BIG DATA ANALYTICS
Data Science Project

Olga Klopp
klopp@essec.edu

Outline

Stages of a data science project

Fixing Goals

Setting the project strategy

Model evaluation and critique

How to present results and document

Practical information

Schedule

Deliverables

Topics

Outline

Stages of a data science project

- Fixing Goals
- Setting the project strategy
- Model evaluation and critique
- How to present results and document

Practical information

- Schedule
- Deliverables
- Topics

Goals

- Clear goals
  - The bank feels that its losing too much money to bad loans and wants to reduce its losses
  - The ultimate business goal is to reduce the bank's losses due to bad loans
  - Your project sponsor envisions a tool to help loan officers more accurately score loan applicants.

Goals

The goal should be specific and measurable!

- "~~We want to get better at finding bad loans~~",
- but "We want to reduce our rate of loan charge-offs by at least 10%, using a model that predicts which loan applicants are likely to default."
  - A concrete goal ⟹ concrete stopping conditions and concrete acceptance criteria
  - The less specific the goal, the likelier that the project will go unbounded

Goals

- We can also run more exploratory projects:
  - ”Is there something in the data that correlates to higher defaults?”
  - ”Should we think about reducing the kinds of loans we give out? Which types might we eliminate?”
- You can still scope the project with concrete stopping conditions, such as a time limit
- The goal is then to come up with candidate hypotheses
- $\Longrightarrow$ concrete questions or goals for a full-scale modeling project

Setting the project strategy

- Design the project steps
- Pick the data sources
- Pick the tools to be used

Collect and manage data

- Identifying the data, exploring it, and conditioning it to be suitable for analysis:
  - What data is available to me?
  - Will it help me solve the problem?
  - Is it enough?
  - Is the data quality good enough?
- The most time-consuming step in the process
- **One of the most important!**

Loan application problem

- A sample of representative loans from the last decade
- Some of the loans have defaulted; most of them (about 70%) have not
- A variety of attributes about each loan application:
  - `Status.of.existing.checking.account` (at time of application)
  - `Duration.in.month` (loan length)
  - `Credit.history`
  - `Purpose` (car loan, student loan, etc.)
  - `Credit.amount` (loan amount)
  - `Savings.Account.or.bonds` (balance/amount)
  - `Present.employment.since`
  - `Personal.status.and.sex`
  - `Installment.rate.in.percentage.of.disposable.income` (the size of the loan payments relative to the borrowers disposable)

Loan application problem

- income
- Cosigners
- Present.residence.since
- ...
- Job (employment type)
- Number.of.dependents
- Good.Loan (dependent variable): GoodLoan was paid off, and a BadLoan defaulted

Collect and manage data

- Initial exploration and visualization of the data
- Clean the data: repair data errors and transform variables
- You may discover that the data isn't suitable for your problem, or that you need other types of information
- You may discover things in the data that raise issues more important than the one you originally planned to address...

Modeling

- You get to statistics and machine learning during the modeling stage
- Here is where you try to extract useful insights from the data

Data science modeling tasks

- **Classification** - Deciding if something belongs to one category or another
- **Scoring** - Predicting or estimating a numeric value, such as a price or probability
- **Ranking** - Learning to order items by preferences
- **Clustering** - Grouping items into most-similar groups
- **Finding relations** - Finding correlations or potential causes of effects seen in the data
- **Characterization** - Very general plotting and report generation from data

Data science modeling tasks

- For each of these tasks, there are several different possible approaches
- The loan application problem is a classification problem
- Three common approaches:
  - Logistic regression
  - Naive Bayes classifiers
  - Decision trees

Evaluate and critique model

- Does the model solve my problem?
- Do the results of the model make sense?
- Is it accurate enough for your needs?
- Does it perform better than the obvious guess? Better than whatever estimate you currently use?

Model evaluation and critique

- If you've answered "no":
  - loop back to the modeling
  - data doesn't support the goal you are trying to achieve
- You can't meet your success criteria with current resources:
  - Defining more realistic goals?
  - Gathering the additional data to achieve original goals?

Null Model: lower bound on performance

The null model represents the lower bound on model performance that you should strive for

- The null model is the "the obvious guess" that your model must outperform:
  - a working model or solution already in place
  - or the simplest possible model:
    - e.g., always guessing GoodLoan
    - e.g., always predicting the mean value
- Its error rate is called the **base error rate**

Loan application example

- 70% of the loan applications in the dataset turned out to be good loans
- A model that labels all loans as GoodLoan would be correct 70% of the time
- To be useful, any model should be better than 70% accurate

Loan application example

- Summary of classifier accuracy: confusion matrix

  Confusion matrix *tabulates actual classifications against predicted ones*

- Assume that the overall accuracy is not enough

- What kinds of mistakes are being made?
  - Is the model missing too many bad loans?
  - Is it marking too many good loans as bad?

Model Evaluation

- **Recall**: measures how many of the bad loans the model can actually find

- **Precision**: measures how many of the loans identified as bad really are bad

- **False positive rate**: measures how many of the good loans are mistakenly identified as bad

- **You want the recall and the precision to be high, and the false positive rate to be low**

- The right balance: trade-off between recall and precision.

How to present results and document

- Different audiences require different kinds of information
- Business-oriented audiences want to understand the impact of your findings in terms of business metrics
- Give this audience your most interesting findings or recommendations.

Loan application example

- Business audiences: how your model will reduce the money that the bank loses to bad loans
  - Your model identified a set of bad loans that amounted to 22% of the total money lost to defaults
  - Your presentation should emphasize that the model can potentially reduce the banks losses by 22%
- Interesting findings or recommendations:
  - new car loans are much riskier than used car loans
  - most losses are tied to bad car loans and bad equipment loans

Outline

Stages of a data science project

- Fixing Goals
- Setting the project strategy
- Model evaluation and critique
- How to present results and document

Practical information

- Schedule
- Deliverables
- Topics

Schedule

- Teams of **4 people**
- There are two main deliverables for the project:
  - Proposal 5%
    November 2 2018
  - Final report + Presentation + CEO 45%
    December 4th 2018

Project Proposal 1/2

- Describe your project:
  - Define the problem you propose to solve
  - How do you plan to achieve it
- One should easily identify the following:
  - Data set(s)
  - Technique(s)
  - Goal(s)
- Formatting and Page Limits:
  - PDF 2 pages
  - Include the title of your project and the names of the members of the team.
  - Add ”Project Proposal” as subtitle

Project Proposal 2/2

- Motivation and Problem Definition
- Dataset
- Methodology
- Evaluation
- References

Final report 1/2

- Formatting and Page Limits:
  - 7-8 pages
  - Include in the header of the report the title of your project and the names of the members of the team
  - Also, indicate that this is the final report

  in PDF format
- Introduction/Motivation: show the need that motivated this project
- State the project goal
- Previous efforts on this problems:
  - What did they do?
  - Why their approaches may or may not work for your problem
  - References

Final report 1/2

- Describe how the project was run:
  - Introduce the input variables (and issues with them)
  - Introduce the model, why you chose it, and issues with it
- Show your results: model performance and other outcomes
- Discuss other key findings, like which variables were most influential on the model
- Limitations of your results
- Listing some improvements and follow-ups that you would like to make
- Include Html (or similar) file with your code and comments.

Presentation

- December 4th or December 11th
- 15 min by team (including discussion)
- Include
  - Motivation and Data Set
  - Goal(s)
  - Data Preparation and DataViz
  - Model. Why do you think it is a good choice?
  - Results: model performance and other outcomes

Data Set 1/2

- Possible topics:
  - Creating a new data set (if important data preparation part such as missing values, outliers etc) + DataViz
  - Study effect of missing mechanism
  - Linear Regression and GAM (on massive data sets)
  - Linear Model selection and Regularization
  - Resampling Methods, Bootstrap
- **Classification projects with none of the previous one are not allowed** (ML cours)

Data Set

- Many data sets available
- **Please be original!**
    - Data sets from science article
    - Census data from most nations are open
    - FRED from the Federal Reserve
    - NASA open source data
    - ...

GOOD LUCK!
YOU