of data. Some users may not like the fact that KXEN works like a ‘black box’ and in most cases, it is difficult to understand and explain the results.

Angoss: Like Salford systems, Angoss has developed its products around classification and regression decision tree algorithms. The tools are easy to learn and use, and the results easy to understand and explain. The GUI is very user friendly and a lot of features have been added over the years to make this a powerful tool.

MATLAB: is statistical computing software developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms and creation of user interfaces. There are many add-on toolboxes that extend MATLAB to specific areas of functionality, such as statistics, finance, image processing, bioinformatics, etc. Matlab is not a free software. However, there are clones like Octave and Scilab which are free and have similar functionality.

Open Source Software

R: R is a programming language and software environment for statistical computing and graphics. The R language is an open source tool and is widely used by the academia. For business users, the programming language does represent a hurdle. However, there are many GUIs available that can sit on R and enhance its userfriendliness.

Weka: Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software, developed at the University of Waikato, New Zealand. Weka, along with R, is amongst the most popular open source software used by the

WHICH IS THE BEST ANALYTICS SOFTWARE TO LEARN?

The next question that arises, especially for someone aspiring to get into the field of analytics, is "Which is the best analytics software to learn?"The analytics software market has numerous players – from billion dollar corporations to 1-man shops- offering highly sophisticated, even self-learning platforms to niche, custom-made solutions.

SAS

SAS is undoubtedly the reigning king in the market. Most large global companies and cash-rich businesses tend to go for this software. In India, most medium and large analytics service providers use SAS for predictive modelling and advanced data mining.

There are many different software available from the SAS institute. Some of them are fairly generic and domain agnostic while some are domain-specific for niche purpose. Base SAS is the cheapest and most widely used. It is a code-based tool that uses the SAS language for coding. While some of the tasks like data import can be performed using the GUI, most tasks require a knowledge of the SAS language.

Being a code-based tool, Base SAS is not very easy to learn. However, other than Excel, it is one of the most popular analytics software in the business market.

SAS has another tool called "E-miner". E-miner is a GUI based version of Base SAS where you can perform complex data manipulation and modelling tasks just by clicks and drags. E-miner has additional advanced analytics capabilities that are not present in the base SAS version. Capabilities to perform market basket analysis, decision trees, neural networks and support vector machines make e-miner a very comprehensive and user-friendly tool.

The biggest drawback of this software is the high price that the SAS institute insists on charging for it.

Deployment of this tool in a mid-size organization can run into millions of dollars. In India, apart from the large MNC banks like Citibank, Barclays etc. few companies are able to afford this tool.

IBM

With some major acquisitions in the last few years, IBM has suddenly become a lead player in the analytics software market. While COGNOS is largely considered as a business intelligence tool, SPSS and IBM modeler (previously SPSS Clementine) are also big players. SPSS is more popular in the market research field than the business analytics field. IBM modeler is comparable to SAS e-miner both in terms of features as well as pricing and hence has the same pros and cons.

SOFTWARES OF THE FUTURE

R

R is the most popular open-source (FREE) analytics software in the world. Within the analytics community, its popularity easily surpasses that of any other tool. And it is easily the tool of choice for most academicians and scientists. Businesses, however, have been slow to adopt R and a lot of this can be because of intellectual property issues arising out of codes and algorithms written on an open source platform.

While R is again a code based package similar to Base SAS, there are GUIs available for it. The Revolution analytics GUI for enterprise R is an awaited new addition to the analytics software market. The GUI cuts down on the learning time for R and even though it is not free, it still makes for a very cost-effective solution.

Numerous companies have expressed a desire to move to R. They are, however, unsure about training their existing resources and building a pipeline of trained resources for the future.

As a new generation of analysts starts to move towards building R skills, companies will surely move to this platform.

With its ever-increasing capability list, and low/no-cost pricing, R is the reigning analytics software for the future.

WPS

WPS is a tool that has been around for some time but has not been able to gain the popularity it deserves. This tool, now acquired by IBM, is virtually a clone of Base SAS. It uses the same SAS language, has a similar interface and identical algorithms. In fact, most of your SAS codes will run on the WPS platform and most SAS users can transition to this tool with a simple, 2-day training.

The software is completely legitimate, having won numerous cases filed by the SAS institute.

WPS is very attractively priced compared to Base SAS saving over 50% in costs for companies if they move from SAS to WPS.

Once IBM gets around to marketing this tool effectively, it seriously dent SAS's stranglehold on the market. With a pool of SAS trained resources available to them, companies will find moving from SAS to WPS a walk in the park.

With such a short learning and adoption time, WPS is the second most exciting tool for the future.

CHAPTER 6 MODELLING TECHNIQUES

REGRESSION MODELLING

Regression models are widely used in analytics, as they are the easiest to understand and interpret. Regression techniques allow the identification and estimation of possible relationships between a pattern or variable, and factors that influence that pattern. For example, a company may be interested in understanding the effectiveness of its marketing strategies. It may deploy a variety of marketing activities in a given time period, like TV, radio and print advertising, and social media campaigns. A regression model can be used to understand and quantify which of its marketing activities actually drive sales, and to what extent. The advantage of regression over simple correlations is that it allows you to control for the impact of multiple factors that influence your variable of interest, or the "target" variable. For example, things like price changes or competitive activities also influence a brand's sales, and the regression model accounts for the impacts of these factors on sales.

Types of Regression Analysis

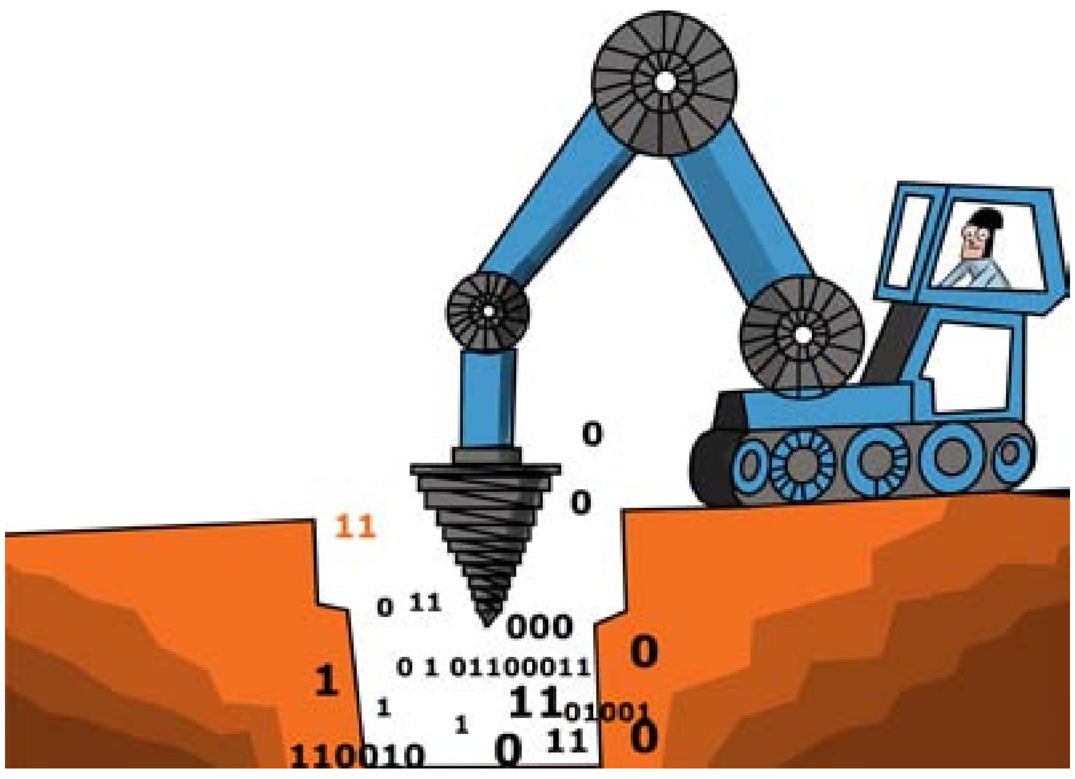
There are several different types of regression techniques, including :

Linear regressions- assumes that there is a linear relationship between the predictors (or the factors) and the target variable.

Non-linear regression- allows modelling of non-linear relationships,

Logistic regressions - useful when your target variable is binomial (1,0 – Accept or Reject)

Time Series Regressions- used to forecast future behaviour of variables based on historical time ordered data.

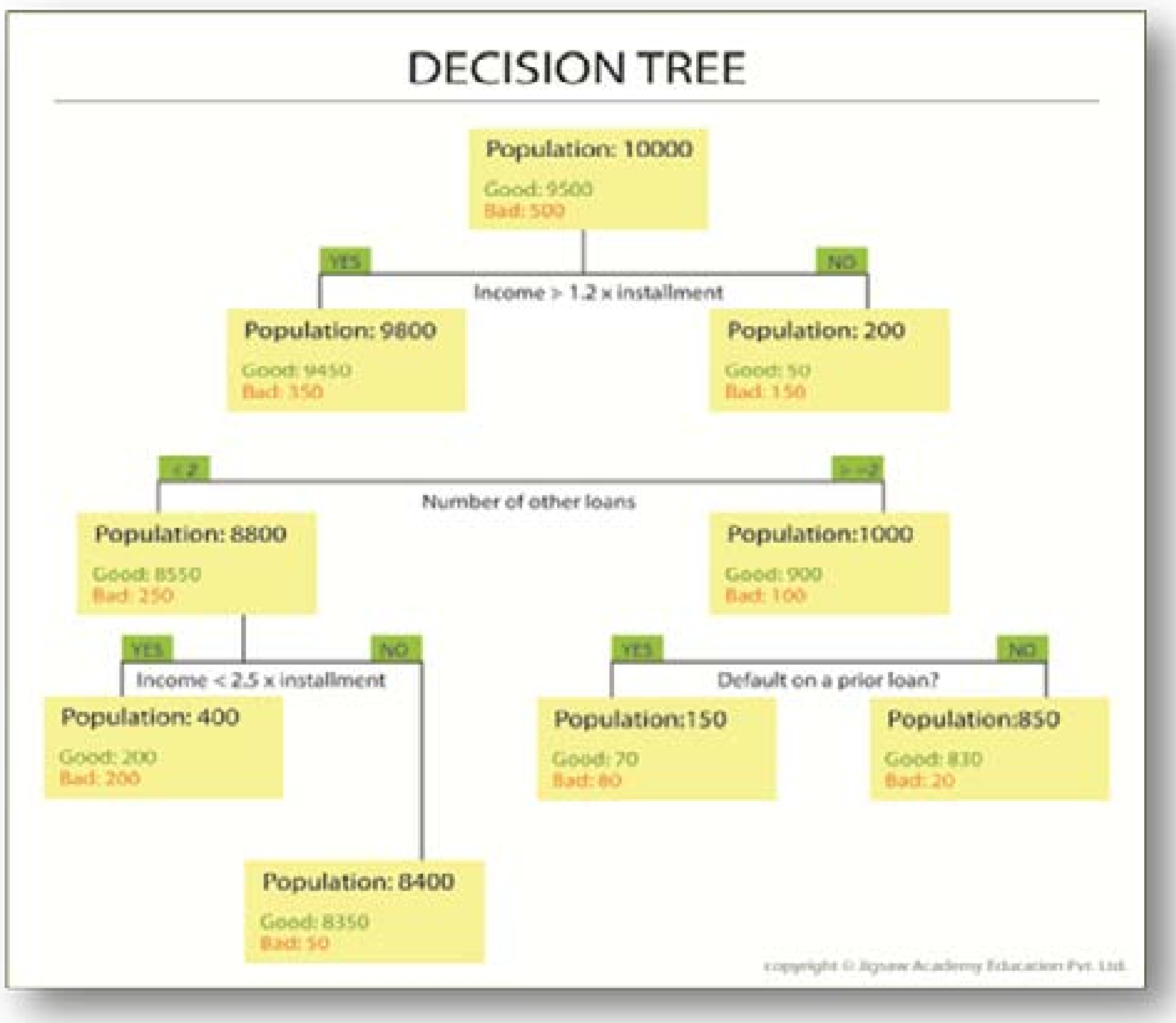


BUSINESS APPLICATIONS

Regression techniques are widely used in various business needs. These models are built to understand historical data and relationships to assess marketing effectiveness, price changes on sales, ranking people on propensity, responds of a direct mailing campaign, to flag potentially fraudulent applications, to assess cross-sell and up-sell opportunities across an existing customer base, to predict attrition or churn, and many more.

DECISION TREES

A decision tree is a predictive model that can be viewed as a tree. The predictions are made on the basis of a series of decision similar to the game of 20 questions. For instance, if a loan company wants to create a set of rules to identify potential defaulters, the resulting decision tree may look something like this.



DECISION TREE ALGORITHMS

The most crucial step in creating a decision tree is to identify the right split (question to ask) at each stage. Most algorithms look at all possible split options and calculate the most effective option .The split will result in 2 or more child nodes from the parent node. The same process is repeated at each child node. All possible options are evaluated to identify the best split as per a pre-set criterion.

Although there are many decision tree algorithms, they basically work on 2 principles. One kind works on increasing the purity of the resulting nodes while the other works on ensuring maximum statistically difference from the parent node. The 3 most commonly used methods are Gini, Chi-square and Information gain. The F test is also used when the target variable is continuous like probability or income. The F test is also called the reduction in variance test.

APPLYING DECISION TREES TO BUSINESS

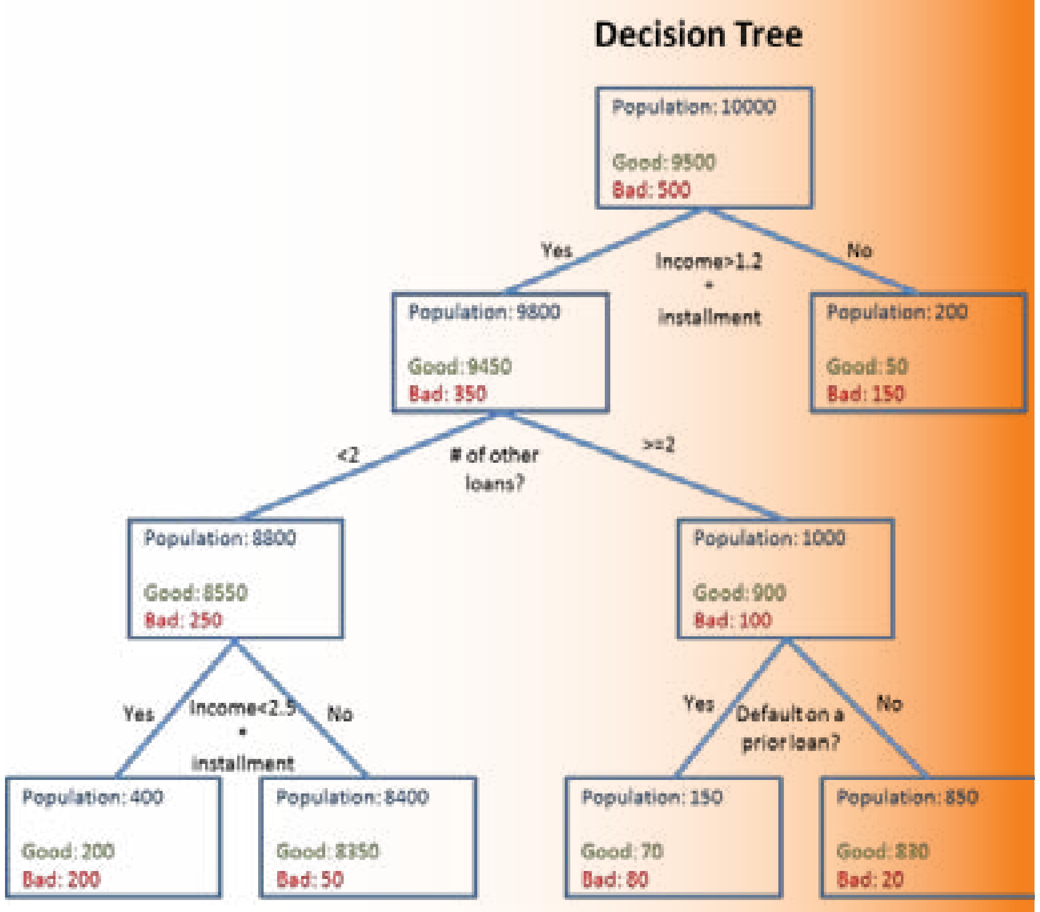
Because of their graphical structure, decision trees are easy to understand and explain making it very popular. Because of this clarity they allow for

DECISION TREE EXAMPLE

PREDICTING LOAN DEFAULTERS USING DECISION TREES

IDIDI Bank wants to improve the profitability of its consumer loan department. The bank has seen a high default rate on its loans between 2 to 5 years of tenure and wants to tighten its credit policy. The bank wants to understand what the profile of customers who default is, and how does it differ from the non-defaulters. Finally, the bank wants to identify characteristics that are peculiar to defaulters and hopes to come up with a set of rules that can help weed out defaulters at an early stage.

Here is an example of such analysis using Decision trees.



The decision tree is built on a population of 10000 customers. 500 of these customers have defaulted i.e. a default rate of .5%. The purpose of creating a decision tree is to identify data based characteristics that separate defaulters from non-defaulters.

For example, in the first step, the tree divides the population into 2 groups. One group has a default rate of .4% while the other has a default rate of 75%. It is a simple criteria for separation. The first group has customers whose monthly income is at least 1.2 times the monthly instalment. The second group consists of customers with a monthly income <1.2 times the monthly instalment. Yet the second group has a default rate that is more than 200 times that of the first group.

It makes clear business sense to create a rule that ensures that no new customer goes into the 2nd group. In other words, a rule that says that a customer's income has to be at least 1.2 times the instalment, will help reduce the default rate for the bank.

Similar inferences and rules can help the bank tighten its credit policy to ensure the default rate goes down over a period time.

more complex profit and ROI models to be easily added on top of the predictive model.

Their high level of automation and the ease of translating these models into SQL the technology has also Their high level of automation and the ease of translating these models into SQL the technology has also proven to be easy to integrate with existing IT processes, requiring little pre-processing and cleansing of the data.

Apart from predictive modelling, decision trees are useful in other ways too. During the data exploration stage of a project, decision trees are a quick and effective way of understanding impact of numerous and important variables .A real life business situation could involve thousands of variables and decision trees are a quick and easy way of reducing to a manageable number.

Decision trees are also very useful in understanding variable interactions. Sometimes two or more variables could combine to become powerful. For example, the target market for Yamaha motorcycles will be males between a certain age group and income band. Just the income, gender or age may not be as powerful in segmenting this market.

Trees can be used to impute missing values. Sometimes trees can help interpret model results that could otherwise be difficult to understand. For example a neural net may generate output that can't be interpreted easily. Applying decision trees using the prediction of neural net as the target variable can generate rules that provide broad understanding of the neural net model.

Features of a good decision tree

A good decision tree follows Occam's razor. It delivers a high level of accuracy using as few variables as possible. A good tree is easy to visualize and interpret. Like all models, it should make intuitive sense.

When to use decision trees

Decision trees are the technique of choice when the problem is binary (0/1, yes/no). It is not the first choice for estimating continuous values. It is widely used in business, especially in situations, where the interpretation of results is more important than the accuracy. Use decision trees when time is critical. Decision trees work the best for initial data exploration in any project.

RANDOM TREE

Random forest is an improvement over the traditional decision tree approach. The technique involves building multiple trees and then chooses the class that is output by most number of trees – thus taking a mode of all the trees.

Random forest is extremely useful for prediction and classification and produces a high accuracy rate. Like decision trees, it requires minimal data preparation and is unaffected by outliers.

CLUSTERING

Clustering is the process of grouping similar observations into smaller groups within a larger population. It has widespread application in business analytics. One of the questions facing businesses is how to organize the huge amounts of data into meaningful structures. Cluster analysis is an exploratory analysis tool which aims at sorting different objects into groups. It analyses the degree of association between two objects; maximal if they belong to the same group and minimal otherwise.

Features of clustering

Clustering is an undirected data mining technique. This means it can be used to identify hidden patterns and structures in the data without formulating a specific hypothesis. There is no target variable in clustering. In the case above, the grocery retailer was not actively trying to identify fresh food lovers at the start of the analysis. It was attempting to understand the different buying behaviors of its customer base.

Clustering is performed to identify similarities with respect to specific behaviors or dimensions. In our example, the objective was to identify customer segments with similar buying behavior.

Cluster analysis can be used to discover structures in data without providing an explanation or interpretation. In other words, cluster analysis simply discovers patterns in data without explaining why they exist. The resulting clusters are meaningless by themselves. They need to be profiled extensively to build their identity i.e. to understand what they represent and how they are different from the parent population.

In the retailer's case, each cluster was profiled on its buying behavior. Customers in cluster 1 spent a quarter of their total spend on fresh, organic produce. This was significantly higher than other customers who spent less than 5% on this category. This segment of customers was called 'Fresh food lovers' as this is what distinguished them from the rest of the customers.

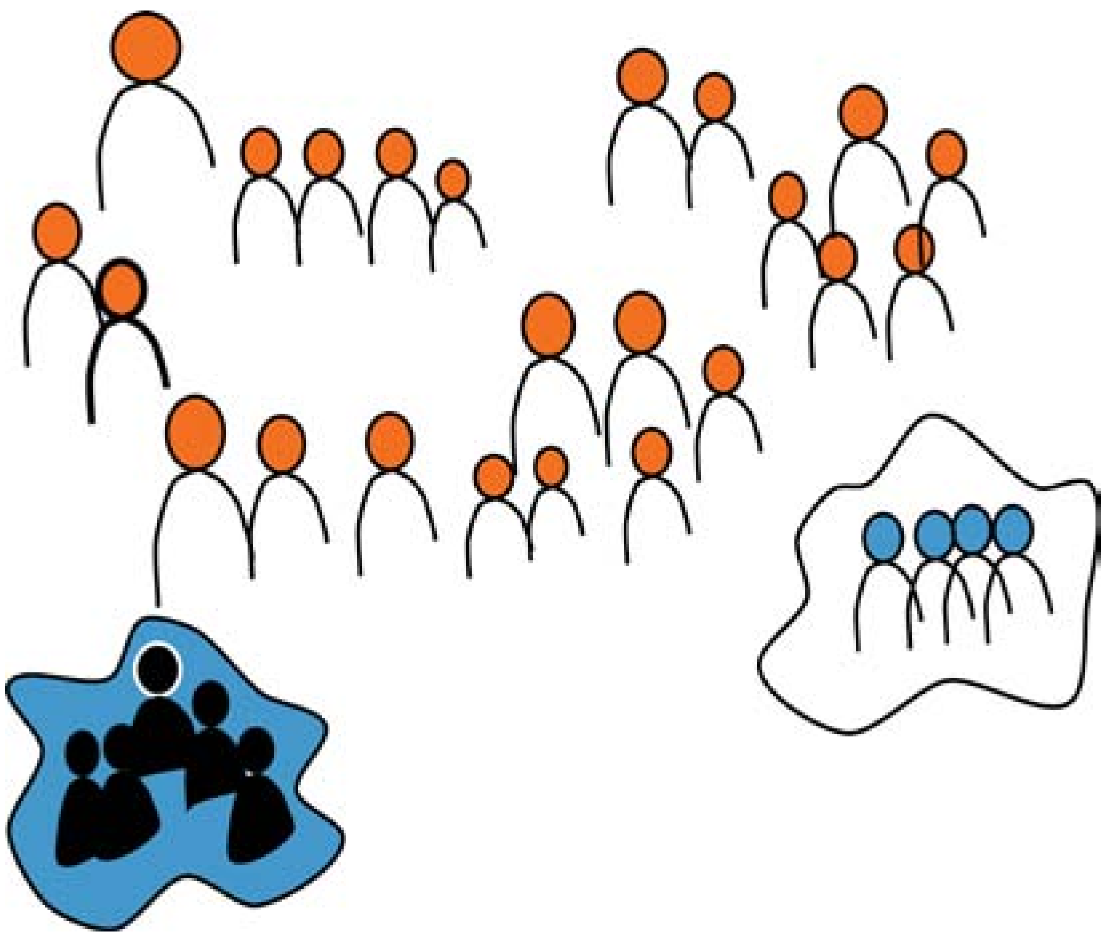
TYPES OF CLUSTERING

There are different algorithms available for clustering, and each of them may give a different set of clusters. The choice of a particular method will depend on the objective of clustering, the type of output desired, the hardware, software facilities available and the size of the dataset.

Clustering techniques may be divided into two categories based on the cluster structure which they produce:

We started the analysis with historical data on 10000 customers, 500 of whom were defaulters, i.e. a default rate of 5%. We asked a series of questions and on the basis of the responses we were able to break the population into smaller groups. In the end we identified 3 smaller groups of population with default rates of 50% or above.

In this data, there are 200 customers whose income is not greater than 1.2 times their loan instalment. 150 of these customers defaulted i.e. 75% of the population. Based on this, it is clear that if a customer's income is not greater than 1.2 times the instalment amount, they are likely to default on the loan. The bank can reject such loan applications in the future to reduce its default rate.



CLUSTERING EXAMPLE

A grocery retailer uses clustering to segment its 1.3MM loyalty card customers into 5 different groups based on their buying behavior. It then adopted customized marketing strategies for each of these segments in order to target them more effectively.

One of the groups was called 'Fresh food lovers'. These are customers who purchase a high proportion of organic food, fresh vegetables, salads etc. A marketing campaign that emphasized the freshness of the fruits and vegetables and year-round availability of organic produce in the stores appealed to this customer group.

Another cluster was called 'Convenience junkies'. They are people who shopped for cooked/semi-cooked, easy-to prepare meals. A marketing campaign focusing on the retailer's in-house line of frozen meals as well as the speed of the check-out counters at the store worked well with this audience.

In this way the retailer was able to deliver the right message to the right customer and maximize the effectiveness of its marketing.

The non-hierarchical methods -divide a dataset of N objects into M clusters. K-means, a non-hierarchical technique, is the most commonly used one in business analytics.

The hierarchical methods- produce a set of nested clusters in which each pair of objects or clusters is progressively nested in a larger cluster until only one cluster remains.

WHEN TO USE CLUSTERING

Clustering is primarily used to perform segmentation, be it customer, product or store. We have already talked about customer segmentation using cluster analysis in the example above. Similarly products can be clustered together into hierarchical groups based on their attributes like use, size, brand, flavor etc.; stores with similar characteristics – similar sales, size, customer base etc., can be clustered together.

Clustering can also be used for anomaly detection, for example, identifying fraud transactions. Cluster detection methods can be used on a sample containing only good transactions to determine the shape and size of the “normal” cluster. When a transaction comes along that falls outside the cluster for any reason, it is suspect. This approach has been used in medicine to detect the presence of abnormal cells in tissue samples and in telecommunications to detect calling patterns indicative of fraud.

Clustering is often used to break large set of data into smaller groups that are more amenable to other techniques. For example, logistic regression results can be improved by performing it separately on smaller clusters that behave differently and may follow slightly different distributions.

ASSOCIATION RULES

Affinity grouping or rule induction is one of the most popular data mining techniques. This unsupervised learning technique works by mining through large databases to identify patterns. These patterns are hidden deep inside the data, creating a set of rules and assigning a measure of strength and likelihood of occurrence for the rules.

One of the most interesting applications of this technique is on the point of sale (POS) data to identify what products sell together. Such groups of products can be useful to make recommendations to customers Customers who have bought one item from a group are more likely to buy the rest of the items within that group.

Web retailers use this tool to make recommendations based on what other people have bought in the past. Amazon is a great example of this. Now more websites are offering this facility.

Brick and mortar retailers can make recommendations in different ways – placing associated products close by in the store, bundling them in promotions etc.

While association rules are very insightful, they have to be used with care. In a typical retail business there are hundreds of products which mean endless combinations are possible. An analyst can quickly get overwhelmed with the number of rules being generated and may not be able to determine the useful ones. A good yardstick to measure the strength or the power of a rule is necessary.

A related problem with association rules is that they provide a lot of redundant information. Most of the generated rules tell us what we already know – People who buy apples are likely to buy bananas too. Or any other fruit for that matter. But we already know this and so fruits are kept together and often promoted together.

But sometimes the analysis can provide deeper information. Wal-Mart may already know about the association between candy and doll sales (both are liked by little children) but the analysis can identify which doll links the most.

Rules can also be used to associate time of day or the day of week with product sales. A convenience store finds that lottery sales peak during lunch hours while newspapers don’t sell much after 11 am. It decides to replace the newspaper rack with a second lottery rack every day before noon. Similarly weekend trends can be identified through association rules as well.

An important thing to keep in mind while using association rules is that rules don’t imply causality. Buying milk doesn’t lead to buying bread even if the rule If milk, then bread is very strong. It just means there is a strong association between milk and bread.

The rules generated by decision trees cover all data points (or every scenario) and there are no overlaps between rules. In other words they are exhaustive as well as mutually exclusive. Induction rules tend not to be exhaustive (at least after the pruning) and are not mutually exclusive. For example, if bread then milk and if bread then eggs are 2 different association rules for the same situation.

NEURAL NETWORKS

Neural networks are another popular data mining technique that can be applied to perform predictive modelling and unsupervised learning in the form of clustering. They have found application in fraud detection, credit scoring and store clustering to name a few.

Are they the best?

Neural networks, or rather artificial neural networks, mimic the way the human brain works – detecting patterns, making predictions and learning from experience. It is tempting to conclude that since neural networks

CROSS TABULATION EXAMPLE

PAY HIGHER INSURANCE IF YOU DRIVE A RED CAR

Car Insurance companies have used risk-based pricing for a long time. Basically, the idea is to charge a higher insurance for cars that are more likely to get into an accident and a lower insurance for cars that are less likely to get into an accident. In the US, for instance, if you are under 25, single and drive a sports car, you are considered a higher risk customer than someone who is, say, 30, married and drives an SUV. And if by chance, your sports car is red in color, then you are even more likely to get into an accident. In this case, your insurance premium is going to be sky-high.

So how do companies decide who is risky and who is not? How do insurance companies infer that a red sports car is more likely to get into an accident than a silver sports car?

This is where analytics comes in. Insurance companies have historical data on hundreds of thousands of cars. They have data on which car filed for an insurance claim (which implies which car got into an accident). They also have other information pertaining to the cars, for example their vintage, location, prior history etc.

CAR INSURANCE IN INDIA

In India, the insurance is for the car and not the driver. This means that insurance companies do not consider factors like the age of the driver, her previous record etc. when it comes to insurance pricing. Only car related factors like the vintage, location etc. are taken into account.

Let us assume a simplistic scenario where we have two factors – the car model and the location. And we want to do build a differential pricing strategy based on these two factors.

Location: Is location a factor that affects risk in this case?For example, is a car more likely to get into an accident in a city like Delhi versus a city like Indore? Now Delhi has more than 100 times the vehicles in Indore. It is therefore possible that accident rates in Delhi are higher than Indore. And thus, a car bought in Delhi could be more likely to get into an accident than a car bought in Indore.

Model: Similarly, the car model. Is a Hyndai Santro less likely to get into an accident than a Tata Indica? Maybe. Without data, we really can’t tell.

Insurance Data

We have historical data on 41,096 policies. The policies cover 12 different car models and sixteen locations. We first examine the data by car model. We rank order the car models based on their risk index.

MODEL	RISK INDEX	RISK CLASSIFICATION
Hyundai Santro	65 %	low
Mitsubishi Cedia	45 %	low
Ford Figo	40 %	low
Maruti Swift	38 %	mid
Honda City	28 %	mid
Tata Indigo	24 %	mid
Honda Civic	24 %	mid
Maruti Wagon-R	12 %	mid
Tata Indica	10 %	high
Honda Accord	-1 %	high
Tata Safari	-13 %	high
Mahindra Scorpio	-14 %	high

Our data tells us that Scorpio and Safari have the highest risk associated with them.

CITY	RISK INDEX	RISK CLASSIFICATION
Bhopal	65 %	low
Kolhapur	58%	low
Mathura	55%	low
Varanasi	52%	low
Indore	47 %	low
Jalandhar	45 %	low
Calicut	29 %	mid
Ahmedabad	25 %	mid
Noida	19 %	mid
Ludhiana	18 %	mid
Calcutta	13%	mid
Guwahati	9 %	high
Gurgaon	8%	high
Bangalore	8%	high
New Delhi	7%	high
Chennai	6%	high

Similarly, Chennai and Delhi are the cities with the highest chance of getting into an accident.

We can now take this one step further and create a matrix of city by model. We can simplify this matrix by classifying all models and locations into a 3*3 grid.

RISK SCORE	BHOPAL KOLHAPUR MATHURA VARANASI INDORE JALANDHAR	CALICUT AHMEDABAD NOIDA LUDHIANA KOLKATTA	GUWAHATI GURGAON BANGALORE NEWDELHI CHENNAI
Hyundai Santro Mitsubishi Cedia Ford Figo	80% low	72% low	64% low
Maruti Swift Honda City Tata Indigo Honda Civic Maruti Wagon-R	68% low	60% mid	52% mid
Tata Indica Honda Accord Tata Safari Mahindra Scorpio	55% mid	33% high	27% high

The grid tells us that based on historical data a Tata Indica in Delhi will be a lot riskier than a Hyundai Santro in Bhopal. Thus, a risk based pricing will lead to a higher insurance premium for someone who buys an Indica in Delhi than someone who buys a Santro in Bhopal.

This overview provides a description of some of the most common modelling techniques in use today.

Please note that this data is not representative of the Indian market. The data is for illustration purpose only.

learn from experience, they will outperform other techniques by improving over time. However, this is not true. Neural networks learn and improve like other techniques do and they do this in a clumsy one-record-at-a-time manner unlike the other techniques.

When to use neural networks?

The results of neural networks are hard to interpret and explain. In fact, this technique works like a black-box. You can get the right answer but you may not understand how exactly you got it. In light of this, it is best to use neural networks when you have a high degree of familiarity with the input variables as well as the predicted variable and you have a clear idea of what you want to model. Apart from prediction, they can be used for clustering, outlier detection and variable selection as well.

Data preparation is often an involved and complex process. The fields need to be chosen carefully and the data needs to be standardized. The predictor variables need to be numeric and the output is also numeric.

Neural networks do not produce easily understood rules that explain how insight was arrived at. Yet, it is possible to understand the relative importance of inputs into the network by using sensitivity analysis. In many business situations, understanding the comparative importance of inputs is almost as good as having clear rules.

There are many analytic tools such as IBM modeler, Statistica and e-miner that have neural network capabilities. R has many packages for this purpose as well.

WHICH ANALYTIC TECHNIQUE TO CHOOSE?

This is often the hardest question at the start of the analysis. There are numerous different analytic techniques that are used to solve different kinds of problems. While different techniques work well with specific kinds of problems, it is often hard to pick one technique as the best solution in real-world cases. Analysts often try multiple approaches and then use the one that makes the most amount of sense. The choice of technique is often governed by availability as well. Most software packages will offer a sub-set of all the techniques and the analyst has to pick the best option from the available ones.

Some tips on modeling

- It is better to pick a simple, stable, easily explainable model than a complex, less stable and more accurate model.
- It is better to do a quick-and-dirty analysis and implement actions quickly than wait to complete a long and refined analysis.
- It is better to spend more time understanding and exploring the data than building numerous sophisticated models.

ASSOCIATION RULES EXAMPLE

People who buy milk and bananas are more likely to buy cornflakes.

If milk and bananas, then cornflakes

People who buy one diet product are interested in other diet products.

If diet coke, then low fat mayo

People pick up chips when they buy beer.

If chips, then beer

Neural networks are like a “black box”. You put in the inputs, tweak some parameters, and you get the end results. It is very hard to explain how the algorithm came to the final results.

The Analyst

The entry level role in the field of analytics is that of an analyst. This role offers you a chance to work on projects in a team, building expertise in the domain as well as the analytical tools and techniques. Basic knowledge of statistics is mandatory. You need to be familiar with widely used statistical concepts like p-value, probability distributions, chi-square etc. In addition, knowledge of common analytical techniques like logistic regression, clustering and decision trees is highly desirable.

The Senior Analyst

You will spend anywhere from 18 to 36 months in your first role before moving to the next level i.e. a senior analyst. If you have a PhD, you could directly join at this level. In this role, you will be expected to work independently on projects, even leading some of the smaller ones. You will continue to learn new methodologies as well as build domain knowledge.

The Team Lead

After spending another 2 to 4 years in this role, you will move to a team lead or lead consultant level. At this stage, many people will move into a people management track where they get to demonstrate their leadership and people management skills. However, some people with a more technical bent may choose to become subject matter experts (SMEs) instead. As an SME, you will provide leadership in defining and executing analytic methodologies for specific projects as well as develop new techniques to improve current processes. At this level, you will also be expected to manage most of the communication with clients regarding the projects you are working on.

The Manager

The next role in your career will be that of a manager. As an analytics manager you will be responsible for a team of 5 to 30 people. You may have 1 or more team leads working under you. If you are on the SME track, you may not have a permanent team assigned to you but you would be expected to provide thought leadership on individual projects.

The Associate Vice President

The next role is that of a senior manager followed by the Associate vice president (AVP).

Growth in analytics is largely merit-based. You would need to demonstrate strong quantitative skills and an analytical aptitude in the early years. However, as you move up, communication and people management skills become differentiating factors.

CHAPTER 8

SALARIES IN ANALYTICS

Analytics is a high growth field and salaries in analytics have been sky rocketing in the last 5 to 8 years .India ranked 3rd in the study – in terms of number of people available with analytical skills – behind the US and China. Yet India has emerged as the global hub for analytics. This is because India has a steady supply of English-speaking analytical talent.

In 2005, a company would hire a graduate or a post-graduate at an entry level position of analyst for about Rs 1.8 lacs per annum. In 2011, campus salaries for analytic roles have been reported at around Rs. 6 lacs going as high as Rs. 12 lakhs at premium colleges.

A senior business analyst with 3 to 5 years of experience would get anywhere between Rs. 8 to 12 lacs per annum. A Team leader or a lead analyst with 5 to 10 years of experience would be making around 10 to 14 lakhs. The salary growth in this field is fairly steep. due to the high demand for experienced professionals .

By the time a person gets to the manager's level, they usually have close to 10 years of experience and would be commanding a salary in the range of Rs. 12 to 20 lacs.

A senior business analyst with 3 to 5 years of experience would get anywhere between Rs. 8 to 12 lacs per annum. A Team leader or a lead analyst with 5 to 10 years of experience would be making around 10 to 14 lakhs.

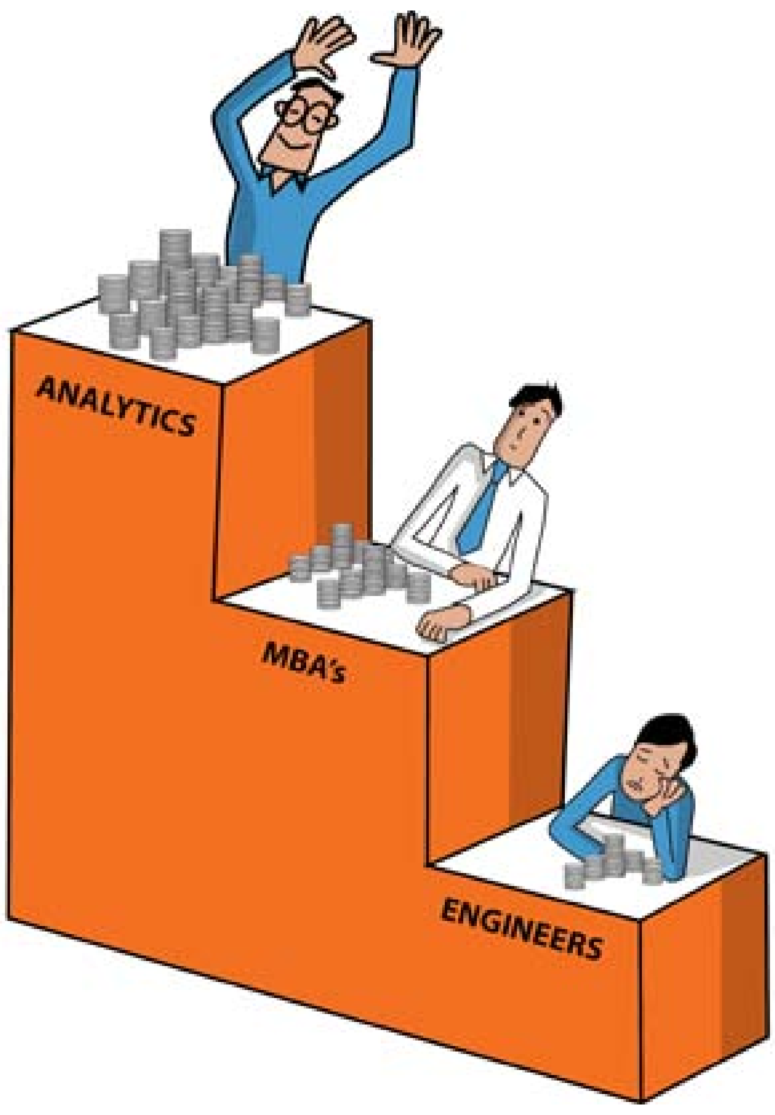
Campus recruitment in Analytics

Analytics companies recruit students in large numbers from various colleges. Initially the trend was to recruit from colleges like ISIs (Indian Statistical Institutes). However, now companies are increasingly going after MBA graduates. Colleges like Goa Institute of Management have become a hot bed for analytics recruitment.

Companies offer anywhere from Rs. 3 to 8 lacs, for campus placements in most MBA colleges.

Captive vs. Non-captive centres

Captive centres are those that are dedicated to a particular company. For example, those working for HSBC analytics or Dell analytics will work exclusively on the HSBC or Dell business. On the other hand, an analyst working in an analytics service company such as Genpact or Marketelligent is likely to work across multiple businesses and clients. The latter work tends to be looked upon as more rewarding since it is more varied. To make up for this, captive centres usually offer slightly higher salaries than non-captives.



Salaries in Analytics in Bangalore vs. Gurgaon vs. Chennai vs. Kolkata

Salaries across major cities like Bangalore, Gurgaon, Chennai and Kolkata are fairly consistent and differ largely due to cost of living differences. Kolkata has a lot of analytic talent and it is not an expensive city to live in. Hence salaries are a little lower compared to other cities.

How will salaries in Analytics grow in the next 5 years?

The dependence of business on data savvy professionals is expected to sky rocket in the coming years. There is a shift in demand from pure statistical talent to professionals who can apply analytics in the context of a business problem and thus a shift from statisticians and mathematicians to data savvy professionals

The starting salaries in analytics are expected to grow at 10 to 20% per annum at least for the next 5 years.

Professionals in the analytics industry can expect annual pay hikes of about 15% every year.

Promotions in analytics tend to be largely merit-driven.

Average time to promotion is 18 to 24 months at the lower levels and 2 to 5 years as you move up the hierarchy.

CONTRIBUTING AUTHORS

GAURAV VOHRA

Gaurav is an MBA from IIM Bangalore and has over a decade of experience in the field of analytics. He has worked across multiple verticals including financial services, retail, FMCG, telecom, pharmaceuticals and leisure industries. Gaurav likes to work with different analytic tools including Knowledgestudio, SAS, SPSS, Statistica, KXEN, WPS and CART.

SARITA DIGUMARTI

Sarita has a Master's degree in Quantitative Economics, from Tufts University, Boston, and a PG Diploma in Management from T.A. Pai Management Institute, Manipal. She has over 10 years of analytics and consulting experience across diverse domains including FMCG, retail and healthcare. She has worked in both India and the US, helping clients tackle complex business problems applying analytical techniques.

ANSHUMAN ACHARYA

Anshuman is a graduate of IIT-Kanpur and IIM-Bangalore with over eleven years of experience. For the past five years, he has been with Wal-Mart, India operations and worked across finance, strategy, merchandising and supply chain. He has four years of experience consulting global CPG manufacturers and US retailers across formats. His other interests are theatre and training.