Alexia Marcous
Springboard Data Science Career Track
August 5th2019 Cohort

Capstone Project 2 – Predicting Water Pump Status
Milestone Report 1

Importance of Topic

According to the 2017 "Tanzania Economic Update" report from the World Bank, only 63% of the population has access to basic and improved water supply services[1], making the functioning of those services vital to the health of millions of people. The same report also said that "Tanzania needs to urgently improve the management of its water resources" to prevent disruption of the country's economic progress and sustain growth rates "long enough to make a significant dent on poverty".[2]

To help improve the management of water resources, a machine learning model will be created using supervised, classification algorithms to predict water pump status as functional, functional needing maintenance and non-functional. The model will also help identify which features contribute most to the prediction.

This model could be used by the Ministry of Water to make their maintenance schedule more efficient and effective. By knowing which pumps are more likely to break down, they can prioritize those pumps for maintenance. This model can also help the Ministry learn what are the factors that are important in determining the pump status, in order to start to address the root problems of pump malfunction.

Data Source

This project will use data collected from the Tanzanian Ministry of Water, which has been organized into datasets by DrivenData. These datasets were made available by DrivenData for a Kaggle competition and on their website, where they host competitions with humanitarian impact.

Data Wrangling

The data was in 2 datasets - one containing the target and the other containing 41 features and 59,400 rows. The target feature, status_group, has 3 possible outcomes: ‘functional’, ‘non functional’ and ‘functional needs repair’.

Visual inspection of the data showed there were two features with very high frequencies of the same data at 85%. These features were dropped. There were also two features that had duplicated data, and were also dropped.

Missing categorical data was evaluated. One feature was missing in almost half of the observations, and represented the name of the group that managed the pump. This feature was dropped. Missing observations in 5% of cases for a boolean feature were assigned randomly in the same frequency as the 95% of present values.

Percentage of zero values for numeric data was evaluated. A feature with 97% zeros was dropped, and other features were evaluated as to whether zero was a valid value, as with altitude (for pumps at sea level).

Three features contained survey-specific data, and were dropped as they did not add predictive value.

Datatypes were reviewed among the 31 remaining features, with 22 categorical and 9 numeric.

Pandas Profiling was run on this reduced dataset, and after the results were reviewed. Given that most of the features in this dataset are categorical, they will need to be one-hot-encoded for use by the machine learning models, which will create x(n-1) features for x features with n values. This makes reducing cardinality important. At the same time, there are many features with varying levels of granularity for the same data. For example, there are 6 features related to the location of the pump. The curse of dimensionality seemed impending.

Removing features and reducing cardinality will reduce the variability of the data, but this could lead to an increase in bias.

In an attempt to balance the bias/variance tradeoff, two datasets were created. The first ‘low-variance’ dataset had a set of features with less granularity, and the second ‘high-variance’ dataset had features with more granularity. For example, the type of pump was represented in both datasets as ‘extraction_type_class’ with 7 unique values in the low-variance set, and ‘extraction_type’ with 18 unique values in the high-variance set.

Fortunately, this did not result in a large difference in the number of features. The low-variance dataset had 11 features, and the high-variance dataset had 13 features.

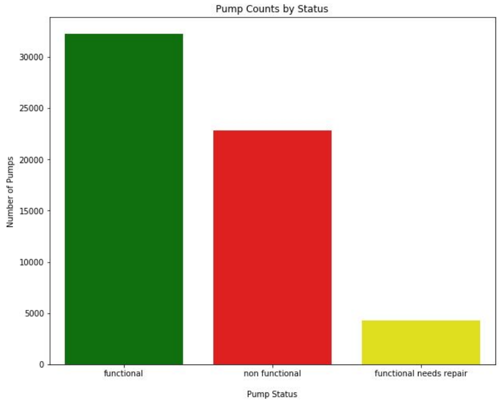
Data Story

Visual data analysis techniques were applied to gain an understanding of the data. 4 main questions were asked:

1. **What is the distribution of pump status values?**

As the straightforward bar chart shows, the distribution of pump status is very uneven.

54% are ‘functional’
39% are ‘non-functional’
7% are ‘functional needs repair’



2. **How does the pump status vary over features that would seem to have an impact on the functional status?**

   We will examine construction year, pump type, management, geographic location, altitude, payment_type, permit status, quantity, source, and waterpoint type?

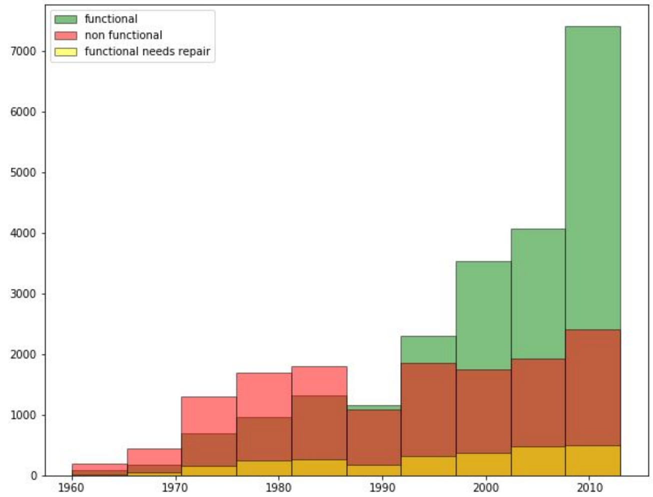
   As these features were being evaluated, opportunities for reducing cardinality were discovered. Highlights of the visualizations are presented below.

   Pump status by construction year

From this histogram of pump status by year constructed, we can see that pumps built before 1990 are non-functional more than they are functional.

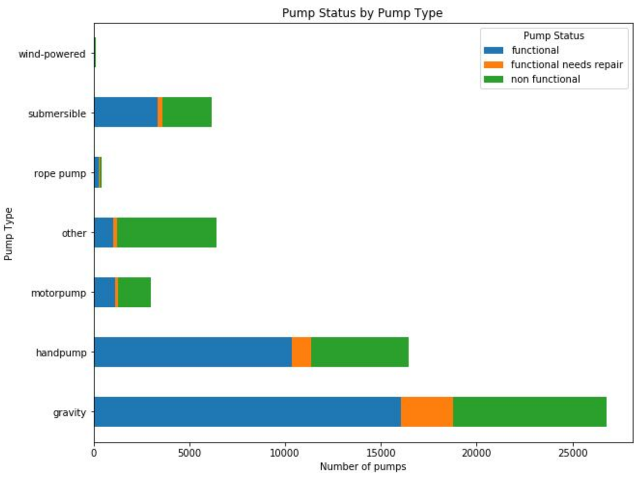
After 1990, the number of functional pumps increases, and the total number of pumps has increased dramatically.

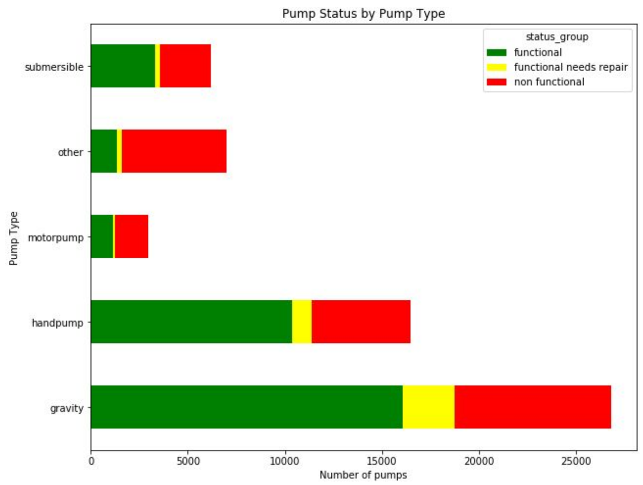
Unfortunately, 30% of the pumps do not have a construction year recorded, so the predictive quality of this feature may not be very strong.



Pump status by pump type

The first visualization of pump types on the left revealed that the majority of the pumps were gravity pumps, and also that there were very few pumps in the ‘wind powered’ and ‘rope pump’ categories. To reduce cardinality, observations in these two categories were combined into the ‘other’ category, resulting in the distribution on the right.





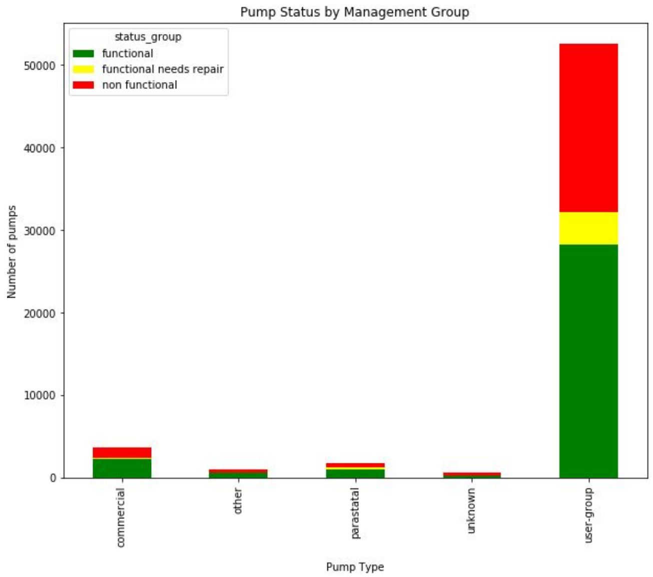
Pump status by management group

At first glance, there didn’t seem to be enough variability in the type of management groups to be of value for predictions.

However, the distribution of pump status is informative in terms of studying the impact of management on pump status. Among the most common type of management, user groups, there is high variability among the pump status. It would be interesting to learn best practices and support structures of the groups with functional pumps.

Also, despite the small number of observations, the commercial group has more than half of their pumps operating, making this another group to learn from.

The visualization also revealed an opportunity to reduce cardinality by combining the three smallest categories into the ‘other’ category.



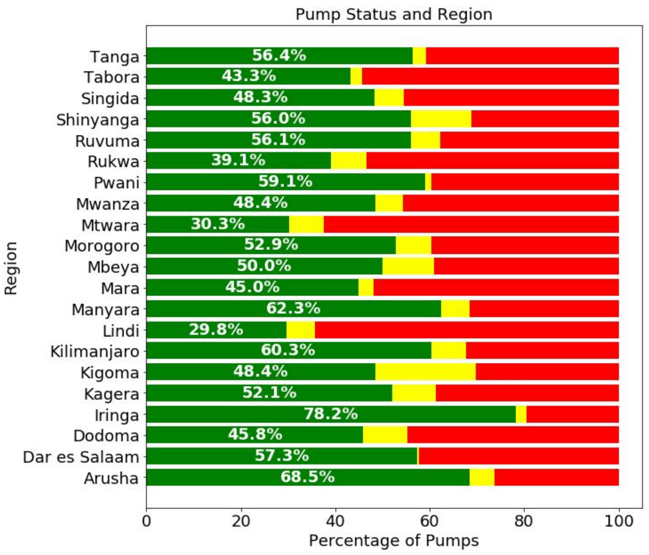
3. **Which features should be chosen when there are multiple levels of granularity for the same data?**

An example of this in this dataset, were 2 numeric, latitude and longitude, and 6 categorical features representing the geographic location of the pump: region, region_code, district_code, lga, ward, subvillage. Having so many features describing the same information, especially of the same datatype, introduces the idea of collinearity, which would have a negative on model results. Also, the lowest level of the hierarchy, subvillage, had 19,288 unique values, guaranteeing the curse of dimensionality. To address this, categorical geographic features at the highest level of granularity were visualized to determine the distribution of pump status across cardinality.

Pump status by region, the top of the location hierarchy, showed that there were 21 regions, with a good amount of variability among pump status types.

Knowing that the 21 regions would be one-hot-encoded into 20 features, this feature was chosen for the low-variable dataset.

Region_code, with 27 different values, was chosen for the high-variable dataset, and latitude and longitude features were kept in this dataset.



Statistical Analysis

Given that the data from the water ministry is mostly categorical, and the goal of the project is multiclass prediction, the statistical analysis of the data was focused on testing for collinearity among features and between features and the target.

To remove any consideration of ranking, Cramer's V was run between all features, a nominal version of Pearson's Chi-Square Test for independence between categorical data. The test returns values between 0 for complete independence and 1 for complete dependence in terms of the effect a category for a feature has on the probability of a category for another feature occuring. Tests were run on the high and low-variance datasets.

These results on the low-variance dataset shows a strong correlation between basin and region. This makes sense, as pumps in the same region would draw from the same basin. There is also a strong correlation between extraction type class and waterpoint type, underscoring the repetition of some of the same data points in the two features.

Of all the features, region had the most correlations with other features. This suggests there are substantial differences among the regions of Tanzania, since as the region changes, so the other features change. This makes sense in a country as diverse as Tanzania, which contains the plains of Serengeti National Park, the mountains of Kilimanjaro National Park, and a coastline close to tropical islands.

Of interest, there is also a correlation between the extraction type class and the source, revealing that certain types of pumps are used in certain types of sources. This is probably common knowledge among pump installers, but it takes statistical analysis to identify that point to those with less contextual knowledge about the data.

The analysis of correlation on the high-variance dataset also shows a strong correlation between basin and region, this time using a feature, region_code, that had more variability. Interestingly, the correlations between region and other features are present but not as strong as in the low-variance dataset.

The same phenomenon appears with extraction type and waterpoint type. The correlation is present but not as strong as in the low-variance dataset.

This change in correlation might imply that the low-variance dataset would be the better dataset for creating the machine learning models.

With regard to correlations between the features and the target, in the high-variance dataset show slightly higher correlation to the target than the features in the low-variance dataset. This implies that the high-variance dataset would be the better set for machine learning. Both sets will be tested and compared in the machine learning phase.

Examining Significance of Data Story Relationships

The visual data analysis revealed potential relationships between pump status and year, pump type, management group and region. The following table shows the Cramer's V test results for these features:

Feature	**Cramer's V low-variance**	**Cramer's V high-variance**
Year	0.18	0.18
Pump Type	0.23	0.25
Management Group	0.045	0.045
Region	0.2	0.2

The test statistics do not show strong correlations, with all p-values near zero. This underscores the importance of following up visual data analysis with statistical analysis to determine if what looks like a relationship is actually statistically significant.
The symmetry of the correlations was evaluated with Theil's U test, also referred to as the Uncertainty Coefficient. This test is based on conditional entropy - given the value of one feature, how many possible

states does another feature have, and how often do they occur. The test output is in the range of [0,1], where 0 means no association and 1 is full association.

Only the ‘quantity’ feature demonstrated an asymmetrical relationship with a Theil’s U value of 0.11. This value was the same in the high and low-variance datasets.

After wrangling, visualizing, and performing statistical analysis on the dataset, it does not appear that there are very strong correlations between any of the features and the target variable. However, machine learning algorithms may be able to uncover as-yet-unseen relationships, and having clean, organized data is the right preparation for this next step.

References

1. “Water Stress Could Hurt Tanzania’s Growth and Poverty Reduction Efforts – New World Bank Report”, accessed on 1/3/2020 at https://www.worldbank.org/en/news/press-release/2017/11/06/water-stress-could-hurt-tanzanias-growth-and-poverty-reduction-efforts---new-world-bank-report

2. Ibid.