Data Science

Prof. Dr. Markus Breunig

Homework
in preparation of

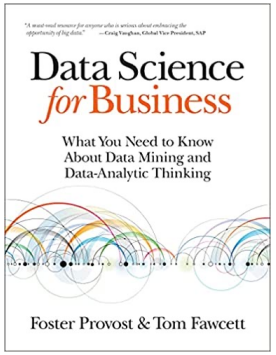
Lecture 2
–
Business Understanding

Photo by Dorothea OLDANI on Unsplash

Reading Assignment

- In this homework, you will be reading two chapters of the book

  Foster Provost, Tom Fawcett:
  Data Science for Business.
  (O'Reilly, 2013)



- The full text of the book is available as an e-book via the University Library (Catalogue OPAC).
- Start here: https://www.th-rosenheim.de/en/th-rosenheim/libraries/library-campus-rosenheim
- Note that both OPAC and O'Reilly's e-book-website do have rather peculiar user interface, which may need some getting used to. Make sure you select the correct version of the book (apart from the e-book version, there are also audio-book and physical versions available).

Introduction to Data Science

(Estimated Time: 1 hour)

- Please read **Chapter 1: Introduction: Data-Analytics Thinking**
- Close the book.
- Now try to answer the questions on the next page in your own words (English or German, your choice) - they are meant to give yourself feedback about how deeply you do understand the materials in the chapter.
- If you cannot answer one of the questions, re-open the book, re-read the corresponding part of the chapter, close the book and try to answer the question in your own words once again.
- When you're done, compare your answers to the answers given on the following pages.

Introduction to Data Science - Questions

- **Questions for Chapter 1**
  - What is the difference between Data Science and Data Mining?
  - Why does it make sense to understand data-analytic thinking even if you do not intend to apply Data Science yourself?
  - The book discusses how Wal-Mart applied data science when a Hurricane was on it's way to Florida. Discuss why this might be useful and what they gained from this project.
  - What is "customer churn" and why is a problem (in many industries, but specifically for mobile telecommunication providers)?
  - What does DDD stand for? Explain the concept. Why would a company want to employ DDD?
  - Why does data science overlap with DDD and not just support it
  - What does Big Data mean?
  - Explain how Signet Bank gained a significant competitive advantage by investing in data!
  - Give an example for the Fundamental concept: "Formulating data mining solutions and evaluating the results involves thinking carefully about the context in which they will be used."

Business Understanding

(Estimated Time: 1 hour)

- Please read **Chapter 2: Business Problems and Data Science Solutions**
- To verify your grasp of this material, please try to answer the questions on the next page in your own words (English or German, your choice).
- If you cannot answer one of the questions, re-open the book, re-read the corresponding part of the chapter, close the book and try to answer the question in your own words once again.
- When you're done, compare your answers to the answers given on the following pages.

Business Understanding - Questions

- **Questions for Chapter 2**
  - Why do we want to decompose a data-analytics problem and into what kind of pieces?
  - Explain the tasks "classification" and "scoring" and how they relate to each other!
  - What is the difference between Classification and Regression?
  - What does Clustering do?
  - Describe the difference between supervised and unsupervised learning. Give one example for method for each type.
  - What are the two phases of a machine learning/data mining system, and what is done in each of them?
  - Draw the CRISP DM process.
  - Explain each step of the CRISP DM process in 1-2 sentences in your own words.
  - Explain the difference between explanatory modelling and predictive modelling for regression analysis.

Introduction to Data Science – Answers (1)

- **Answers for the Questions for Chapter 1**
  - What is the difference between Data Science and Data Mining?
    Answer: They are very similar and often used interchangeably. Both are about extracting knowledge from data. If you want to make a distinction, Data science is more about the process and the underlying principles that guide the extraction of knowledge from data. Data mining is more about the technologies used, e.g. specific machine learning methods applied in the data science process.

  - Why does it make sense to understand data-analytic thinking even if you do not intend to apply Data Science yourself?
    Answer: Data-analytic thinking enables you to evaluate proposed data mining projects - you should be able to spot obvious flaws, unrealistic assumptions, and missing pieces.

  - The book discusses how Wal-Mart applied data science when a Hurricane was on it's way to Florida. Discuss why this might be useful and what they gained from this project.
    Answer: There are certain obvious items that will be bought before a Hurricane hits (like bottled water or flashlights) - these are well known, so a data science project to find these does not add value. But being able to predict the additional amount/demand for these items is helpful in managing stock levels. In addition, other, less obvious items are also sold in much higher-than-usual quantities, predicting these and the increase is again helpful to manage stock levels.

  - What is "customer churn" and why is a problem (in many industries, but specifically for mobile telecommunication providers)?
    Answer: Customer churn is when customers stop using a companies product or service and switch to a competitor. Companies try to reduce customer churn as it typically more expensive to gain a new customer than to keep a current one. This is particularly the case in saturated markets, i.e. markets with very few new customers.

Introduction to Data Science – Answers (2)

- **Answers for the Questions for Chapter 1**
  - What does DDD stand for? Explain the concept. Why would a company want to employ DDD?
    Answer: DDD = data-driven decision-makeing. Instead of deciding based on intuition/past experiences, decision are made based on hard data. Reasons for DDD: around 5% higher productivity, correlation with higher return on assets/equity/market value.
  - Why does data science overlap with DDD and not just support it?
    Answer: More and more decision are made automatically by systems based on data science/predictive analytics.
  - What does Big Data mean?
    Answer: Big Data means that a dataset is so large that it cannot be handled by a traditional data processing system any more.
  - Explain how Signet Bank gained a significant competitive advantage by investing in data!
    Answer: They only offered one set of credit card terms and only to customers that seemed credit worthy. They wanted to offer different terms to different customers in a way to optimized profitability, but modelling this was not possible with the data at hand. They invested in data by defining a number of different terms and randomly offering these to customers, and observing that happened. Naturally, this decreased their profitability, this was the cost of generating this data. Once the had the data however, they were able to successfully implement the optimized term/customer approach and became one of the largest and most profitable credit card company.
  - Give an example for the Fundamental concept: "Formulating data mining solutions and evaluating the results involves thinking carefully about the context in which they will be used."
    Answer: For the churn example, you need to decide if the expected value of a customer should be part of the project, or just the likelihood of him leaving. And you need to compare your solution to random decisions and also a very simple but smart "default" alternative.

Business Understanding – Answers (1)

- **Answers for the Questions for Chapter 2**
  - Why do we want to decompose a data-analytics problem and into what kind of pieces?
    Answer: We want to decompose it into pieces such that each pieces matches a known machine learning/data science task for which we have a tool available to solve it.
  - Explain the tasks "classification" and "scoring" and how they relate to each other!
    Answer: Classification produces a model which determines for a new example, which class it belongs to. Scoring determines a kind numerical value for each possible class, indicating the probability that the example belongs to this class. Most algorithms doing one of classification or scoring can be changed to do the other one.
  - What is the difference between Classification and Regression?
    Answer: Classification predicts a class (i.e. if something will happen), Regression a numerical value (how much something will happen).
  - What does Clustering do?
    Answer: It groups examples based on their similarity.
  - Describe the difference between supervised and unsupervised learning. Give one example for method for each type.
    Answer: In supervised learning, the learning algorithm is provided with the label/target value for each example. In unsupervised learning, such a label/target value is not provided to the learner. Examples for supervised learning: classification, regression; for unsupervised learning: clustering, dimensionality reduction, outlier detection.

Business Understanding – Answers (2)

- **Answers for the Questions for Chapter 2**
  - What are the two phases of a machine learning/data mining system, and what is done in each of them?
    Answer: Phase one is training the model, i.e. finding a model/the parameters of this model based on a (large) number if example observations. Phase two is using the model, i.e. applying the model to new data points/examples observations.

  - Draw the CRISP DM process.
    Answer: See Figure 2-2 on page 27 of the book.

  - Explain each step of the CRISP DM process in 1-2 sentences in your own words.
    Answer:
    Business Understanding: divide the business objective into one or more data science/machine learning problems.
    Data Understanding: getting an overview of available data and data that can be acquired. deciding on the match between machine learning method from the business problem and what data to use for it based on a deep understanding of the data.
    Data Preparation: Cleaning and restructuring the data so it works as input for the chosen machine learning algorithm, in particular identifying leaks in the data. Often the most time consuming step.
    Modelling: training the model.
    Evaluation: assess the quality of the results of modelling, in particular in respect to the business problem that needs to be solved.
    Deployment: getting the model or even the machine learning method itself to retrain the mode on new data into a production environment.

  - Explain the difference between explanatory modelling and predictive modelling for regression analysis.
    Answer: Explanatory modelling tries to find patterns in the data that are valid for the data given. Predictive modelling tries to find patterns that generalize to so-far-unseen data.