Produced by:
Dr. Brenda Mullally bmullally@wit.ie
Ruth Barry rbarry@wit.ie
Department Computing Maths and Physics
Waterford Institute of Technology
www.wit.ie
moodle.wit.ie

Business Intelligence Information Insights II

Introduction

- **Data mining** attempts to discover patterns, trends, and relationships among data, especially nonobvious and unexpected patterns.
- The place to start is with a **data warehouse**—a huge database that is designed specifically to study patterns in data.
- A data mart is another source of data for data mining.

Data Mining Model

- Describing what happened:
  - Characterisation - is a summarisation of general features of objects in a target class (e.g. characterise the OurVideoStore customers who regularly rent more than 30 movies a year)
  - Patterns/associations/correlations
    - frequent itemset, (e.g. set of items that appear together – milk and bread)
    - frequent subsequences (sequential pattern – customers buy a laptop and then a digital camera)
  - Clustering analysis – identify natural groups based on their known characteristics
    - Outlier detection – dataset may contain objects that don't comply
- Predicting the future:
  - Classification and regression

The Data Mining Model: Describing What Happened

- **Describe** what happened
- *Descriptive techniques* are used to look for patterns with no outcome variable defined.
- Descriptive techniques can be of two types:
  - *Association – establishes relationships about items that occur together in a given record*
  - *Clustering – partition data into segments (segmentation) in which the members of data segment share similar qualities, discovery of groups, then they are understood and named by exploring distributions of variables in the groups.*

Association/Clustering Techniques

Association Techniques

Goal	Input Variables (Predictor)	Output Variables (Outcome)	Statistical Technique	Examples
Find large groups of cases in large data files that are similar on a small set of input characteristics,	Continuous or Discrete	No outcome variable	K-means Cluster Analysis	• Customer segments for marketing • Groups of similar insurance claims
To create large cluster memberships			Kohonen Neural Networks	• Cluster customers into segments based on demographics and buying patterns
Create small set associations and look for patterns between many categories	Logical	No outcome variable	Market Basket or Association Analysis with Apriori	• Identify which products are likely to be purchased together • Identify which courses students are likely to take together
Create small set associations and look for patterns between many categories	Logical or numeric	No outcome variable	Market Basket or Association Analysis with GRI	• Identify which courses students are likely to take together
To create linkages between sets of items to display complex relationships	Continuous or Discrete	No outcome variable	Link Analysis	• To identify a relationship between a network of physicians and their prescriptions

Cluster Analysis for Data Mining

- Used for automatic identification of natural groupings of things (e.g. customers)
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data
- There is no output variable
- Also known as segmentation
- Most common clustering algorithm – K-means

Association Rule Mining

- A very popular DM method in business
- Finds interesting relationships (affinities) between variables (items or events)
- Part of machine learning family
- Employs unsupervised learning
- There is no output variable
- Also known as market basket analysis
- Often used as an example to describe DM to ordinary people, such as the famous "relationship between diapers and beers!"
- Most common algorithm - Apriori

Association Rule Mining

- Are all association rules interesting and useful?

  A Generic Rule: **X ⇒ Y [S%, C%]**

  X, Y: products and/or services

  X: Left-hand-side (LHS)

  Y: Right-hand-side (RHS)

  S: Support: how often **X** and **Y** go together

  C: Confidence: how often **Y** go together with the **X**

  Example: {Laptop Computer, Antivirus Software} ⇒ {Extended Service Plan} [30%, 70%]

Association Rule Mining

- Input: the simple point-of-sale transaction data
- Output: Most frequent relationships among items
- Example: according to the transaction data…

  "Customer who bought a laptop computer and a virus protection software, also bought extended service plan 30 percent of the time" "70% of the transactions for the sale of extended service plan also purchased a laptop and virus protection"
- How do you use such a pattern/knowledge?
  - Put the items next to each other for ease of finding
  - Promote the items as a package (do not put one on sale if the other(s) are on sale)
  - Place items far apart from each other so that the customer has to walk the aisles to search for it, and by doing so potentially see and buy other items

Association Rule Mining

- Representative applications of association rule mining include
  - In business: cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - In medicine: relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)

Are all patterns interesting?

- What makes a pattern interesting?
- Can a system generate only the interesting ones?

  - Is it easily understood by humans?
  - Is it valid on new or test data with some degree of certainty?
  - Is it potentially useful?
  - Is it novel?

  - It is also interesting if a pattern validates a hypothesis that the user sought to confirm.
  - An interesting pattern represents knowledge.

Data Mining Model: Predicting what will happen

- To predict what will happen means to develop a model that uses historical data to predict an outcome based on a set of input characteristics.
- Predictive techniques require the use of past history with the intent to predict future behaviour.
- DM techniques in this area serve to classify the outcome variable into predefined categories.

Data Model: Predicting what will happen

- **Prediction/Forecasting** estimates future values based on patterns within large sets of data, this prediction can be labeled for determining weather forecast as 'sunny' or 'rainy' . use input to produce some classification of output, e.g.
  - A pattern has been found already in a set of data.
  - Given a new set of data, you can predict which of these classes it belongs too.
- **Regression** is a well-known statistical technique that is used to map data to a prediction value e.g. a real number 65℉
  - How accurate am I with this? Use classification model to give an actual measure with how close you are to the target – 95% correct

Data Mining Model - Prediction

- Classification

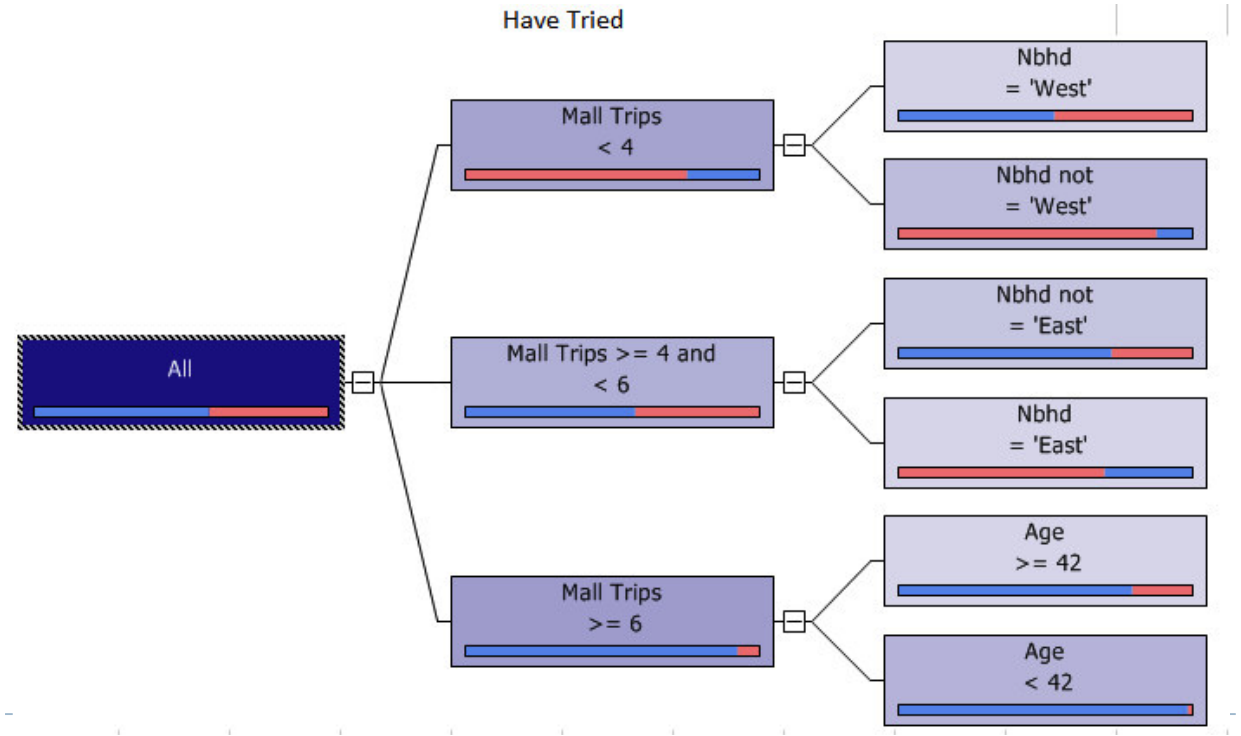
  Supervised induction used to analyze the historical data stored in a database and to generate a model that can predict future behavior
- Part of the machine-learning family
- Employ supervised learning
- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature
- Most common algorithm/technique: Decision trees

Predicting the future

- Classification Trees (decision trees), nonlinear relationships discovered
  - The basic idea of classification trees is to split a box of observations into two or more boxes so that each box is more "pure" than the original box, meaning that each box is more nearly Yes than No, or vice versa.
    - Each of these boxes can be split on another variable (or even the same variable) to make them purer.
    - This split continues until the boxes are either sufficiently pure or they contain very few cases.
  - The attractive aspect of this method is that the final result is a set of simple rules for classification.

Classification Trees

- The final tree might look like the one below.
  - Each box has a bar that shows the purity of the corresponding box, where blue corresponds to Yes values and red corresponds to No values.



Predicting the future

- **Data partitioning** plays an important role in classification.
  - The data set is partitioned into two or even three distinct subsets before algorithms are applied.
    - The first subset, usually with about 70% to 80% of the records, is called the **training** set. The algorithm is trained with data in the training set.
    - The second subset, called the **testing** set, usually contains the rest of the data. The model from the training set is tested on the testing set.
    - Some software packages might also let you specify a third subset, often called a **prediction** set, where the values of the dependent variables are unknown. Then you can use the model to classify these unknown values.

Further Data Mining Concepts and Applications

- **Sequence discovery**

  The identification of associations over time
- **Visualization** can be used in conjunction with data mining to gain a clearer understanding of many underlying relationships

Further Data Mining Concepts and Applications

- **Hypothesis-driven data mining**

  Begins with a proposition by the user, who then seeks to validate the truthfulness of the proposition

- **Discovery-driven data mining**

  Finds patterns, associations, and relationships among the data in order to uncover facts that were previously unknown or not even contemplated by an organization

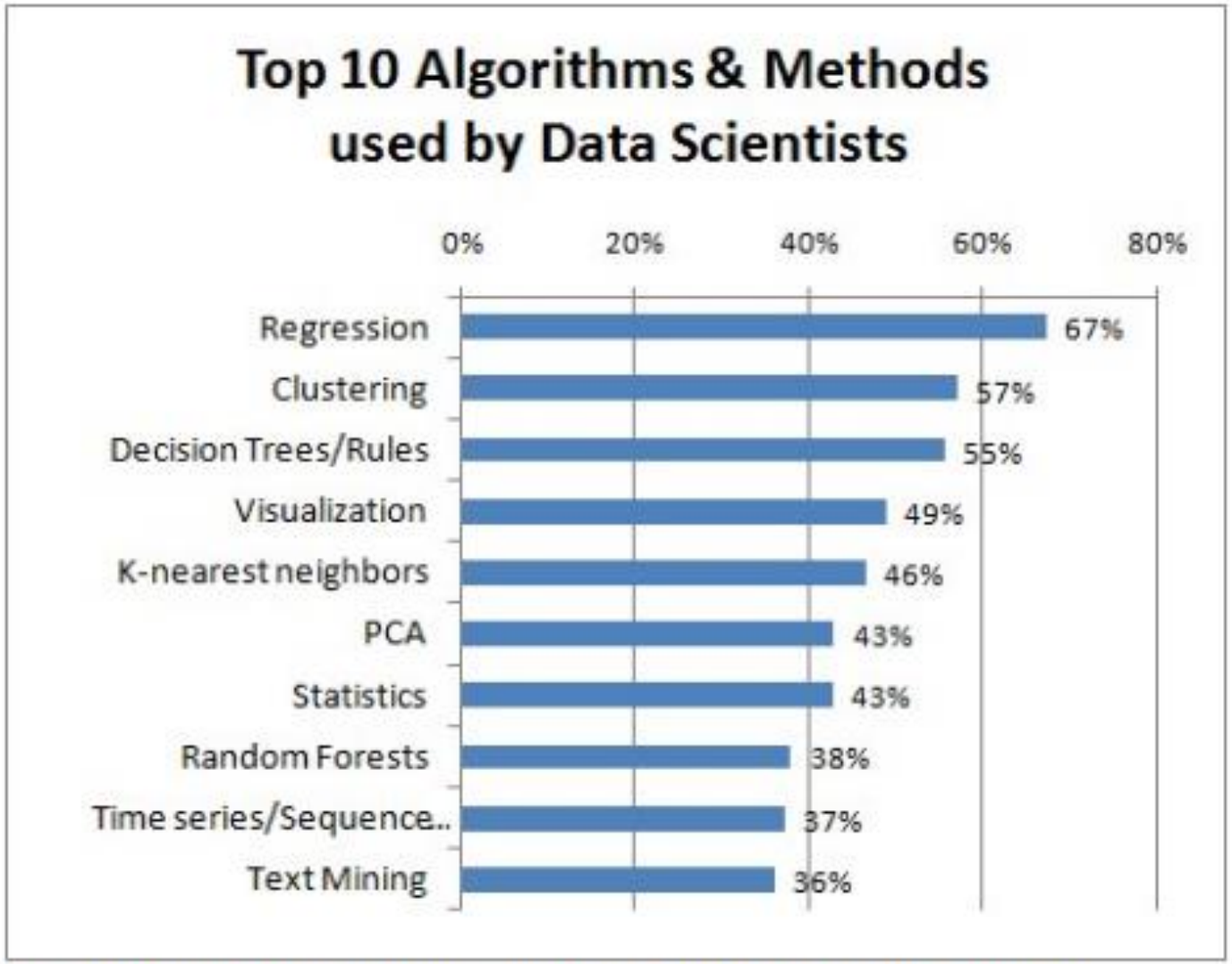


Fig. 1: Top 10 algorithms & methods used by Data Scientists.
See full table of all algorithms and methods at the end of the post.

- www.kdnuggest.com

Tools

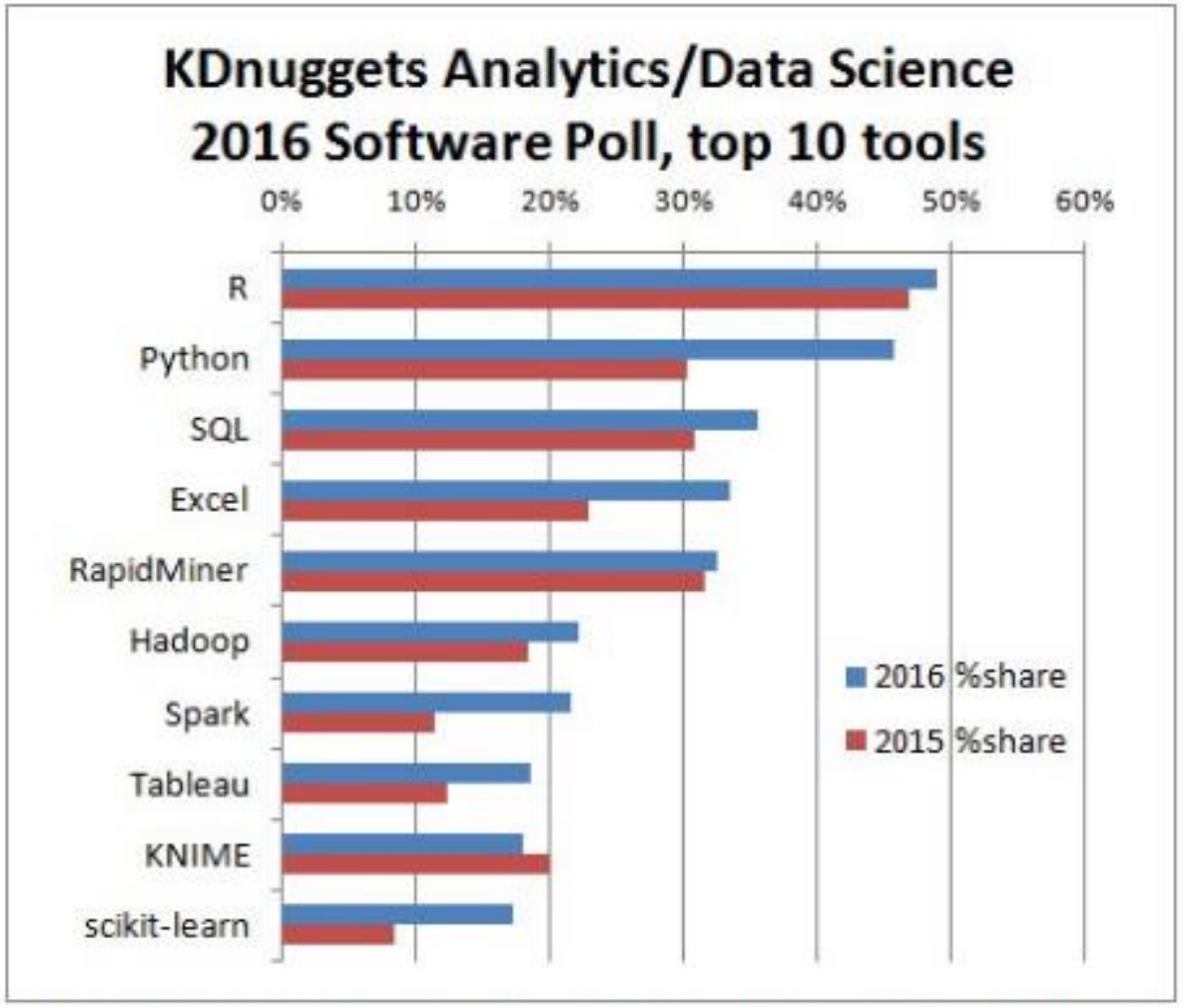


Fig 1: KDnuggets Analytics/Data Science 2016 Software Poll: top 10 most popular tools in 2016

www.kdnuggets.com

3 year comparison

Careercast.com

- Top 10 careers 2016
  - Data Scientist made it to the list in 2015 and now tops number 1 for 2016.
  - Statistician is second.
  - Mathematician is sixth
  - Actuary is tenth.
- Booming market for those that deal with numbers.

- U.S study conducted each year using Bureau of Labour Statistics(BLS) using environmental (work week, emotional, physical), income (start, midlevel, top) , outlook (growth, unemployment) and stress (travel, deadlines, competitiveness etc) factors.

Data Mining Myths

- Data mining …
  - provides instant solutions/predictions.
  - is not yet viable for business applications.
  - requires a separate, dedicated database.
  - can only be done by those with advanced degrees.
  - is only for large firms that have lots of customer data.
  - is another name for good-old statistics.

Data Mining?

"Not everything that can be counted counts, and not everything that counts can be counted" William Bruce Cameron 1963

"Prediction is very difficult, especially about the future" Neil Bohr 1918

"If we have data, let's look at data. If all we have are opinions, let's go with mine." – Jim Barksdale, former CEO of Netscape Communications Corporation.

"Torture the data, and it will confess to anything." – Ronald Coase, Economics, Nobel Prize Laureate

- Ted Talks what do we do with all this data?
- Ted talks How to use data to make a hit TV show
- https://www.ted.com/talks/alessandro_acquisti_why_privacy_matters