Pump it Up!

DATA MINING THE WATER TABLE

Pump it Up!

DATA MINING THE WATER TABLE

I own none of the images in this powerpoint. Don't sue.

Pumps fail

- The goal of the project was to predict which water pumps would are faulty or were on track to fail.
- Data came from Taarifa and the Tanzanian Ministry of Water.
- Part of a data science competition through Driven Data.

Personal Background

- Economic and Environmental development work for the Clinton Foundation
- Dual Masters graduate in environmental science and environmental and public policy analysis from SPEA at Indiana University
- Currently an environmental subject matter expert for the Next Gen Strategic Innovation Group at Booz Allen Hamilton
- Volunteer with the Climate Science Legal Defense Fund
- Areas of expertise: Climate Change Policy and Energy Issues, Applied Ecology, and Water Resources

So what sort of problem was this?

- This was a classification problem
- There were three possible outcomes: functional, non functional, and needs repair
- The decision was made to use random forest via scikit learn to develop the classification model.

Working with the data

- The dataset made available for training had 40 columns and over 50000 rows.
- Lots of data was missing, and there was even column that lacked any kind of description.
- The biggest challenge was cleaning the data
- I removed seven rows that I felt as though would have limited utility in developing a predictive algorithm.
- 3000 rows were missing categorical data from important columns, and these rows were dropped.

Preparing the data for random forest

- Column removal was done in excel as opposed to inside of python.
  - Dataset was small enough to easily navigate in excel
  - This was just faster
- Label Encoder was used to change categorical variables into numbers.
- Instead of writing many lines of code for the categorical variables, I wrote a script to run label encoder on multiple columns.
- The training data was split 60/40 for an initial training and test set.

Random Forest Model

- Initial predictive score for training set was 93% and test set was 80%
- Grid search was performed to improve the model.
- Not much was gained.

- The model was applied to the submission set.
- Odd ball problem occurred: couldn't get label encoder to do the inverse transform on the random forest output.
  - Pulled the output into excel and wrote an IF loop to translate the output into submission document friendly variables.

Result:

Submissions

BEST SCORE	CURRENT RANK	# COMPETITORS	SUBS. TODAY
0.7817	283	1796	1 / 3

Contact Information

- roundce@gmail.com
- Round_Christopher@bah.com