[Lecture Note] Introduction

MGS3701: Data Mining

Chungil CHae

Wed, 26 February 2025

This book has by far the most comprehensive review of business analytics methods that the author have ever seen, covering everything from classical approaches such as linear and logistic regression, through to modern methods like neural networks, bagging and boosting, and even much more business specific procedures such as social network analysis and text mining. If not the bible, it is at the least a definitive manual on the subject. However, just as important as the list of topics, is the way that they are all presented in an applied fashion using business applications. Indeed the last chapter is entirely dedicated to 10 separate cases where business analytics approaches can be applied.

Learning Objectives

1. What Is Business Analytics?
2. What Is Data Mining?
3. Data Mining and Related Terms
4. Big Data
5. Data Science
6. Why Are There So Many Different Methods?
7. Terminology and Notation
8. Road Maps to This Book (Shmueli et al., 2017)

SHORT VIDEO INTRODUCTION

https://www.youtube.com/watch?v=diaZdX1s5L4

What Is Business Analytics?

Business Analytics

- Business Analytics (BA) involves using quantitative data to aid decision-making, with its applications and interpretations varying across different organizations.
  - For example, a British tabloid once utilized it to test which images on their website, like cats or dogs, garnered more views.
  - In contrast, the Washington Post uses analytics to target specific influential audiences, such as defense contractors, by tracking reader behaviors like time of day and subscription details.
  - BA encompasses simple data analysis techniques such as counting and basic arithmetic.
- However, it also includes more advanced practices known as Business Intelligence (BI),which focuses on data visualization and dynamic reporting to help understand past and current events.
  - BI tools have evolved from static reports to interactive dashboards that provide real-time data interaction, enhancing managerial decision-making.
- Furthermore, Business Analytics has expanded to include sophisticated statistical and data mining methods aimed at exploring data relationships, making predictions, and forecasting future trends.
  - Techniques like regression models describe relationships and predict outcomes.
  - The field of BA is also supported by methods like A-B testing used in pricing strategies, demonstrating its practical implications in various business contexts.
  - However, successful deployment of BA requires a clear understanding of the business context and the functionality of analytics tools to avoid misapplication.
- Overall, Business Analytics has evolved from basic data reporting (BI) to a comprehensive toolkit that includes advanced analytics, emphasizing the necessity for strategic application aligned with business objectives.

Who Uses Predictive Analytics?

The integration of predictive analytics into various sectors has significantly enhanced organizational capabilities due to the growing availability of data. Key examples include:

- Credit Scoring:
  - Credit scoring utilizes predictive modeling to assess an individual's likelihood of repaying debts. Rather than being an arbitrary measure, it derives from historical data analysis to forecast future repayment behaviors. This established method helps financial institutions determine creditworthiness efficiently.
- Future Purchases:

  - An instance of the application of predictive models in marketing is demonstrated by Target's method to infer whether a customer is likely pregnant based on their shopping patterns. This insight allows Target to send tailored promotions to potential mothers at the early stages of pregnancy, optimizing marketing efforts and potentially increasing sales during a crucial buying period.

- Tax Evasion:

  - The U.S. Internal Revenue Service (IRS) has leveraged predictive analytics to enhance its enforcement strategies. By employing predictive models, the IRS is reportedly 25 times more successful in identifying cases of tax evasion. This focused approach allows them to concentrate resources on auditing individuals who are statistically more likely to have evaded taxes, making their processes more effective and efficient.

What Is Data Mining?

Data mining is

- In the context of this book, data mining extends beyond simple counting, descriptive statistics, and rule-based methods, venturing into more sophisticated realms of business analytics.
- While data visualization serves as an introductory tool in advanced analytics, the primary focus of the book is on deeper, more complex data analytics tools.
- These include both statistical and machine-learning techniques designed to automate and enhance decision-making processes, with a strong emphasis on prediction at an individual level.
- For instance, rather than merely analyzing broad relationships like the connection between advertising and sales, the book explores targeted questions such as which specific advertisement or recommendation should be presented to a particular online shopper in real-time.
- Additionally, it covers methods for clustering customers into distinct groups or "personas" tailored for different marketing strategies, and details how new prospects can be assigned to these personas for more effective engagement.
- The rise of Big Data has further propelled data mining into prominence.
- These methods are particularly adept at handling vast datasets and are key to unlocking valuable insights from the data deluge, thanks to their powerful analytical capabilities and automation potential.
- This shift highlights the advanced level of analytics discussed in the book, which prioritizes actionable insights and customization in business strategies.

Data Mining and Related Terms

- The analytics field is expanding rapidly in both the scope of its applications and the prevalence of its use across organizations. This growth has led to considerable inconsistencies and overlaps in terminology.
- For example, "data mining" is defined differently by different groups:

  - the general public may view it as an invasive search of personal data;

  - a major consulting firm might focus on identifying trends in historical data under a "data mining department," while relegating more advanced predictive modeling to an "advanced analytics department."

- Data mining occupies a space at the intersection of statistics and machine learning (also referred to as artificial intelligence).
- It employs a range of techniques, many of which have roots in long-established statistical methods like linear regression, logistic regression, discriminant analysis, and principal components analysis.
- However, unlike classical statistics where data scarcity and computation limitations are primary concerns, data mining operates in environments where data and computational power are abundant. This shift leads to what Daryl Pregibon describes as "statistics at scale and speed."
- Significantly, there are fundamental differences between statistics and machine learning.
  - Statistics typically focuses on inferring from a sample to the broader population about an average effect
  - Machine learning concentrates on predicting outcomes for individual instances.
- Furthermore, while statistical methods lean heavily on inference to determine if observed patterns occur by chance, data mining sidesteps such inference, which occasionally leads to overfitting.
- This occurs when a model so closely fits the sample data that it also captures random noise rather than just the underlying pattern.
- In this text, the term "machine learning" is used specifically for algorithms that learn directly from data, spotting local patterns in a layered or iterative manner.
- Conversely, "statistical models" refer to approaches applying a global structure to the data, such as linear regression, as opposed to algorithms like k-nearest-neighbors which focus on nearby data points for predictions.
- Ultimately, many IT and computer science professionals use the term "machine learning" more inclusively to refer to all the methodologies discussed in this book, highlighting the variations in how these terms are understood and applied across different sectors.

Big Data

- Data mining and Big Data are inextricably linked, with the landscape of Big Data defined by the "four V's": volume, velocity, variety, and veracity. These characteristics outline the challenges and opportunities presented by modern data sets:
  - Volume addresses the sheer amount of data.
  - Velocity pertains to the rapid rate at which data is generated and updated.
  - Variety indicates the different types of data being collected, from numerical to textual and beyond.
  - Veracity highlights the unreliability and organic nature of data generation processes, which often lack the rigorous controls found in traditional data collection.

  The scale of Big Data can be astonishing. Comparatively, if traditional statistical data (like those from a small-scale study) were the size of a period at the end of a sentence, a database

like Walmart's could be equated to the size of a football field. This doesn't even consider additional unstructured data from sources such as social media.

- The challenge of handling Big Data is significant but the potential rewards are great, as demonstrated by several practical examples:
  - OKCupid utilizes statistical models to predict which messaging strategies are most likely to engage users.
  - Telenor utilized predictive models to reduce subscriber turnover by 37%, identifying at-risk customers early and proactively engaging them.
  - Allstate improved the accuracy of predicting injury liability claims by threefold by incorporating detailed information about vehicle types.
- Big Data has enabled new technologies that were previously unfeasible.
  - For instance, the evolution of Google's search technology showcases this shift.
  - Initially, searches returned general results based on keywords.
  - However, as Google's database grew and incorporated user interaction data, searches became remarkably specific, capable of returning highly relevant results based on complex queries, such as pinpointing a humorous scene in an "I Love Lucy" episode based on a mixed-language script.

Thus, Big Data not only presents complex challenges but also unprecedented opportunities to derive deep insights and create value across various domains.

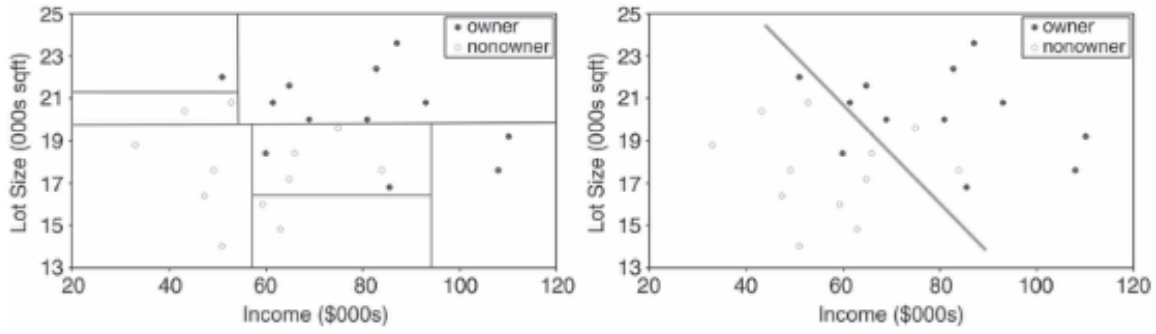
Data Science

- The emergence of Big Data has led to the creation of an entirely new professional field: data science.
- This discipline is broadly defined and encompassing, requiring a diverse set of skills ranging from statistics and machine learning to mathematics, programming, business, and IT.
- The term data scientist reflects this wide array of competencies, with most practitioners possessing what is described as a 'T'-shaped skill set.
- This includes deep expertise in one particular area (the vertical bar of the T) complemented by broader, shallower knowledge across other relevant disciplines (the horizontal top of the T).
- The role of programming within data science is debated; during the Strata+Hadoop World conference in October 2014, a majority of attendees considered programming essential, though a significant minority disagreed.
- Notably, despite the big data motivator, most data scientists do not usually deal with terabyte-size datasets on a daily basis.
- Instead, such massive datasets more commonly come into play during the deployment stage of a model, which is predominantly managed by IT professionals who tackle data-handling and systems integration challenges.
- The focus of our book, however, is on the earlier stages of data science work, which involve piloting and prototyping.
- This phase includes developing statistical and machine learning models that will later be integrated into larger systems.

- Our exploration centers on understanding which methods are suitable for different types of data and problems, how these methods function, their strengths and weaknesses, their requirements, and how to evaluate their effectiveness.
- Thus, data science, as a field, operates at the intersection of numerous disciplines, requiring a versatile skill set to navigate the complexities of both the creation and implementation of data-driven models and systems.
- The profession not only prioritizes technical proficiency but also necessitates an understanding of business and operational contexts to effectively leverage Big Data.

Why Are There So Many Different Methods?

- As explored in this book, and indeed across the broader literature on data mining, there is a plethora of methods available for prediction and classification.
- The coexistence of these diverse methods often raises questions regarding their relative merits and applicabilities.
- The reason for this variety is that **each method comes with its own set of strengths and weaknesses, making them suitable for different scenarios depending on several factors.**
- These factors include the size of the dataset, the types of patterns present in the data, adherence to the assumptions underlying each method, the level of noise in the data, and the specific objectives of the analysis.
- For instance, in a simplified example illustrated (as would be in Figure 1.1 of a hypothetical book), where the task is to differentiate between buyers and non-buyers of riding mowers based on household income and lot size, two different methods might be applied.
- One method might employ straightforward horizontal and vertical lines to segregate buyers from non-buyers, while another might utilize a diagonal line for separation.



- These differing approaches can yield varying results, and their effectiveness can change based on the context in which they are used.
- Consequently, it is typical practice in data mining to experiment with multiple methods to determine which one best meets the specific needs of the analysis at hand.
- This approach ensures a more tailored and potentially more accurate outcome in predicting or classifying data according to the challenge posed by the dataset.

- This pragmatic and flexible strategy highlights the inherently adaptive and investigative nature of data mining, where method selection is critical to achieving optimal results.

Terminology and Notation

Terms

Due to the hybrid origins and interdisciplinary nature of data mining, the terminology used by its practitioners often varies depending on their background in fields like machine learning (artificial intelligence) or statistics. Here's a list of commonly used terms in data mining, along with descriptions of how they might be referred to in different fields:

- Algorithm : A specific procedure for implementing a data mining technique, such as a classification tree or discriminant analysis. Attribute : Also referred to as Predictor.
- Case : Also known as Observation.
- Confidence : In the context of association rules, this is the conditional probability that an event will occur given another event. In statistics, it also refers to the concept of confidence intervals, which deal with the variability expected from one sample to another.
- Dependent Variable : Known in machine learning as the Response.
- Estimation : Also referred to as Prediction.
- Feature : Another term for Predictor.
- Holdout Data : A subset of data not used in training a model but reserved for testing its performance. Also called the validation set or test set.
- Input Variable : Also known as Predictor.
- Model : An algorithm applied to a dataset, complete with its parameter settings.
- Observation : The unit of analysis, also called instance, sample, example, case, record, pattern, or row.
- Outcome Variable : Another term for Response.
- Output Variable : Also known as Response.
- P (A | B) : The probability of event A given event B has occurred.
- Prediction : The act of predicting the value of a continuous variable; also called estimation.
- Predictor : An input variable (denoted by X) used in a predictive model. Also called feature, input variable, independent variable, or field.
- Profile : A set of measurements on an observation.
- Record : Synonymous with Observation.

- Response : The variable being predicted in supervised learning, also known as the dependent variable, output variable, target variable, or outcome variable.
- Sample : Used in statistics to denote a collection of observations, whereas in machine learning it typically refers to a single observation.
- Score : A predicted value or class. Scoring involves using a model developed with training data to predict values in new data.
- Success Class : In binary outcomes, this refers to the class of interest (e.g., purchasers in a purchase/ no purchase outcome).
- Supervised Learning : A type of machine learning where the model is trained on data where the outcome is known in order to learn to predict the outcome on new data.
- Target : Synonymous with Response. Test Data : Data reserved for testing the final model's performance, used only at the end stages of model building and selection.
- Training Data : Data used initially to fit and train a model. Unsupervised Learning : Learning patterns from data that do not involve predicting a specific output value.
- Validation Data : Data used to check how well a model fits, to tweak models, and to select the best among tried models.

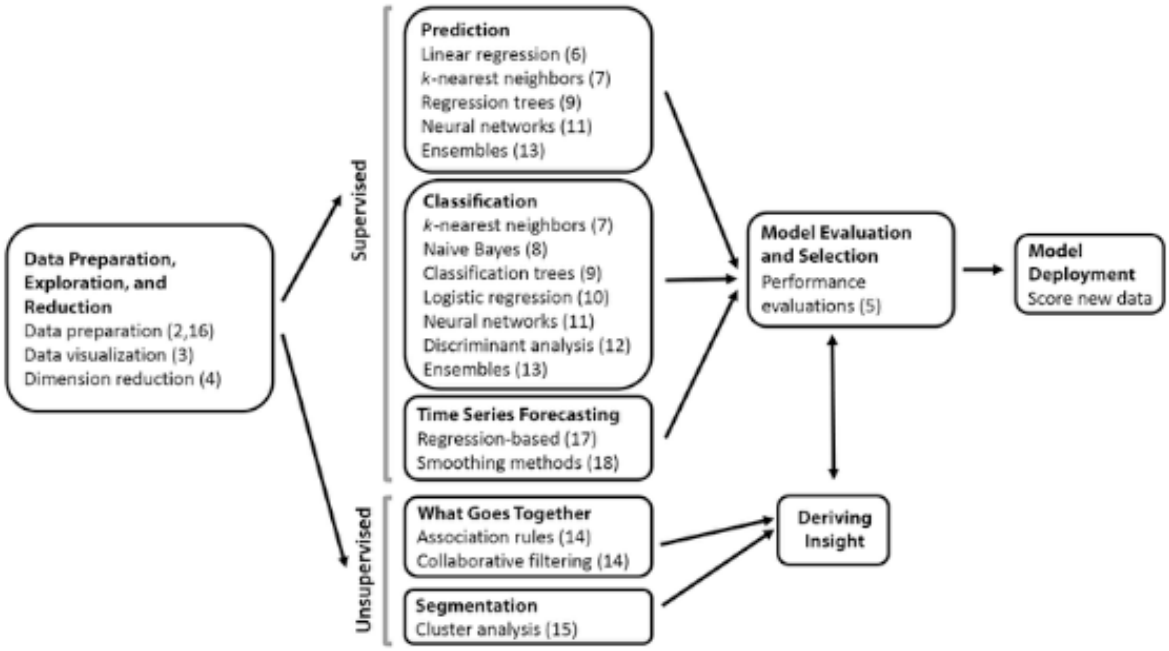
Road Maps to This Book

Order of Topics

The book is structured into eight parts, each focusing on distinct aspects and applications of data mining:

- Part I (Chapters 1–2): This section provides a broad introduction to data mining, outlining its key components and the fundamental concepts underpinning the field. It serves as the foundational groundwork for the more detailed explorations that follow.
- Part II (Chapters 3–4): Here, the focus shifts to the preliminary stages of data analysis, specifically on data exploration and dimension reduction. These chapters help readers understand how to streamline complex datasets into more manageable and interpretable forms.
- Part III (Chapter 5): Although it consists of only one chapter, this part dives deep into performance evaluation, covering everything from predictive performance metrics to the costs associated with misclassification. The principles discussed here are critical for accurately evaluating and comparing different supervised learning methodologies.
- Part IV (Chapters 6–13): This substantial segment discusses various popular supervised learning methods used for classification and prediction. The chapters are organized by the complexity of the algorithms, their popularity, and their accessibility. The concluding chapter in this part introduces the concept of ensembles and method combinations, which can enhance prediction accuracy.

- Part V (Chapters 14–15): Focused on unsupervised learning, this part examines methods for mining relationships through association rules and collaborative filtering, as well as cluster analysis. These techniques are vital for discovering patterns and groupings in data without predefined labels.
- Part VI (Chapters 16–18): These chapters are devoted to forecasting time series data. The initial chapter addresses general issues related to handling and interpreting time series data, followed by chapters on regression-based forecasting and smoothing methods. These approaches are essential for making predictions about future events based on historical data.
- Part VII (Chapters 19–20): This section explores specialized applications of data mining in social network analysis and text mining. These chapters demonstrate how data mining techniques can be adapted to analyze data from specific structures like social networks and textual content.
- Part VIII: The final part of the book presents a collection of case studies that illustrate the practical application of the techniques discussed in earlier chapters.
- While the chapters within the book are designed to stand alone, enabling readers to focus on topics of particular interest without requiring sequential reading, it is recommended that Parts I–III be read first to establish a solid understanding of the basics before delving into the more advanced topics in Parts IV–VIII.
- Additionally, Chapter 16 should ideally be read before proceeding with other chapters in Part VI to ensure a proper understanding of time series analysis fundamentals. This structured approach allows readers to build their knowledge progressively and effectively.



	Supervised		Unsupervised
	Continuous Response	**Categorical Response**	**No Response**
Continuous predictors	Linear regression (6) Neural nets (11) k-Nearest neighbors (7) Ensembles (13)	Logistic regression (10) Neural nets (11) Discriminant analysis (12) k-Nearest neighbors (7) Ensembles (13)	Principal components (4) Cluster analysis (15) Collaborative filtering (14)
Categorical predictors	Linear regression (6) Neural nets (11) Regression trees (9) Ensembles (13)	Neural nets (11) Classification trees (9) Logistic regression (10) Naive Bayes (8) Ensembles (13)	Association rules (14) Collaborative filtering (14)

***Numbers in parentheses indicate chapter number.**

House management

More to Read

Assignment

- What to do
- Requirement
  - PDF format
  - file name should be include your student id and name
    - stuID_name_title.pdf (e.g. 1111111_ChungilChae_SelfIntroduction.pdf)
- Due date
  - by DATE 11:59PM
  - NO LATE SUBMISSION ALLOWED!!!!

Reference

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data mining for business analytics: Concepts, techniques, and applications in r*. John Wiley & Sons.

