Ensemble Learning Report

Authored by Kshitiz Gupta.

Contents

Using Ensemble ML Models in Predicting WaterPoint failures using Sensor Data

- The Driven Data Competition demonstrated how effective survey data (collected by Taarifa) can be in predicting waterpoint failues but other data like pump's 3-axis acceleration and pump's basin gauge pressure, collected through sensors can also prove to be powerful predictors. To better understand the potential of this sensor data in improving ML models' predictions we can look towards the PLOS One white paper: *Ensemble machine learning and forecasting can achieve 99% uptime for rural handpumps*[1].
- The aim of this project was to analyze a group of 42 Afridev-brand handpumps (maintained by The Water Project) in western Kenya and create a data-adaptive system capable of predicting failure well enough in advance to allow preventive maintenance, repair, or replacement.
- The model to identify pump failures was very simple: uncharacteristically long gaps in pump use were used as a pump failure heuristic, and service was dispatched accordingly. This pump maintenance model was termed the **"ambulance model"** and resulted in an average of **91%** of the pump fleet functioning at any given time compared to just 56% of pumps functioning under the industry-standard service model[2].
- To boost up time of the pumps from **91%** to **>99%**, the authors of the project developed a new preventative maintenance framework that services pumps as soon as, or ideally before, they fail. This framework called condition-based maintenance helps agencies identify pumps at high risk for failure and service them as-needed and "just-in-time" before they break. The most important advantage of condition-based maintenance is the ability to allocate limited maintenance resources where they are needed instead of spreading maintenance resources evenly, including where they may not be needed.
- It is important to point out that the authors of the study conservatively assumed that a preventative maintenance service visit only prevents failures that would have happened within one week of the service visit.
- The Water Project's records which indicate when a pump site was visited, what was wrong, and when the pump was repaired provided the Ground Truth Data. This Ground Truth Data combined with Sensor Data was used to classify a training set spanning the entire duration of the sensor-observed dataset.
- After Feature Extraction, the features were broke down into four categories: features based on the number of pumping events per day, the pump's flow rate, the duration of pumping events, and the ratio of a pump's flow rate to amount of handle motion (i.e. volume of water per human effort).

- Further analysis of the features found that there is great potential value in features that rate historically-relative properties of a pump rather than just the properties themselves.
- ML algorithms used in the final model included supervised ensemble machine learning tool, Super Learner [3], for predicting pump failures. Super Learner employs an ensemble of robust machine learning classification techniques, using cross-validation methods to tune model parameters. A number of other machine learning algorithms including simple regression models, Support Vector Machines, Multivariate Adaptive Regression Splines and Random Forests were also used. The final output of the machine learning model is a thresholded binary outcome: **to dispatch a service person or not**.

Additional Features

In addition to the important features discovered in Driven Data Compeition and PLOS One paper there are some more important features that some WASH research and non-profits have recommended looking into:

- To better track how functional the water source is over time, mWater[3] suggests that organizations develop separate indicators for **access** (such as the round trip time to collect water) and **reliability** (frequency of outages or breakdowns).
- Rossa O'Keeffe-O'Donovan's research[4] finds that it is useful to add new feature **'Nearby pumps similar'** which indicates whether majority of other pumps within a certain mile radius are of the same type or not. He also suggests that Standardization of pump technologies might be something worth taking a look into.

Conclusion

The final findings reinforce views that a multifaceted range of conditions is critical for the sustainability of community-managed hand pumps and that Governments and development partners must significantly strengthen post-construction support for operation and maintenance systems, and greater efforts are needed to test and evaluate alternative models for managing hand pump water supplies. Finally, the successes of Driven Data competition and PLOS One Project indicate that a ML model that integrates survey data with satellite and sensor data has great potential to robustly predict waterpoint failures before they occur.

References

1. Ensemble machine learning and forecasting can achieve 99% uptime for rural hand pumps Wilson DL, Coyle JR, Thomas EA (2017) Ensemble machine learning and forecasting can achieve 99% uptime for rural handpumps. PLOS ONE 12(11): e0188808. https://doi.org/10.1371/journal.pone.0188808
2. Nagel C, Beach J, Iribagiza C, Thomas EA. Evaluating Cellular Instrumentation on Rural Handpumps to Improve Service Delivery—A Longitudinal Study in Rural Rwanda. Environmental Science & Technology. 2015;49(24):14292–14300.
3. https://medium.com/mwater-technology-for-water-and-health/sharing-200-000-water-points-how-we-did-it-e11e9a9957c
4. https://www.ircwash.org/blog/waterspillovers-and-free-riding-economics-pump-functionality-tanzania