Module 3: Final Project

Water Pump Operational Status Classification

Competitive Data Science Project: Pump it Up: Data Mining the Water Table, hosted by DRIVENDATA

by Steven Contreras

The problem

Context

“Taarifa is an open source platform for the crowd-sourced reporting and triaging of infrastructure related issues. Think of it as a bug tracker for the real world which helps to engage citizens with their local government. We are currently working on an Innovation Project in Tanzania, with various partners.”

- Taarifa

The particular goal is to bring filtered-water, sourced and pumped from local water-sources, to disparate and impoverished geographic locations.

Problem statement

This is a multi-class classification machine learning problem.

The goal is to predict the operating condition of a waterpoint for each record in the dataset, which can be one of the three operational status classes: “*functional*”, “*functional needs repair*”, and “*non functional*”.

Solution

Ensemble (`VotingClassifier`) of classification algorithms:
`RandomForestClassifier` and `CatboostClassifier`

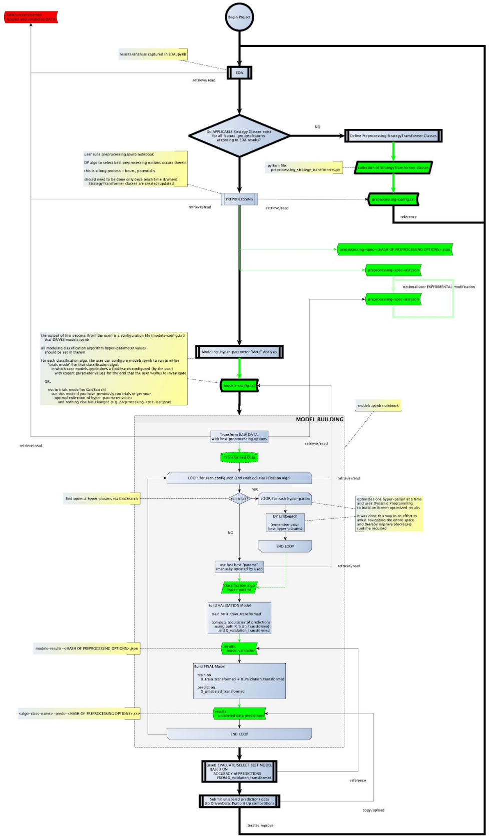
Competition Leaderboard:

- *Model Accuracy*: **81.87%**
- *Competition Rank*: **837/9751**
- *Competition Percentile*: **> 91st**

Overview

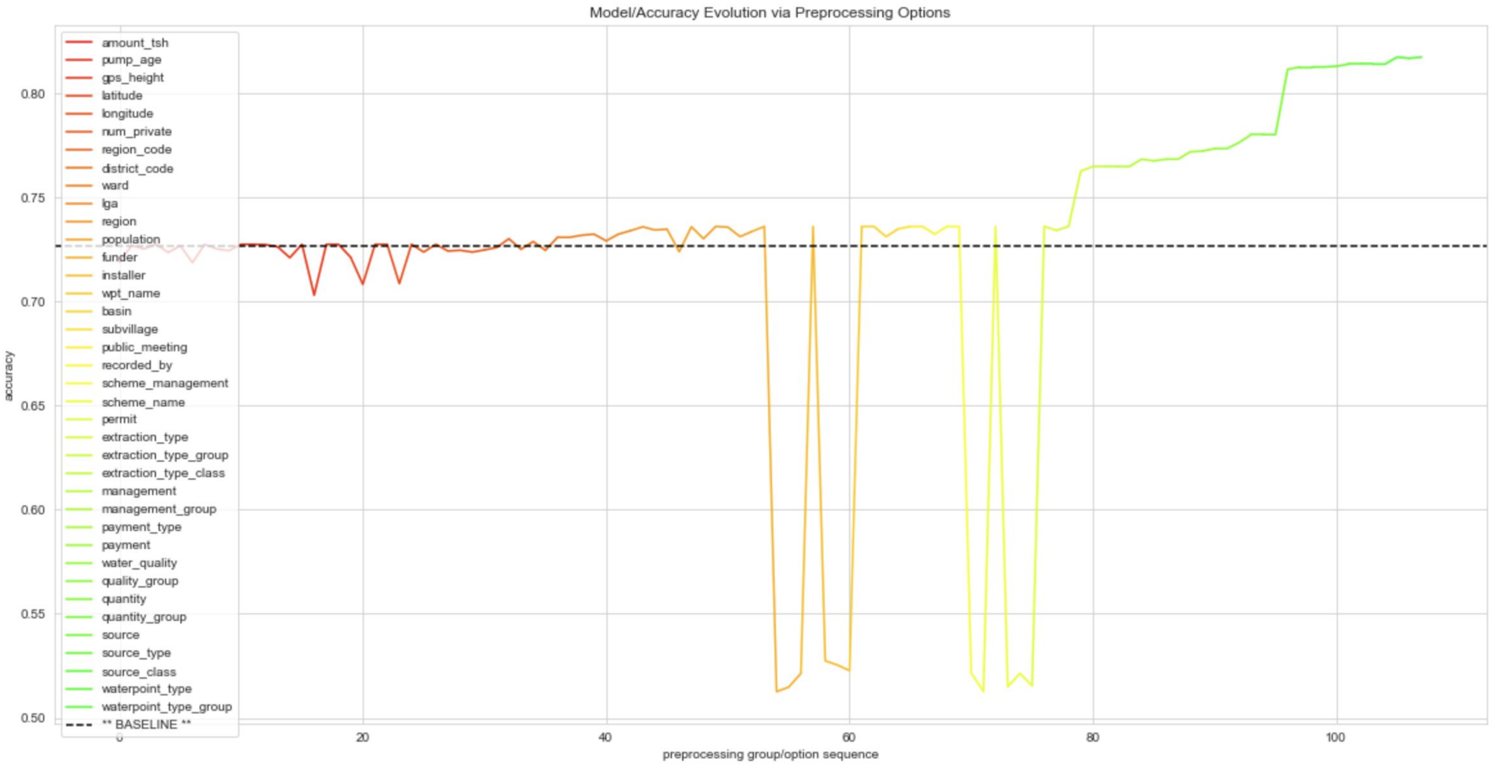
- **Features**
  - 38 starting features (predictors), not including *id*
  - Feature groups and data "overlap" - see feature descriptions at https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25
  - Dropped 21 features not contributing to classification, 17 features contributed to classification
- **Approach**
  - Very thorough EDA time-investment
  - Greedy Algorithms: feature selection and transformation-option selection, hyper-parameter tuning
  - Development/use of ensembling techniques provided best accuracy
- **Links**:
  - Github Repo: https://github.com/sacontreras/dsc-mod-3-project-v2-1-online-ds-sp-000
  - Blog: https://sacontreras.github.io/my_first_data_science_competition

Workflow: Adapted from OSEMN

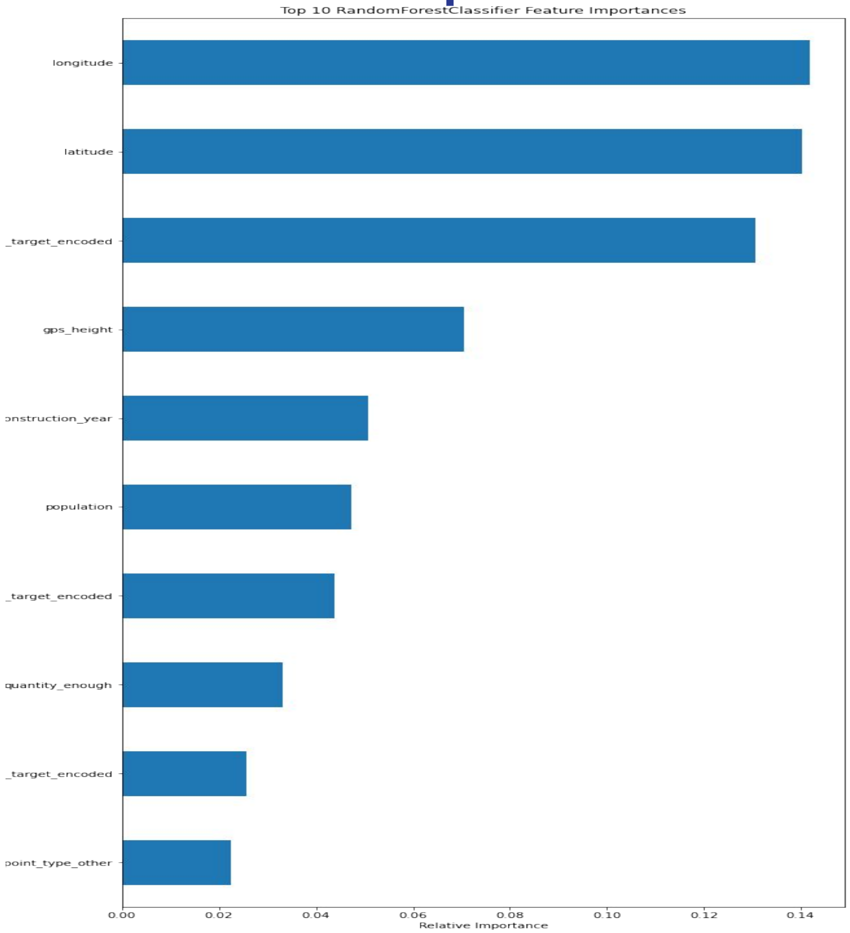


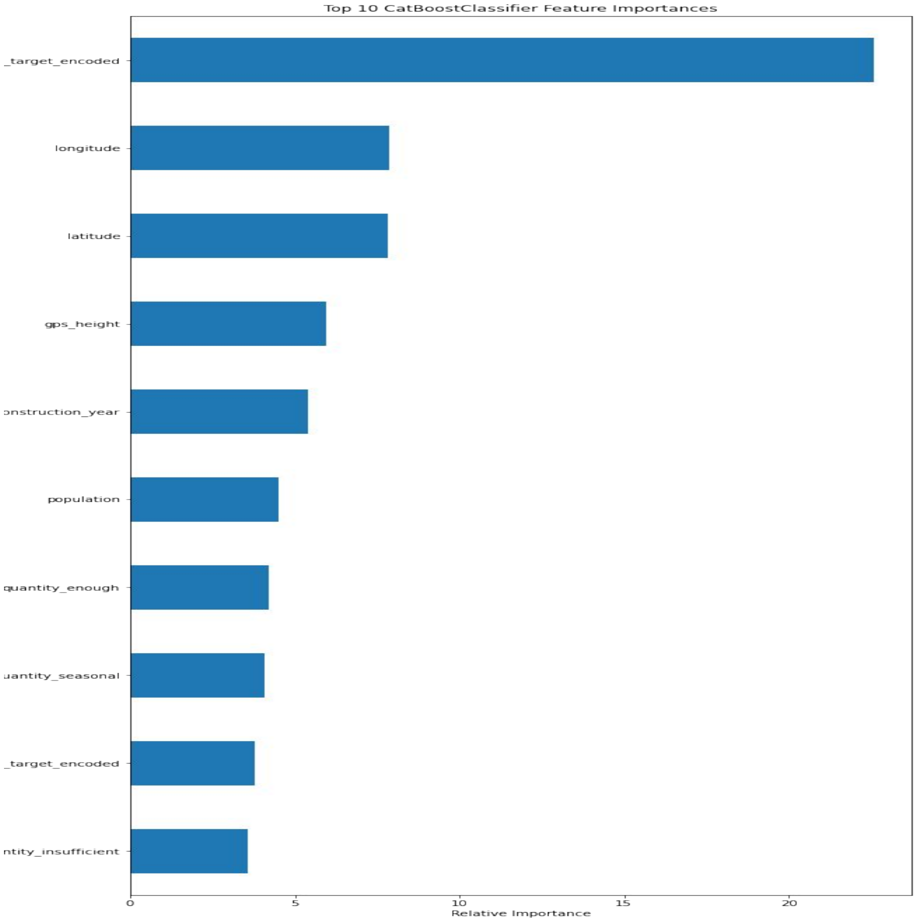
View the diagram in your browser here.

Preprocessing (Optimization)



Feature Importances





Interpretation and Recommendations

1. Geographic Location (*ward, gps-coordinates*) features are the most important predictors: focus routine maintenance in locations with current greatest occurrence of *functional needs repair* status in order to preempt (avoid) this status in the future. The same rationale applies to those pumps with greater relative *population* as well as relatively older pumps.
2. For pumps/wells with greater *gps_height* values and with status *non functional* or *functional needs repair* statuses, consider refitting with more powerful/robust pump-motor installation (and related components).
3. Water *quantity* and *amount_tsh* ("Total static head (amount water available to waterpoint)") - i.e. **flow** - is a factor; for those pumps falling into the *quantity_insufficient* or quantity_seasonal categories, consider reducing power to the pump as a response to lower flow in order to address *functional needs repair* status or prevent *non functional* status.

Conclusion and Future Work Consideration

Thank you so much for your time!

I really enjoyed this project and I learned MANY new and powerful techniques.

Possible Future Work:

- **Preprocessing algorithm identified missing preprocessing options - low-cardinality categoricals below threshold should be one-hot encoded (vs. current anomalous target encoding)... correct this for further increase to leadboard accuracy**
- **More advanced ensemble techniques:**
  - **Stacking**
  - **Blending**
- **Use a different classification algorithm (e.g. `CatboostClassifier`) for baselining within the preprocessing.ipynb notebook for comparison**
- **Experiment more with `RandomizedSearchCV` for better resolution of hyper-parameter tuning candidate values with, for example, `RandomForestClassifier` which may lead to even higher competition accuracy**

Supplemental: The Future is Now!

The result of switching low-cardinality categoricals below threshold to OneHot Encoding (previously Target Encoded):

Competition Leaderboard:

- *Model Accuracy*: **81.99%**
- *Competition Rank*: **719/9857**
- *Competition Percentile*: **Top 8%**