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
- Data Set
- Final report

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
Data Set
Final report

Goals

- Clear goals (usually given by the project sponsor)
  - Suppose you're working for a bank. 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, and so reduce the number of bad loans made.

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 begets concrete stopping conditions and concrete acceptance criteria
  - The less specific the goal, the likelier that the project will go unbounded, because no result will be good enough.

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
- These hypotheses can then be turned into 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 you need, 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?
- This stage is often 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 (excluding home loans)
- 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)

In this data, Good.Loan takes on two possible values: GoodLoan and BadLoan. Assume that a 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, as needed
- You may discover that the data isn't suitable for your problem, or that you need other types of information as well
- 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, or analysis, stage
- Here is where you try to extract useful insights from the data in order to achieve your goals.

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
- We'll cover some of the most common approaches to the different tasks in this course
- Other approaches: Machine Learning course

Data science modeling tasks

- For each of these tasks, there are several different possible approaches
- We'll cover some of the most common approaches to the different tasks in this course
- Other approaches: Machine Learning course
- The loan application problem is a classification problem: you want to identify loan applicants who are likely to default. Three common approaches:
  - Logistic regression
  - Naive Bayes classifiers
  - Decision trees

Evaluate and critique model

Once you have a model, you need to determine if it meets your goals:

- Does the model solve my problem?
- Do the results of the model (coefficients, clusters, rules) make sense in the context of the problem domain?
- 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", it's time to loop back to the modeling step - or decide that the data doesn't support the goal you are trying to achieve
- Maybe you will understand that you can't meet your success criteria with current resources:
  - Defining more realistic goals?
  - Gathering the additional data or other resources that you need to achieve your original goals?

Null Model: lower bound on performance

- The null model is the "the obvious guess" that your model must outperform:
  - if there is a working model or solution already in place that you're trying to improve, the null model is the existing solution
  - In situations where there's no existing model or solution, the null model is the simplest possible model:
    - always guessing GoodLoan
    - or always predicting the mean value of the output, when you're trying to predict a numerical value
- Since the null model is the simplest possible model, its error rate is called the base error rate

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

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, using only the existing process to classify loans would be correct 70% of the time
- Any actual model that you fit to the data should be better than 70% accurate to be useful

Loan application example

- A good 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
- **Ideally, you want the recall and the precision to be high, and the false positive rate to be low**
- Often, the right balance requires some trade-off between recall and precision.

How to present results and document

Once you have a model that meets your success criteria, you'll present your results:

- Different audiences require different kinds of information:
  - Business-oriented audiences want to understand the impact of your findings in terms of business metrics

Loan application example

- The most important thing to present to business audiences is how your model will reduce the money that the bank loses to bad loans:
  - Suppose your model identified a set of bad loans that amounted to 22% of the total money lost to defaults
  - Your presentation or executive summary should emphasize that the model can potentially reduce the banks losses by that amount
- 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

Schedule

- Create groups of 4 people. **Deadline: next week**
- Choose the data set
- Short presentation:
  - Data Set
  - Goal(s)
  - DataViz
  - Model. Why do you think it is a good choice?

  November 13th 2017
- 2 deliverables:
  - A technical report
  - A CEO (Chief Executive Officer) summary (2 pages)

  Deadline: December 30th 2017

Data Set 1/2

A regression data set among:

- Airfoil Self-Noise
- Bike Sharing Dataset
- Combined Cycle Power Plant
- Communities and Crime
- Forest Fires
- Housing
- Individual household electric power consumption
- Parkinsons Telemonitoring
- Student Performance
- Wine Quality

on the website: **http://archive.ics.uci.edu/ml/**

Data Set 2/2

Other possibilities:

- A classification data set on the website:
  http://archive.ics.uci.edu/ml/
- Another data set on the website:
  http://www.stat.ufl.edu/ winner/datasets.html
- Another data set you could propose

Final report 1/2

Your final report should contain:

- Modeling task: show the need that motivated this project
- State the project goal (e.g. in terms of the business motivation)
- You may also discuss previous efforts on this problems:
  - What did they do?
  - Why their approaches may or may not work for your problem
  - Cite who did the work, and where you found out about it

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.