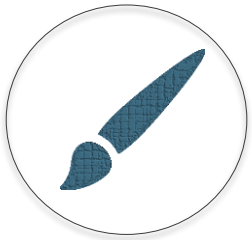




Introduction to Data Science

Overview

Objective

After completing this lesson you will be able to:

- Describe business analytics
- Explain the components of business analytics
- Explain the usage of business analytics in various domains

Business Analytics–Definition

- Business analytics (BA) refers to the tools, techniques and processes for continuous exploration and investigation of past data to gain insights and help in decision making.

- Business Analytics is an integration between science, technology and business context that assist data driven decision making.

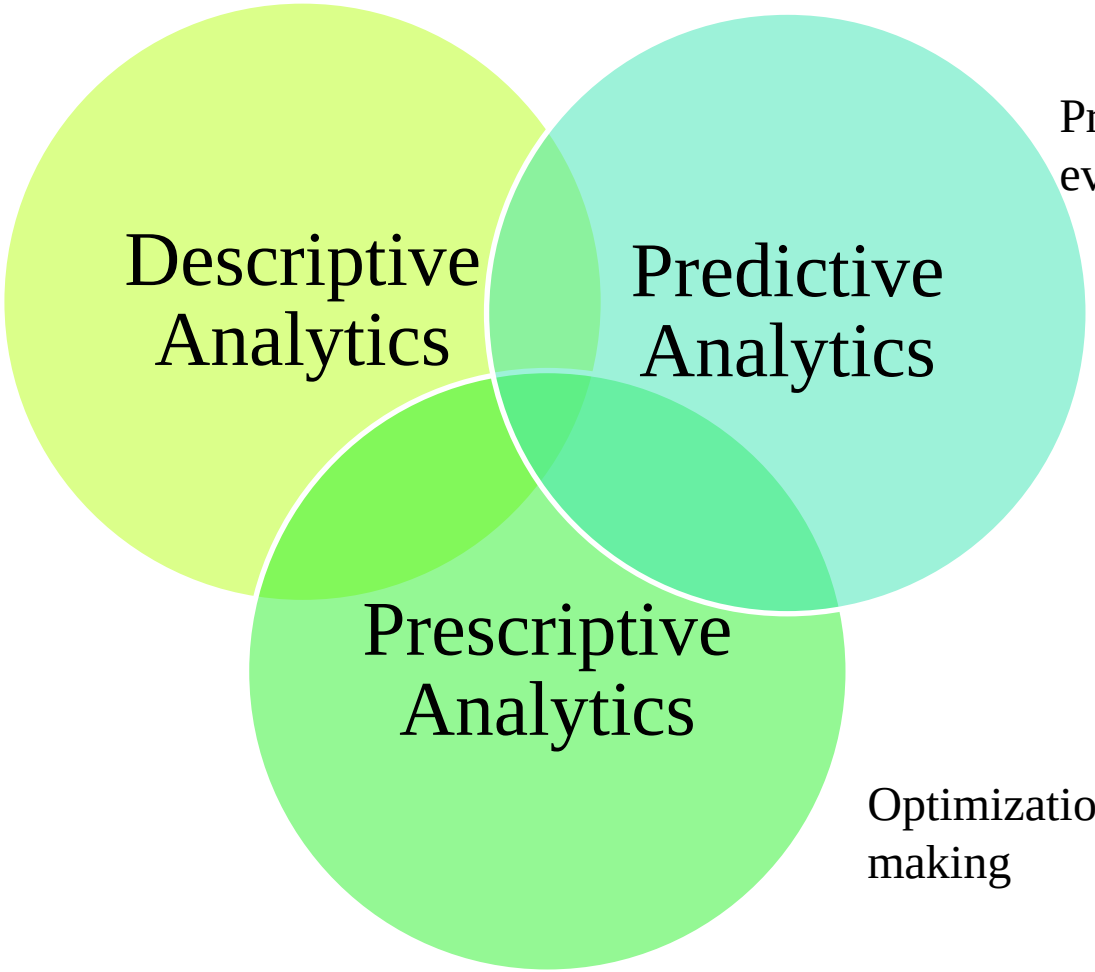
Science	Technology	Business Context
Data imputation techniques	Python	Which use case to work on?
Regression	SPSS	What business challenge solved?
Logistic regression	Minitab	What is the business benefit?
Visualization techniques	R	Maximizing Revenue/Profit
Data sampling startegies	Julia	Does it impacts topline?
Neural Net	Tableau	Does it impact bottom line?
Ensemble Models	PowerBI	
Deep Learning		

Analytics in Use

- Ecommerce – Forecast demand for each SKU!!!
- Bank – Understand and predict the NPA. Who will default on the loan!!!
- Insurance- What premium to charge for a vehicle policy renewal!!!
- Healthcare – What is the package price for different surgeries!!!
- Human Resource – Who will leave the organization in next quarter!!!
- Marketing – Segment customer!!! What Net Promoter Score indicates for a product/ service!!!
- Sentiments of customers on a new launch in automobile sector!!!
- Automobile - Warranty forecasting for spare parts!!!

Components of Business Analytics

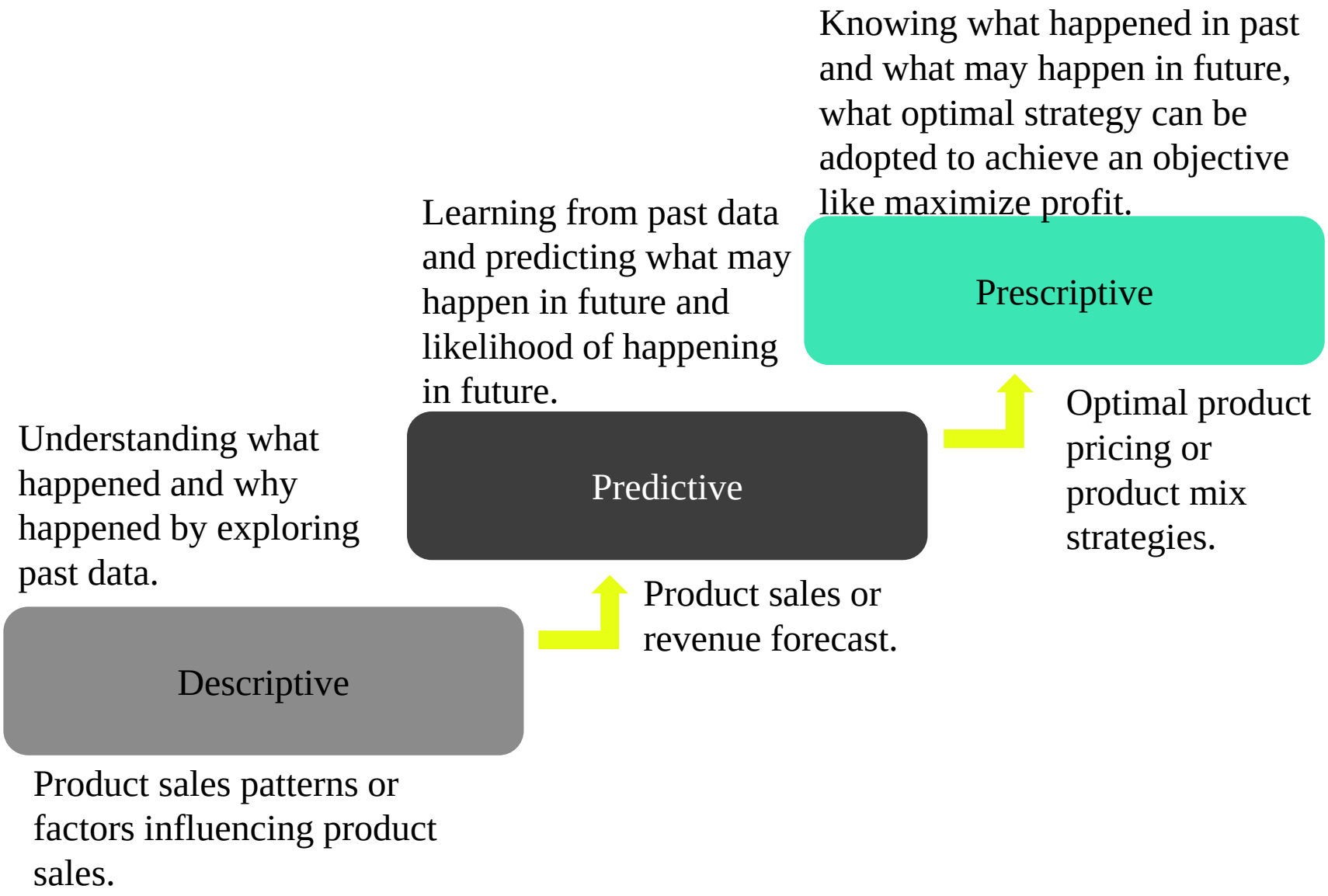
Data synthesis and Visualization



Predicting future events

Optimization and decision making

Components of Business Analytics



Broad Classification in Predictive Analytics

- Supervised Learning
  - Input (X's) and Output (Y) both are known features

- Unsupervised Learning
  - Input (X's) is known but Output (Y) is unknown

What Tools are available?

R Vs. Python Vs. Julia

Python (1991)	R (1997)	Julia (2012)
Guido van Rossum at Centrum Wiskunde & Informatica (CWI) in the Netherlands. Stable v1.0 in 1994	Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. Stable v1.0 in 2000.	Jeff Bezanson, Stefan Karpinski, Viral B. Shah at MIT, United States. Stable v1.0 in 2018.
Interpreter based language. Vectorized code makes mathematical operations faster as it helps run faster 'for' loops in optimized C routines (internally).	Interpreter based language and thus translates one statement at a time to machine code. Vectorized code makes mathematical operations faster.	Complier based language. Translates the entire program into machine code at once (creates an exe). No need to write vectorized code as all the functions in Julia are written in Julia. It is as fast as C language.
Dynamically typed language. Data type to variable assigned at runtime.	Dynamically typed language. Data type to variable assigned at runtime.	Julia's type system is dynamic, but gains some of the advantages of static type systems by making it possible to indicate that certain values are of specific types.
Does not generate intermediary code or machine code and thus efficient in terms of memory use but overall time to execute the code is slow.	Does not generate intermediary code or machine code and thus efficient in terms of memory use but overall time to execute the code is slow.	Generates intermediary code and thus consumes memory but faster to run compared to interpreter based language.
General Purpose Language which makes deployment and production easy. Turned into a machine learning mainstream over years but was not necessarily built for it. However, easy adaption in analysis work and deployable solutions	Built for Statistical Analysis and popular in academics and research. Steep learning curve. Difficult to learn in beginning as it is unlike a general programming language.	General Purpose Language primarily built for data analysis and at the same time can be used to build applications and write microservices. Syntax would feel familiar to users of Python and R.
More than a lac libraries but not as many as compared to that in R for data analysis. Good libraries for machine and deep learning.	More than 12000 libraries for data analysis. Many advanced packages for statistical analysis and inferences.	Over 5000 libraries but catching up with R and Python

An interpreter produces a result from a program, while a compiler produces a program written in assembly language. More here.

Data Science Stack

Purpose	R	Python	Julia
Data load, clean-up and manipulation	data.table, dplyr, tidyr	Pandas, Numpy	CSV.jl, DataFrames.jl, Spark.jl (for integration to spark ecosystem). Storge and organize data JLD.jl (HDF5 format)
Data visualization	ggplot2	Matplotlib, seaborn	Gadfly.jl, Plots.jl, PyPlot.jl
Statistical Learning	stats, caret	scipy, statsmodel	StatsKit.jl, More packages here. Stats with Julia here., MultivariateStats.jl
Machine Learning	caret, h2o	sk-learn, h2o	MLJ.jl, ScikitLearn.jl
NPL	tidytext, spacyr	spacy, nltk	TextAnalysis.jl, Word2Vec.jl, Embeddings.jl
Deep learning	Keras, TensorFlow	Keras, TensorFlow	Flux.jl, keras.jl, TensorFlow.jl, Mocha.jl

For installation and useful material in Python, R, Julia: refer this link

- This is not a comprehensive list but a few packages from each languages to draw an analogy of their use while analyzing data.
- Documentation for most packages (dplyr, tidyr, ggplot2, tidytext) listed for R is in this link and this one for caret package. More about data.table here.

Framework For Decision Making

Opportunity Identification

- Domain knowledge is very important at this stage of the analytics project. This will be a major challenge for many companies who do not know the capabilities of analytics.

Collection of relevant data

- Once the problem is defined clearly, the project team should identify and collect the relevant data. This may be an interactive process since "relevant data" may not be known in advance in many analytics projects. The existence of ERP systems will be very useful at this stage.

Data Pre-processing

- This would include data imputation and the creation of additional variables such as interaction variables and dummy variables in the case of predictive analytics projects.

Model Building

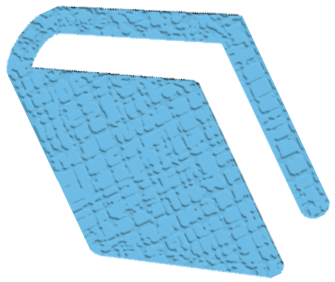
- Analytics model building is an iterative process that aims to find the best model. Several analytical tools and solution procedures will be used to find the best analytical model in this stage.

Communication of the data analysis

- The communication of the analytics output to the top management and clients plays a crucial role. Innovative data visualization techniques may be used in this stage.

Summary

Summary of the topics covered in this lesson:



- With the data explosion across industry, the usage of analytics in decision making will become the most critical factor for being competitive in business.
- Descriptive analytics becomes the stepping stone to all the complex problems which can be solved using analytics.

End of Lesson–Introduction to Business Analytics