Group Assignment 2017

Machine Learning II

Master in Business Analytics and Big Data

Jesús Renero (jrenero@faculty.ie.edu)

DRIVENDATA

“At DrivenData, we want to bring cutting-edge practices in data science and crowdsourcing to some of the world's biggest social challenges and the organizations taking them on. We host online challenges, usually lasting 2-3 months, where a global community of data scientists competes to come up with the best statistical model for difficult predictive problems that make a difference.”

Pump it Up: Data Mining the Water Table

Can you predict which water pumps are faulty?

Using data from Taarifa and the Tanzanian Ministry of Water, can you predict which pumps are functional, which need some repairs, and which don't work at all?

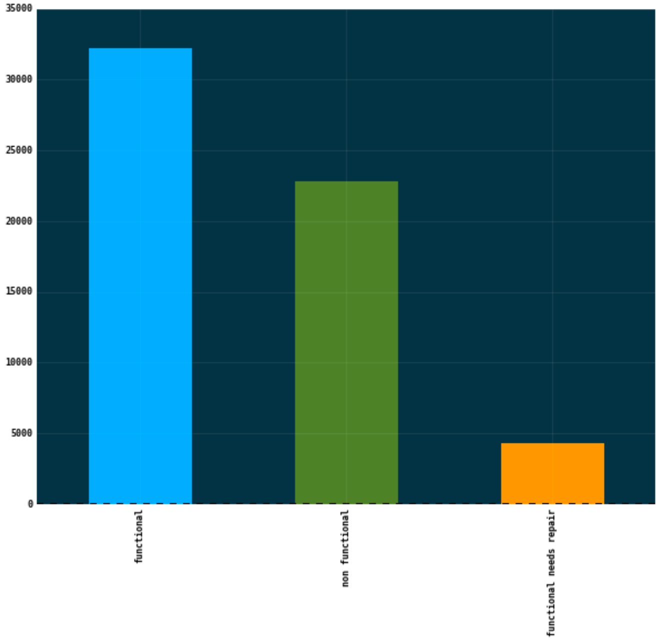
This is an intermediate-level practice competition. Predict one of these three classes based on a number of variables about what kind of pump is operating, when it was installed, and how it is managed. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Resources

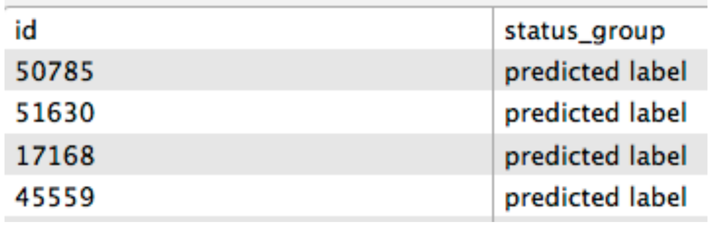
- An interactive course exploring this dataset is currently offered by DataCamp.com
- There’s an online discussion at DrivenData, if you register, where you can learn from what other participants are trying.
- The ongoing competition will end by Sept. 28, 2017, 11:59 p.m.
- The winner team in the internal competition will register its solution to the DrivenData competition.

Labels

- The labels in this dataset are simple.
- There are three possible values:
  - **functional** - the waterpoint is operational and there are no repairs needed
  - **functional needs repair** - the waterpoint is operational, but needs repairs
  - **non functional** - the waterpoint is not operational



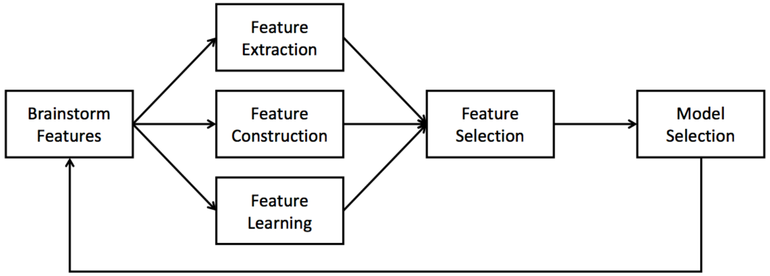
Code preparation



id	status_group
50785	predicted label
51630	predicted label
17168	predicted label
45559	predicted label

- Your code will include a function called ‘`evaluate(filePath)`’
  - The output of this function will be a dataframe with two columns: ‘id’ and ‘status_group’ (see fig. above), with the classification result for the entries in the test file passed as argument (filePath).
  - This function is the only entry point that will be used to produce the final score.
- Classification of entries in the final evaluation test will be made using your model:
  - A **confussion matrix** will be built, and **classification accuracy** will be used (how many correct classifications are produced, divided by the total number of samples in the test set), to determine the winner.

Rules



- Only the methods covered during 1st and 2nd quarters are allowed.
  - In case you decide to use SVM, only the linear version is allowed (no kernel).
- Submission file (`MLO`$_{1/2}$`GroupID.zip`):
  1. PDF file (No HTML) with the code chunks and output that explain the whole machine learning process (see fig. above).
     - IMPORTANT: No log/console messages output. Only relevant output is allowed. Spurious output, not working Rmd files or unreasonably large PDF files without relevant information on how the problem is solved will have a penalty of 1/10 points.
  2. The `.Rmd` source Notebook file.
- Scores will be assigned considering
  - 50% - from the rank in the competition
  - 50% - form the overall quality of the document and the approach.

Evaluation

- The models submitted will be evaluated agains an additional **portion of the data that has not been shared with you**.
  - Your model is **blind** with respect to that dataset, so you will have to build the model which is producing the best generalization possible.
- The final evaluation test sample contains ~ 6000 entries and labels.

Deadlines

- **O1** group deadline
  - **March 30th 2017, 11:59 pm**
- **O2** group deadline
  - **March 28th 2017, 11:59 pm**