IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

INTRODUCTION

LECTURE 01

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

INTRODUCTION

- What is Business Analytics?
  - "Business analytics is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states." - Gartner IT Glossary
  - Includes data mining, predictive analytics, applied analytics and statistics, and is delivered as an application suitable for a business user
- Analytics can be classified as:
  - Descriptive analytics
  - Predictive analytics
  - Prescriptive analytics

INTRODUCTION

- Descriptive analytics involve gathering, organizing, tabulating, and depicting data and then describing the characteristics of what you are studying.
  - Also called reporting in managerial lingo
  - First phase of analytics
  - Though useful, it doesn't inform you about why the results happen or what can happen in future.

INTRODUCTION

- Predictive analytics use the past to predict the future.
  - Identify associations among different variables and predict the likelihood of a phenomenon reoccurring on the basis of those relationships
- Correlation vs. Causation
- Prescriptive analytics suggest a course of action.
  - Recommends decisions entailing mathematical and computational models
  - Final phase of analytics

INTRODUCTION

- Methods from statistics, forecasting, data mining, experimental design are used in Business Analytics
- What is Data Mining?
  - “Extracting useful information from large datasets” - Hand et al. 2001
  - “The process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques.” - Gartner IT Glossary

INTRODUCTION

- Where is Data Mining Used?
  - Variety of domains:
    - Predicting the response of a drug or a medical treatment on the patient suffering from a serious disease
    - Predicting whether an intercepted communication is about a potential terror attack
    - Predicting whether a packet of network data can pose a cybersecurity threat

INTRODUCTION

- Where is Data Mining Used?
  - Common business questions:
    - Which customers are most likely to respond to the marketing or promotional offer?
    - Which customers are most likely to default on loan?
    - Which customers are most likely to subscribe to a magazine?

INTRODUCTION

- Data Mining Genesis
  - An interdisciplinary subfield of computer science
  - Originates from the fields of machine learning and statistics
  - Data mining as “statistics at scale and speed” – Pregibon 1999
  - Extension: “statistics at scale, speed, and simplicity” – Shmueli et al. 2010

INTRODUCTION

Classical Statistical Setting

- Data scarcity and Computational difficulty
- Same sample is used to compute an estimate and to check its reliability
- Logic of inference: confidence intervals and hypothesis tests (Inference is determining whether a pattern or result might have happened by chance)

Data Mining Paradigm

- Large datasets and fast computing powers
- Fitting a model with one sample and evaluating the performance using another sample
- Machine learning techniques, such as trees and neural networks are less structured and more computationally intensive in comparison to statistical techniques

INTRODUCTION

- Rapid Growth of Data
  - Millions of transactions on a daily basis
    - Organized retailers such as Shoppers Stop, Big Bazaar, and Pantaloons
    - E-commerce retailers such as Flipkart, Amazon, and Snapdeal
  - Growing economy and Internet growth
  - Decreasing cost and increasing availability of automatic data capture mechanisms, e.g., Bar codes, POS devices, click-stream data, GPS data
  - Operational databases to data warehouse and data marts
  - Constant declining cost of data storage and improving processing capabilities

INTRODUCTION

- Core of this course focuses on
  - Predictive Analytics consisting of tasks of
    - Prediction,
    - Classification,
    - Association rules

- In Data mining, typically several different methods are applied for a particular goal and the most useful is selected

INTRODUCTION

- Usefulness of a method
  - Goal of the analysis
  - Underlying assumptions of the method
  - Size of the dataset
  - Types of pattern in the dataset
- Dataset Example: Sedan Car owner
  - Goal: Income level and Household Area is used to classify whether a household owns a sedan car

INTRODUCTION

- Dataset format
  - Tabular or matrix format: variables in columns and observations in rows
  - Each row represents a household (unit of analysis) in SedanCar dataset
- R and RStudio
  - R is a programming language and software environment for statistical computing and graphics.
  - It is widely used by statisticians and data miners
  - RStudio is the most commonly used integrated development environment (IDE) for R.

INTRODUCTION

- Key Terms
  - Algorithm
    - A specific sequence of actions or set of rules to be followed to perform a task.
    - Algorithms are used to implement data mining techniques such as trees, neural networks etc.
  - Model
    - By model, we mean data mining model here
    - A data mining model is an application of a data mining technique on dataset

INTRODUCTION

- Key Terms
  - Variable
    - Operationalized way of representing a characteristic of an object, event, or phenomenon
    - A variable can take different values in different situations.
  - Input variable, Independent variable, Feature, Field, Attribute, or Predictor
    - Input variable is an input to the model

INTRODUCTION

- Key Terms
  - Output variable, Outcome variable , Dependent variable, Target variable, or Response
    - Output variable is an output of the model
  - Record, observation, case, row
    - Observation is the unit of analysis on which the variable measurements are taken such as a customer, a household, an organization, an industry etc.

INTRODUCTION

- In Data Mining and related domains, generally two types of variables are used:
  - Categorical
    - Nominal
    - Ordinal
  - Continuous
    - Interval
    - Ratio

INTRODUCTION

- Understanding the type of variables in a dataset is important
  - To identify an appropriate statistical or data mining technique
  - Proper interpretation of the data analysis results
- Data of these variable types are either quantitative or qualitative in nature
  - Quantitative data measure numeric values and are expressed in number
  - Qualitative data measure types and are expressed by a label, or a numeric code

INTRODUCTION

- Structure of these variable types increases from nominal to ratio in a hierarchical fashion

- Nominal
  - Values indicate distinct types, e.g., gender, nationality, religion, PIN code, employee ID
  - Only two operations = and ≠ are supported

INTRODUCTION

- Ordinal
  - Values indicate a natural order or sequence, e.g., academic grades, Likert scale, quality of a food item
  - Four additional operations $<$, $\leq$, $>$, $\geq$ are supported
- Interval
  - Difference between two values is also meaningful
  - Values may be in reference to a somewhat arbitrary zero point
  - Celsius temperature, Fahrenheit temperature, location variables: Distance from landmarks, geographical coordinates (latitude & longitude), calendar dates

INTRODUCTION

- Interval
  - Two additional operations +, - are supported
- Ratio
  - Ratio of two values is also meaningful. Values are in reference to an absolute zero point
  - Kelvin temperature, age, length, weight, height, income
  - Two additional operations ×, ÷ are supported

INTRODUCTION

- Conversion from one variable type to other
  - High structure variable type can be converted into low structure variable type
  - For example, a ratio variable ‘age’ can be converted into an ordinal variable ‘age group’

INTRODUCTION

- Course Roadmap
  - Module I: General Overview of Data Mining and its Components
  - Module II: Data Preparation and Exploration
  - Module III: Performance Metrics and Assessment
  - Module IV: Supervised Learning Methods
  - Module V: Unsupervised Learning Methods
  - Module VI: Time Series Forecasting
  - Module VII: Conclusion

INTRODUCTION

- Supplementary Lectures
  - Introduction to R
  - Basic Statistical Methods

Key References

- HBR Video (Business Analytics Defined by Thomas H. Davenport)
- Gartner IT Glossary
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…