API 222 Prediction Competition

Machine Learning and Big Data Analytics

Due Before 11:45 am on April 11, 2018

Under the Kennedy School Academic Code, this assignment is a Type II assignment. You are encouraged to work in a study group, but must submit your own hand- or type-written solutions. It is not acceptable to work on one electronic document as a group and submit identical, or nearly identical versions.

1 Overview

The following is a real problem posted as a competition on DrivenData. It represents a real way in which policy makers are leveraging the tools of machine learning to gain insights to improve lives.

Access to clean drinking water is vital. A smart understanding of which water points will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania. You will use data from Taarifa and the Tanzanian Ministry of Water to predict the state of water sources. There are three possible classifications: functional - the water point is operational and there are no repairs needed, functional needs repair - the water point is operational, but needs repairs and non functional - the water point is not operational. Beyond Tanzania, many countries face this same problem of maintaining clean water points. Models predicting which water points are likely to be functional can help influence future maintenance as well as help determine areas that may be particularly suited for water pumps. If you can generate an improvement, the new algorithm could be implemented in other countries around the world.

2 Assignment

Your objective is to develop a model that accurately predicts whether the water pump is functional, semi-functional or non-functional. Your score will be determined using the macro F_1 Score on a holdout set of data. Professor Saghafian will announce in class the first, second, and third place students who have been able to achieve the highest macro F_1 scores using a Machine Learning algorithm.

The macro F_1 score is an aggregate measure for classification errors. It calculates three F_1 scores (because the target can take three unique values) and then averages them. For example, the first F_1 score for functional water points, will treat functional as positives and functional needs repair and non-functional collectively as negatives. The F_1 score for a binary classification problem is defined as:

$$F_1 = \frac{2 \cdot \text{ true positives}}{2 \cdot \text{ true positives} + \text{ false negatives } + \text{ false positives}} \tag{1}$$

Unlike with the online version, your score will also be determined by an accompanying write up that clearly explains the process you went through of choosing a model and describes your final approach.

You will be required to submit working code (we must be able to run it by changing only one file path line), a CSV file of predictions for a hold out set of data that will be released 48 hours before the submission deadline, and a write-up between 1.5 and 2.5 pages single spaced, size 12 font Times New Roman, 1 inch margins. **Your write-up should be geared toward a member of the Tanzania Ministry of Water staff who has some familiarity with Machine Learning but who is not an expert.** The goal of the written portion of this assignment is to get you familiar with explaining the process of model selection to a broad audience in a clear way. This will be an important skill in facilitating the adoption of high-performing yet new or unfamiliar methods in the types of organizations where many of you will work after graduation.

3 Prediction Competition Rules

You may not use any data other than the data provided by the course instructors in developing your model. **Anyone who uses any additional data will receive zero credit for the assignment and faces possible disciplinary charges.** However, you may do whatever you like with the data provided, such as generating new features through interactions, non-linear transformations, etc.

4 Grading

This competition is worth 15% of your course grade. Therefore, the assignment will be worth 15 points, which will be broken down into three evenly weighted components (e.g. 5 points each):

1. The write-up, which has
   (a) A thorough description of the process you took to arrive at your final model
   (b) A clear description of your final model, including any data manipulation or feature engineering
   (c) A discussion of your approach as it pertains to algorithmic bias and transparency. This section should contain some numbers illustrating how your model performance varies along salient characteristics, such as demographic and geographic characteristics. It should also concretely discuss the trade-offs your model makes between predictive performance and interpretability / transparency.
2. Clean code that:
   (a) Trains your model
   (b) Produces a CSV of predictions for the holdout data

   The teaching staff must be able to successfully run the code by changing only one line of the file path. **Code that we cannot run without further edits will receive at most 1 of the 5 possible points.**
3. A CSV file of predictions. We will order students in terms of predictive accuracy on the holdout data.
   (a) Students in the top one-fifth of the class on this measure will receive 5 out of 5 points on this component.
   (b) Students in the second-to-top one-fifth will receive 4 out of 5 points on this component.
   ⋮
   (c) Students in the bottom fifth will receive 1 out of 5 points on this component.

   We will provide you with a sample submission CSV, which will have two columns:
   (a) `Id` - An ID column that maps to the holdout data released 48 hours before the submission deadline
   (b) `Prediction` - A column that contains your predicted functionality for each water point in the holdout sample

   You should submit a CSV file with the same two columns, and those columns should be named `Id` and `Prediction`. The filename should be `lastname_prediction.csv`, where you replace `lastname` with your actual last name (for example Prof. Saghafian's file would be `saghafian_prediction.csv`). **Any submissions that do not include these two columns (with the correct column names) or have the wrong file name will receive zero points on this section.**

5 Data Description

Number of Observations	54,619
Number of Predictors	40
ID Variables	`id`
Response Variable	`status_group`

5.1 Core Data Fields

- `id` - a unique identifier for each row.
- `status_group` - the status group is an ordinal variable indicating functionality of the water point
  - `functional` - the water point is operational and there are no repairs needed
  - `functional needs repair` - the water point is operational, but needs repairs
  - `non functional` - the water point is not operational

5.2 All Data fields

Variable name - Variable description

```
amount_tsh - Total static head (amount water available to water point)
year_recorded - The year the row was entered
lat_X - Xth percentile of latitude of GPS coordinate
long_X - Xth percentile of longitude of GPS coordinate
p_X -Xth percentile of height of GPS coordinate
basin - Geographic water basin
region - Geographic location
region_code - Geographic location (coded)
district_code - Geographic location (coded)
lga - Geographic location
ward - Geographic location
population - Population around the well
public_meeting - True/False
recorded_by - Group entering this row of data
VWC_management - VWC operates the water point
permit - If the water point is permitted
construction_year - Year the water point was constructed
extraction_type - The kind of extraction the water point uses
extraction_typegroup - The kind of extraction the water point uses
extraction_type_class - The kind of extraction the water point uses
payment - What the water costs
payment_type - What the water costs
water_quality - The quality of the water
quality_group - The quality of the water
quantity - The quantity of water
quantity_group - The quantity of water
source - The source of the water
source_type - The source of the water
source_class - The source of the water
waterpoint_type - The kind of water point
waterpoint_type_group - The kind of water point group
```