MTAT.03.319

Business Data Analytics

Lecture 8: Course Summary and Data Analytics Projects

Marlon Dumas and Veronika Plotnikova

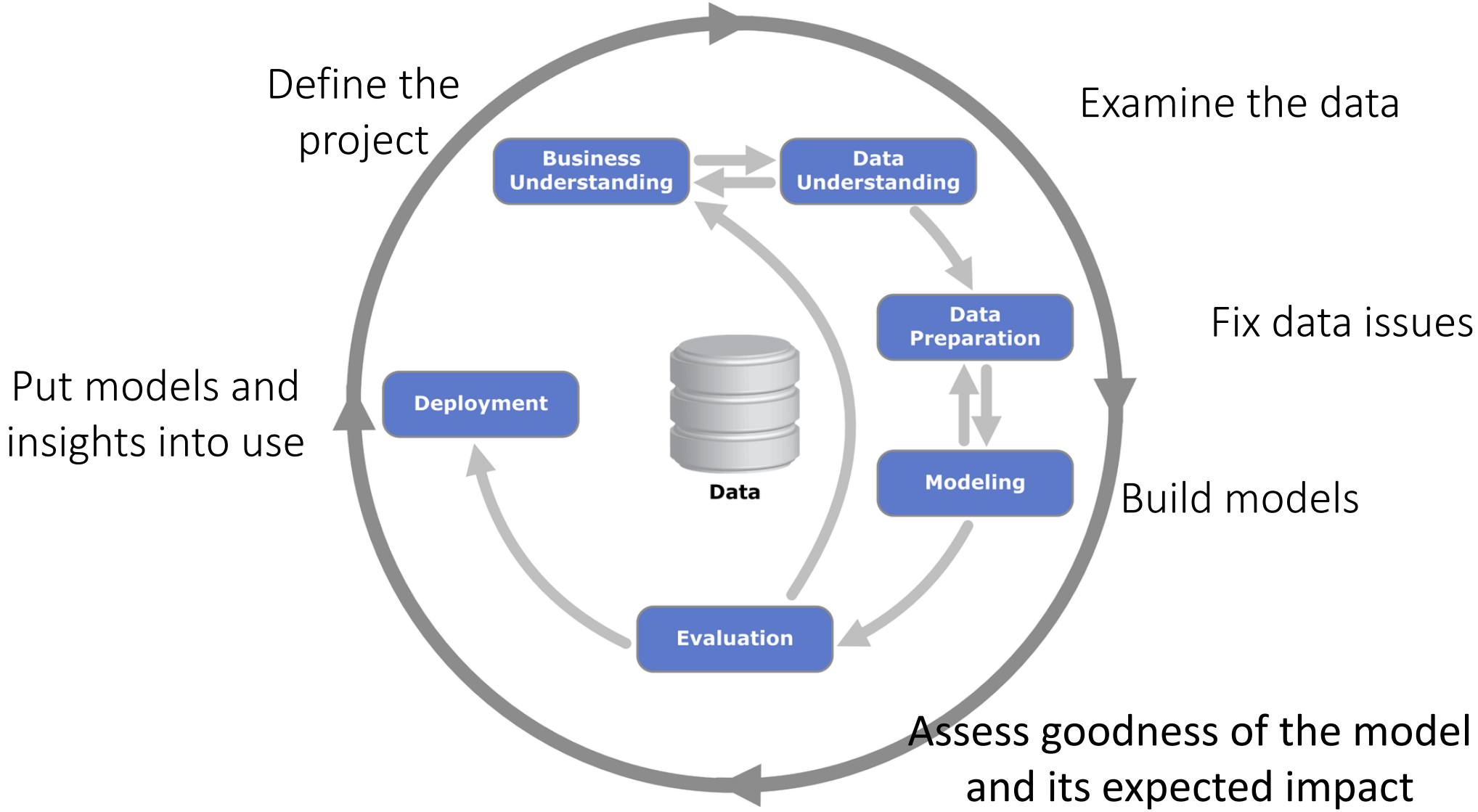
Outline for Today

- Recap
- Exam Preparation
- Data Analytics in Practice (guest lecture)

Recap

1. What is data analytics? Why we need it? How to approach it?
2. Data exploration: visualization & descriptive analysis
3. Customer segmentation
4. CLM – regression
5. CLM – classification (propensity, churn)
6. CLM – recommender systems (cross-sell/up-sell)
7. Brand monitoring – opinion mining

Recap: CRISP-DM
Cross-Industry Standard for Data Mining



Business Understanding

- Define the business objective
- Formulate the question(s)
- Identify target variable & attributes
- Define the success criteria
- Cost/benefit analysis

Who is involved?

- Business sponsor
- Domain expert(s)
- Analytics expert
- Data steward & DB expert

Data Understanding

- Data Collection
  - Identify data sources
  - Write queries
- Data Description
  - Document data quality issues
  - Compute basic statistics
- Data Exploration
  - Simple univariate data plots/distributions
  - Investigate attribute interactions
  - Data Quality Issues
    - Missing Values
    - Skewed Distributions

Data Preparation (cont.)

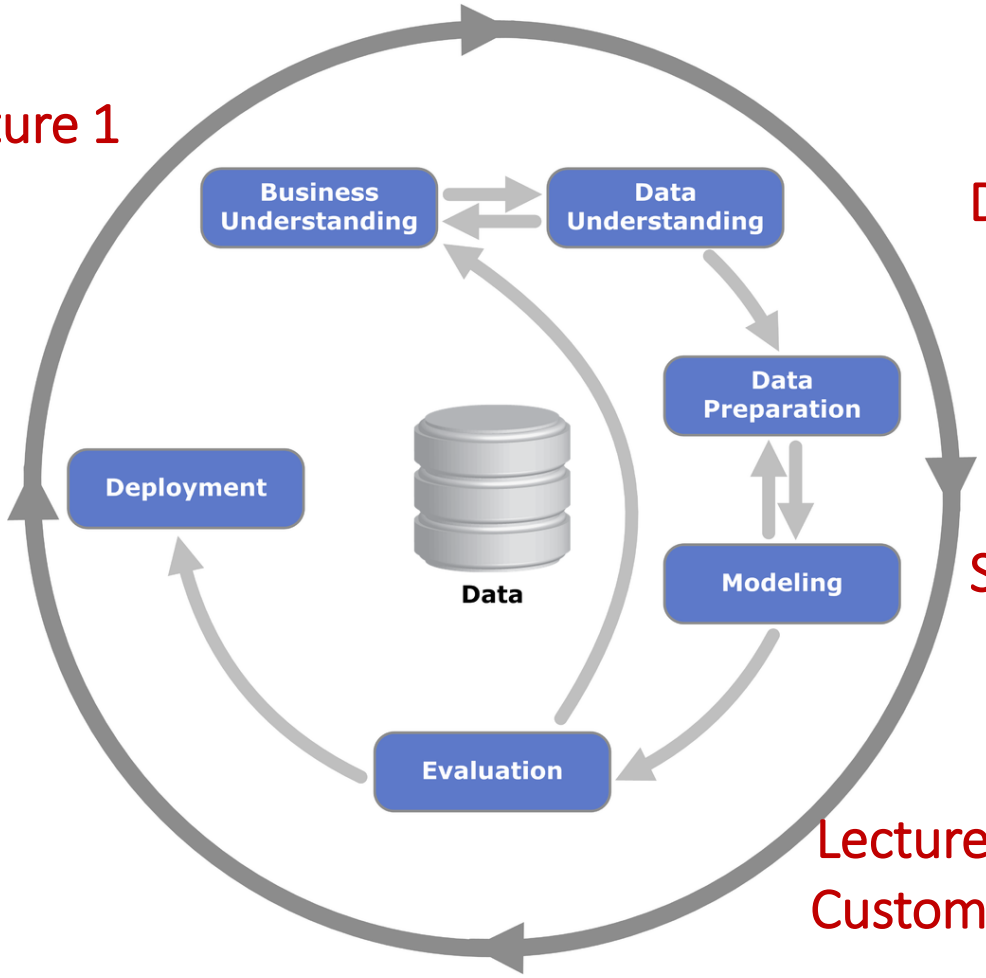
- Integrate Data
  - Joining multiple data tables/frames
  - Summarisation/aggregation of data
- Select Data
  - Attribute subset selection
  - Sampling (sometimes useful for large datasets)
- Transform data
  - Using functions such as log
  - Normalization/Discretisation/Binning
- Clean Data
  - Handling missing values/Outliers
- Enrich Data
  - Calculate derived attributes

Modeling

- Select modeling technique depending on type of problem/output
  - Supervised versus unsupervised
  - Regression versus classification
- Develop a testing regime
  - Select measures of model quality
  - Sampling (train versus test)
- Build Model
- Assess the model

CRISP-DM & Course Structure

Lecture 1



Lecture 2: Descriptive Analysis

Lecture 3: Customer Segmentation

Lectures 4-7: Modelling Techniques for Customer Lifecycle Management (CLM)

Recap – Customer Segmentation

- RFM model
  - What does it stand for? What is it useful for? How can it be used to group customers?
- Clustering
  - K-means clustering and hierarchical clustering
    - What are they? What do they need as input? What they provide as output?
    - What are their relative advantages and drawbacks?
  - How do we determine the *k* in k-means clustering?

Recap – Regression in CLM

- What is regression?
  - What is the input? What is the output?
- How do we train a regression model?
- How do we measure how good a regression model is?
- How can regression be used in Customer Lifecycle Management (CLM)?
- What is CLV (or CLTV)?

Recap – Classification in CLM

- What is classification?
  - What is the input? What is the output?
- How do we train a classification model?
  - Which methods are there? How to use them?
  - What is the difference between a white-box and a black-box classification?
- How do we measure how good a classification model is?
- What is over-fitting? How can we detect it?
- What is class imbalance? How does it impact classification?

Recommender systems for cross-and up-selling

- Market-basket analysis
  - What is it? What is it useful for?
- What is the relation between market-basket analysis and association rule mining?
- What is the input of output of association rule mining?
- How do we measure the goodness of association rules?
- Collaborative filtering: user-based versus item-based
  - What is it? What is it useful for? What is the tradeoff between user-based versus item-based collaborative filtering
- Tradeoffs between market-basket analysis and collaborative filtering

Brand Value Monitoring

- Opinion Mining / Sentiment Analysis
  - Dictionary-based technique
  - Documents-based technique
  - K-NN with similarity based on Term Frequency & Inverse Document Frequency
    - Calculate k nearest neighbours, take the closest one
- Text pre-processing
  - Tokenization: lower-casing, removing punctuation, numbers, etc.

    → list of words (tokens)
  - Stop words removal
  - Stemming

Exam Preparation

Exam structure

- 30 points
- 15 questions
  - 1 to 3 points per question
- Allowed: pen, pencil, calculator (no mobile phones)
- No notes/cheat sheet allowed
- 3 hours
  - But probably 2 hours will be enough

Types of Questions

1) Multiple choice questions
2) Fill in the blanks questions
3) Simple calculation questions (calculators allowed)
4) Analyzing/comparing plots: Answers the questions based on provides data/plots.
5) Problem-solving questions

Question 1: Simple multiple choice question

- (1 point) Mark the correct option(s) with a circle. ◯

Supervised learning differs from unsupervised learning in that supervised learning requires:

1. At least one input feature.
2. Input features to be categorical.
3. At least one output feature.
4. Output features to be categorical.

NOTE: a wrong selection cancels out a correct selection, e.g. two correct selections and one incorrect = one correct selection.

Question 2: Multiple choice question

- (2 points) Mark the correct option(s) with a circle. ◯

You work for an online shop of second-hand appliances. You built a regression model to predict the monetary value (y) of an appliance listed in the online shop. The features are: 1) initial price, 2) age of the product, 3) number of posted pictures of the product, 4) seller response time (in minutes). For each type of appliance you have a separate model. Let's say that for the vacuum cleaners your model looks like this (all features significant, index indicates the feature from the list):

$y = 40 + 3x_1 - 2x_2 + 1.18x_3 + 1.8x_4$

A. By increasing initial price by 1 euro, the final price increases on average by 40 euros
B. By increasing number of pictures by 1, the final price increases on average by 1.18
C. The number of pictures and initial price increase the estimated value of the product, while age of the product decreases the estimate.
D. An increase in response time by 1 min, decreases the estimated value in average by 1.8 euros.

Question 3: Interpreting a plot

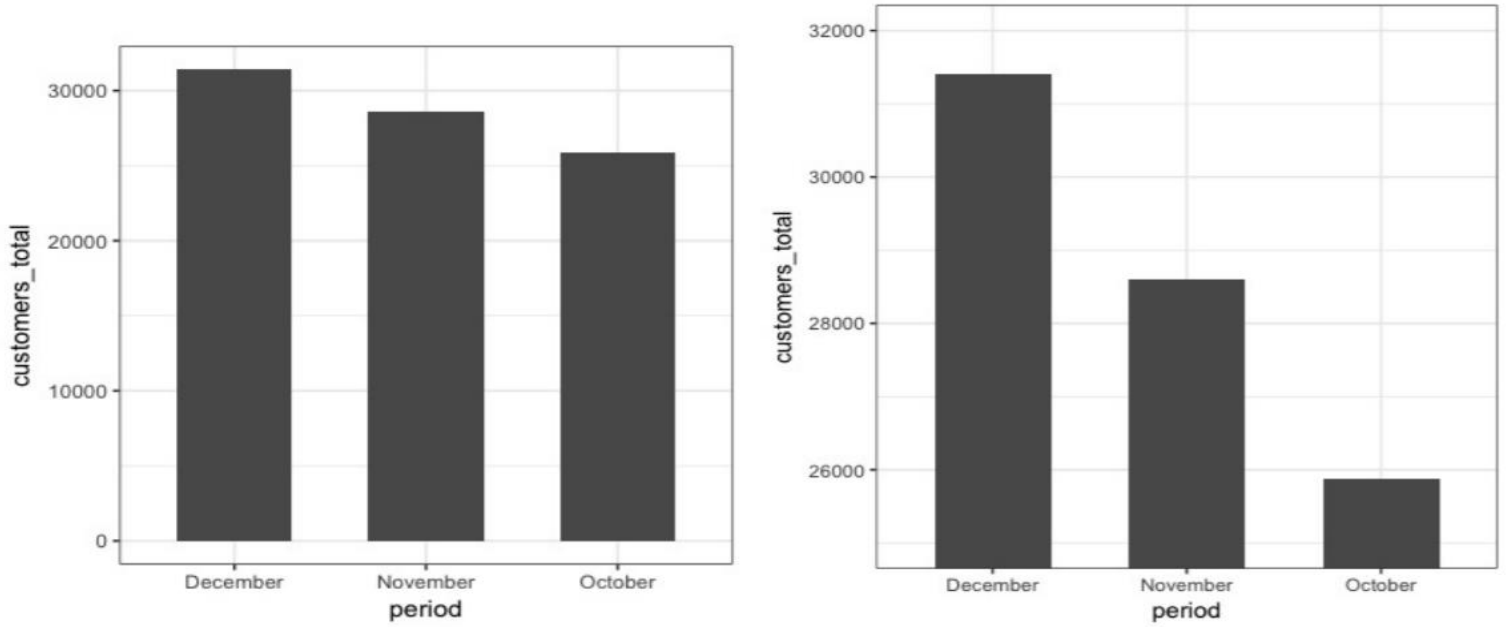
(1 point) Mark the correct option(s).

A. plot (a) as it does not skew the data showing the adequate difference

B. plot (b) as it highlights the difference between months

C. plot (a) as the last y-axis tick is closer to the maximum

D. Both plots are bad



Question 4: Fill in the Blanks (simple, definitional question)

- (1 point) RFM analysis ranks customers by considering ..... of their orders.

Question 5: Simple calculation question

- (2 points) Consider the following confusion matrix below:

	Positive	Negative	Total
Positive	50	30	80
Negative	25	15	40
total	75	45	120

- Calculate Precision and Recall based on these numbers.

Question 6: Problem-Solving (3 points)

Reflect on the following case. What modelling techniques to use and for what purpose? What features could be extracted to build the model?

- You are inventory manager in an e-commerce retail company that sells furniture products
- Your goal is to minimize:
  - Carrying cost (cost of holding inventory)
  - Lost sales revenue due to OOS (out-of-stock)
- The company has data about
  - All sales and all shipments for the past 5 years
  - All purchases from suppliers and all deliveries to the warehouse
- The number of Out-Of-Stock (OOS) events has increased by 5% in the past 2 years. The goal is to reduce OOS events, while capital inventory cost has been stable.