STK-INF4000 - Mid Term Project

Evaluation Checklist

Version 0.1, March 21st 2017

Business Idea

- Project should have a precise application in mind.
  - *Bad*: "We want to sell books."
    - What kind? To whom? Why?
  - *Good*: "We want to help aid organizations predict shortages of vaccine supply."
- Idea should have a selling point.
  - Why is it important?
  - What's better about your strategy than the usual ones?
  - Is it a new idea?
    - No need to come up with something new, but one should research if there are existing solutions.
- Clear audience. Be specific.
  - Transportation company: Not good enough.
  - Bus companies in smaller towns: Better.
  - Marketing department of bus companies in smaller towns: Excellent!

Data Ingestion

- Some slightly advanced techniques should be applied.
  - Do **not** simply read a `.csv` file.
  - If your data is in `.csv`, combine it with other data sets.
- Likewise, make sure you **do not** do the *obvious* analysis.
  - Don't download/get you hands on e.g. the wine quality data set and predict the quality. *The aim is to be creative.*
  - Explanation: There are lots of purpose-built data sets e.g. "predicting airplane delays". Just downloading this data set and performing the analysis the data was collected for is *not good enough.*
- Why do you think the data will help to solve the business problem?
  - How can you prove this?
- Due diligence.

  - Are there anomalies?
    * Why?
  - Is there missing data?
    * Why?
  - Are the inputs categorical?
    * If so, did you use one-hot-encoding?
    * Is there one suspiciously large or small class?
      · Does it still make sense to use that variable?

Modeling

- How can machine learning help solving the business problem?
- How can you show this?
  - Show it. Make plots. Use colors.
- Do you go for high prediction accuracy or interpretable models?
- Find a good way/ways to present the models you've made.
  - Visualization of a decision tree.
  - Plot of error vs. model complexity.
  - Plot of predictor importance and uncertainties.
- How do you prevent over-fitting?

Deliverables

- Computer code.
  - Enough code to prove you did all the steps from data ingestion to modeling, including analysis of data quality (c.f. 'Due diligence' above).
  - Format: Notebooks are fine. So are `git` repositories with `README`s.
- Presentation.
  - The presentation given should be delivered in electronic form (`PPT` is okay, `PDF` even better).