Pump it Up: Data Mining the Water Table

Peder Norr
May 23, 2021

Outline

- Business Problem
- Data
- Methods
- Findings
- Results
- Conclusions

Business Problem

- Tanzanian Ministry of Water wants to improve water pump maintenance operations
- Needs a way to better predict functionality status of water pumps
- Needs to determine what characteristics might indicate a non functional pump in the future

Data

- Data sourced from Taarifa and DrivenData competition site
- Dataset contains 41 variables describing pump functionality status (the target variable), pump geographic location, what kind of pump is operating, when it was installed, how it is managed, etc.
- Dataset encompassed 59,400 pumps from 2011-2013

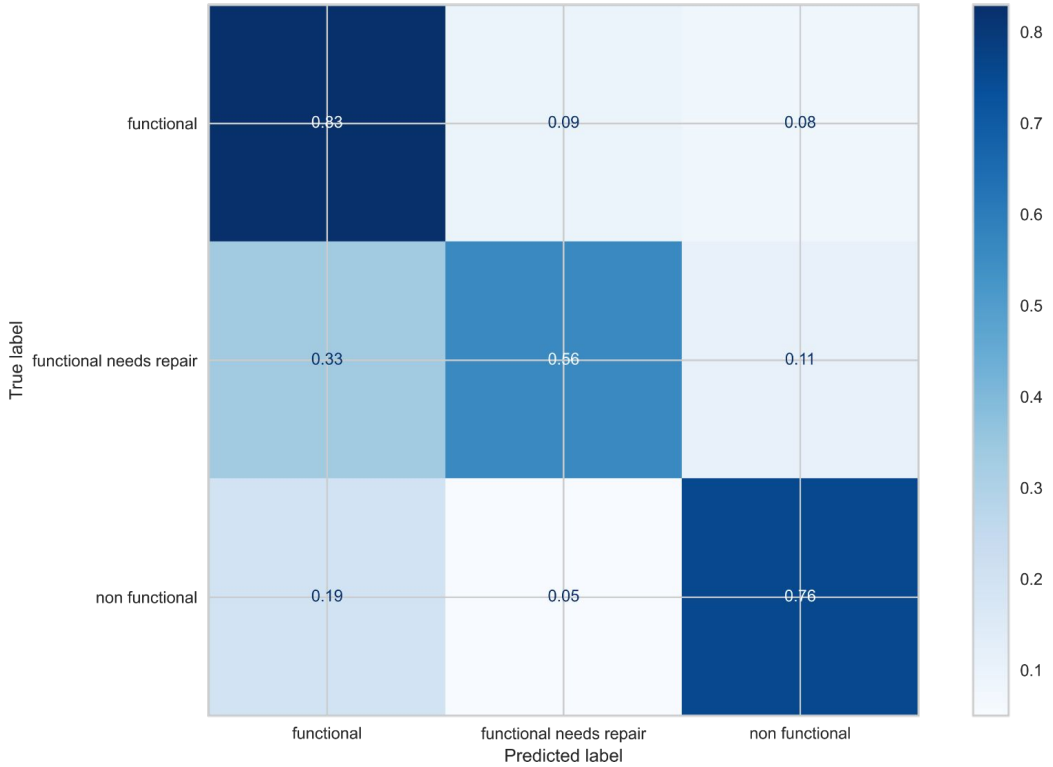
DRIVENDATA

Methods

- Created RandomForest classifier model and XGBoost classifier model
- Resulting RandomForest model had an overall accuracy of 78%, meaning it could accurately predict the status of a given pump 78% of the time
- Resulting XGBoost model had an overall accuracy of 75%, meaning it could accurately predict the status of a given pump 75% of the time

Methods

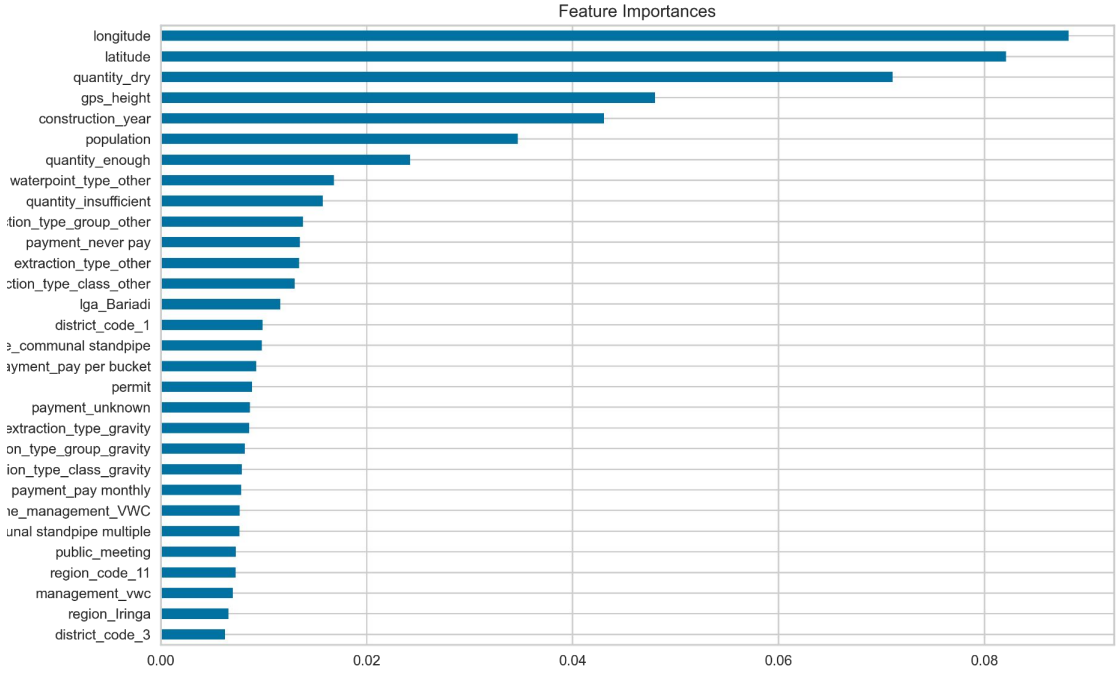
- Figure shows the accuracy of the RandomForest model for predicting each status group
- Correctly predicts functional pumps 83% of the time
- Correctly predicts needs repair pumps 56% of the time
- Correctly predicts non functional pumps 76% of the time



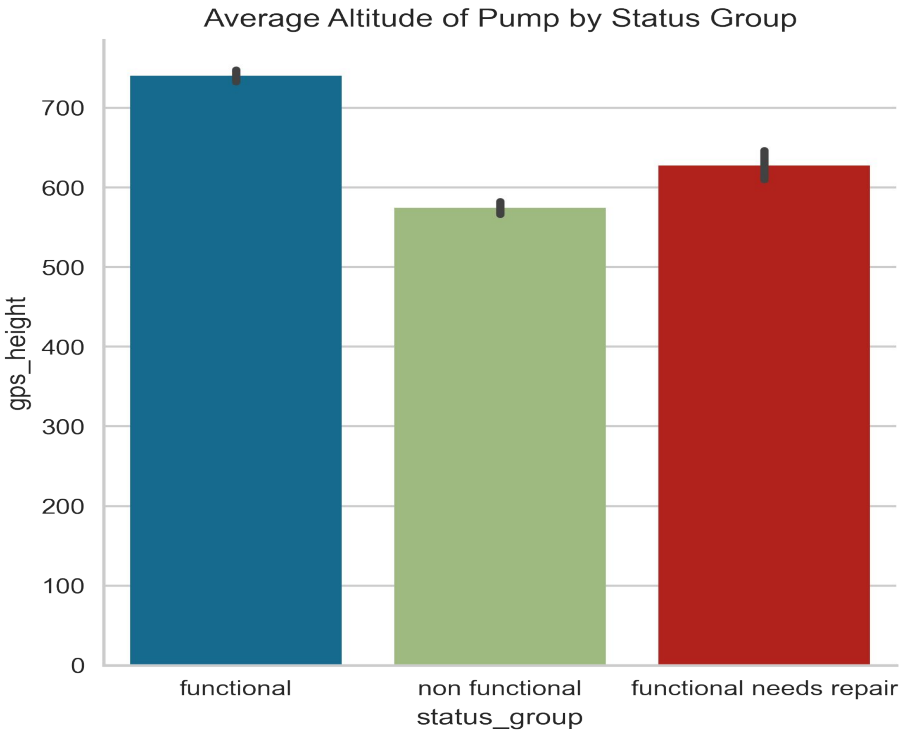
Findings

Random forest classifier model analysis of Tanzanian water pump data identifies several characteristics that are most important in identifying pump status:

- Pump location
- Pump water quantity
- Population surrounding pump
- Age of pump

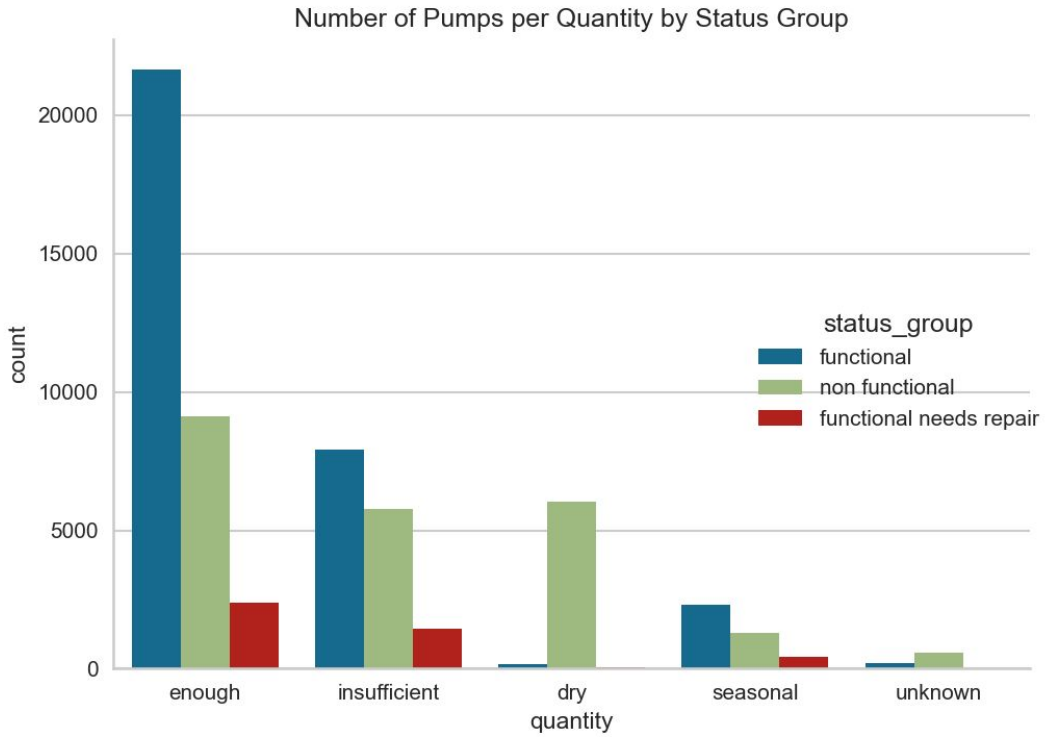


Results



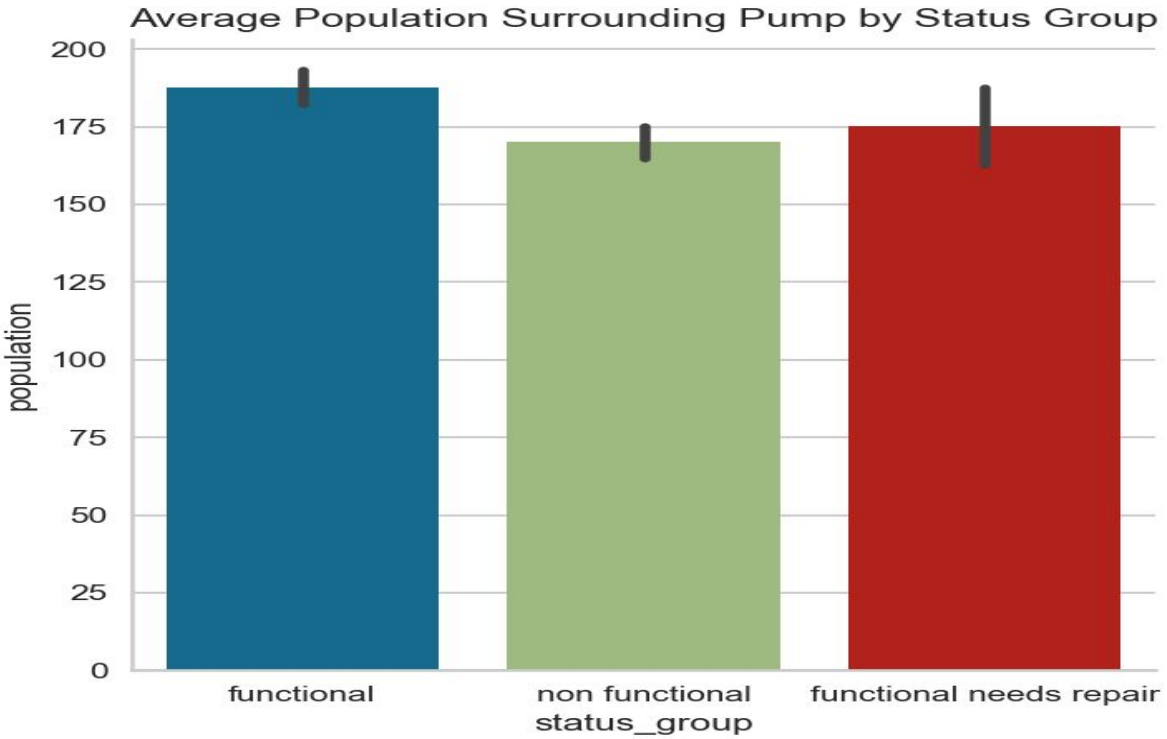
- The figure shows that on average, pumps at lower altitudes are more likely non functional or needing repair

Results



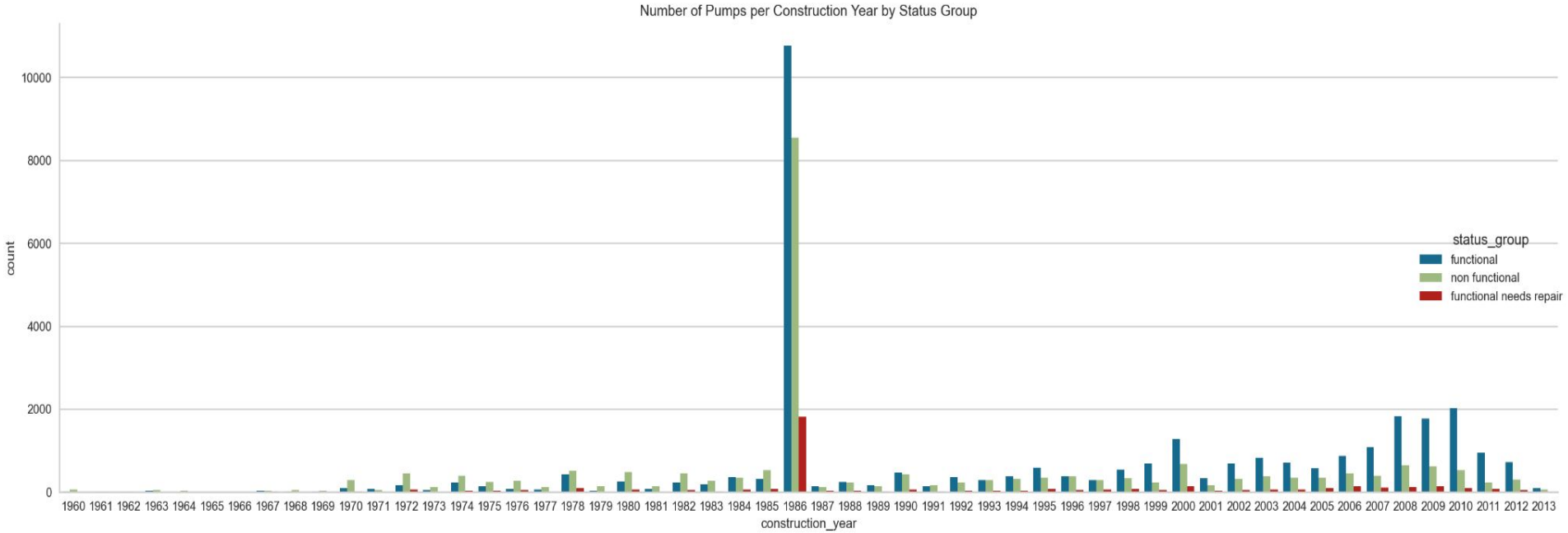
- Pumps with lower water quantities may be more likely to be non functional or needing repair.

Results



- Pumps in lower population areas may be more likely to be non functional or needing repair.

Results



- Older pumps may be more likely to be non functional or needing repair.

Conclusions

- **Location**: The Ministry should focus resources on lower altitude pumps.
- **Quantity**: The Ministry should focus resources on pumps with low quantities of water.
- **Population**: The Ministry of Water should focus resources on low population areas, as they may not be receiving enough.
- **Construction Year**: The Ministry should focus resources on modernizing older pumps

Next Steps

- The model and analysis are not complete solutions
- Model still struggles with identifying ‘functional needs repair’ pumps
- Model is overfit
- Scrub data further, create more features
- Use LightGBM, or Catboost model to attempt to improve accuracy, reduce overfitting, and reduce computation time

Thank You!

Email: norr.peder@gmail.com
GitHub: @pederknorr
LinkedIn: linkedin.com/in/pedernorr/