MTAT.03.319

Business Data Analytics

Lecture 14: Course Summary and Data Analytics in Practice

The slides are available under creative common license. The original owner of these slides is the University of Tartu.
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. Evaluation of Model’s Quality
7. CLM – recommender systems (cross-sell/up-sell) (2 Lectures)
8. Financial Forecasting
9. A/B Testing
10. Brand value monitoring
11. Networks in Finance
12. Fraud Detection
Recap: CRISP-DM
Cross-Industry Standard for Data Mining

- Define the project
- Examine the data
- Fix data issues
- Build models
- Assess goodness of the model and its expected impact
- Put models and insights into use
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 & Preparation

• Data Collection and Integration
  • Identify data sources
  • Write queries: Joining multiple data tables/frames

• Select Data
  • Attribute subset selection
  • Sampling (sometimes useful for large datasets)

• 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

• Transform data
  • Using functions such as log
  • 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
  • Supervised: 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
- Business Understanding
- Data Understanding
- Deployment
- Evaluation

Lecture 2: Descriptive Analysis
Lecture 3: Customer Segmentation
Lectures 4-5, 7-9, 11-13: 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?
Recap – Evaluation of Models

• 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?
  • Random Undersampling: Bias
  • Random Oversampling: Variance
  • SMOTE

• Missing data
  • Ignore it
  • Mean Imputation
  • Hot deck Imputation

• Metrics
  • Accuracy, Recall, Precision, F1-Score, AUC, logloss
Recap – Fraud Detection

• Fraud, Corruption ...
• We analysed using traditional machine learning + network science
• Imbalanced data: Apply balancing techniques
• How boosting is different from bagging?
  • Example of boosting: XgBoost.
  • Example of bagging: Random Forest.
Recap – 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
Recap – Financial Forecasting

• What is (univariate) time series?
• Components of a time series: Trend, Seasonality, Cycle, Noise
• Additive Vs. Multiplicative time series
• When to apply?
• Forecasting techniques: Qualitative & Quantitative
• Quantitative: ARIMA
  • Check if the series is Stationary?
Recap – A/B Testing

• Why we do A/B Testing?
  • Conversion rate

• A/B Testing Process
  • Study data and identify problem areas with the potential improvement
  • Create tools to for creating Hypothesis
  • Construct A/B test
  • If the results inconclusive, rework your hypothesis.

• 2 Types of Tests: 2 Tail or 1 Tail
Brand Value Monitoring

• Opinion Mining / Sentiment Analysis
  • Dictionary-based technique
  • Documents-based technique
  • K-NN & SVM with similarity based on Term Frequency & Inverse Document Frequency

• Text pre-processing
  • Tokenization: lower-casing, removing punctuation, numbers, etc.
    → list of words (tokens)
  • Stop words removal
  • Stemming
Recap – Networks in Finance

- Networks
- Graph: Nodes and Edges
- Directed Vs. Undirected.
  - Directed: Hubs and Authority, Reciprocity
- Weighted Vs. Unweighted.

- Metrics
  - Local and Global metrics
  - What is the interpretation?

- Communities
  - Bridge & weak ties
  - Cut vertex and edge
  - Cliques
  - Communities (modularity)

- Single Vs. MultiLayer N/W
- Linear Threshold model.
Small Discussion about Tools

• **R**: Developed by statisticians, libraries even for current packages.

• **Python**: New trending. Strong in new technologies but weaker in established old theoretical concepts.
  • R and python: Gives you more liberty in terms of data analysis.

• **SAS (more statistical SPSS)**

• **Weka**: UI Based.
DISCLAIMER

• It was only recap/summarization of the lectures.
• It should NOT be taken as possible topics/questions of the examination.
  • In other words if we missed something today from some lecture that does not mean it will not come in the exam.
• Everything discussed in the class during 4 hours during all the lectures is part of the examination.
• You will not be asked to write any R Code in the Exam.
Exam Preparation

Rule

**MUST BRING:** Estonian ID card or Resident Permit (picture and name)
You will be asked to leave in absence of identity card.
Exam structure

• One A4 cheat sheet **allowed**. You can write on both sides.
• 40 points
• ~20 questions
• Allowed: Pen, Calculator (**NO** mobile)
  • Final version of the submission paper should have pen version (not pencil)
• 3 hours
  • But probably 1.5 to 2 hours will be enough
• You will not be asked to write any R Code in the Exam.
• **MUST BRING:** Estonian ID card or Resident Permit (picture and name)
  • You will be asked to leave in absence of identity card.
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

**NOTE:** Some questions invite negative questions. Please read them carefully. When options are provided, they naturally invite negative marking.
Question 1: Simple multiple choice question

• (1 point) Mark the correct option(s) with a circle. 〇

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. 〇

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 euros
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:

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>50</td>
<td>30</td>
<td>80</td>
</tr>
<tr>
<td>Negative</td>
<td>25</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>total</td>
<td>75</td>
<td>45</td>
<td>120</td>
</tr>
</tbody>
</table>

• 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.
No such possibility
IMPORTANT

• Questions will be based on concepts / reasoning/ analytical skills.
  • Why, Why, Why ?

• **NO** correlation between number of correct options and marks. For example, If a question has 3 points, does not mean it has 3 correct options.
More doubts before Exam?

• Before you post a question on Piazza, please note following:
  • Which Lecture/Topic?
  • Which slide(s) you are referring to?
  • Have you listened to video and you could not find in it?

• For a productive discussion: please provide all the above information above when you post a question.
Exam Dates and Results

• Exam Dates:
  • First Exam: 27.05.2019, 10.15 am to 1.15 pm (3 hours), Room 404
  • Second Exam: 03.06.2019, 10.15 am to 1.15 pm (3 hours), Room 122
  • Resit: 10.06.2019, 10.15 am -- 1.15 pm (3 hours), Room 224

• Result (Exam 1 and Exam 2):
  • 04.06. 2019: (Before) 9 am.
  • If not satisfied with the marks: Consultation: 04.06: Room 119 (Timing ??)
  • Write an email before consultation (to avoid waiting)