Water Pumps Functionality Prediction in Tanzania!

1. Overview

Water is critical to a country's development, as it is not only used in agriculture but also for industrial development. Whether it's a drink from the tap, a toilet that works, or a place to wash our hands, having water is a basic necessity that many of us take for granted. But in many parts of the world, women and children spend more than four hours walking for water each day. Some African countries like Tanzania have large water resources, but they still face the dilemmas where many areas have no reliable access to clean water because these resources are distributed unevenly. During the dry season, which usually lasts from June to October, even large rivers can dry up or their flow declines substantially.

In this project, we will use Machine Learning methods to help the Government of Tanzania with predicting which pumps are most likely to stop working so they can be prepared for the pump failure in advance. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania and the results of this project will aid the maintenance, process and costs.

The Random Forest model is able to predict 85% of the pump statuses correctly and the results of this model can be used to skip checking the functional pumps manually throughout Tanzania which would save us $1,446,300 using a 90% functional rate that can be used to lower the cost of the faulty pumps replacements in the future. If a pump is non functional, it requires immediate attention as the population dependent on that water source cannot access clean water.

1.1 Problem Statement:

Hand-driven and gravity fed pumps are still a main source of potable water in Tanzania and the maintenance of these pumps is an ongoing issue. Because many communities either don't have the funds or the knowledge to maintain these pumps, they're constantly breaking! The purpose of this challenge is to predict which pumps will not be functional in the future, in order to expedite the process of locating, examining, and fixing the pumps.

1.2 The client and why do they care about this problem:

The client for this problem is the Government of Tanzania. They should care about this problem because using this algorithm would be life saving to many communities in Tanzania. It will help ensure most pumps remain functional, and let them know which communities they need to give more attention to maintain their pumps. Not only will this be useful in Tanzania, but we can use it in any other country that relies on a similar source for water.

2. Data

Our approach is to find those features which would be important in predicting the functional status of a pump. For example, the location and the population around any particular water point might play a vital role in the status of that water point. The quantity and construction year for any particular water pump would play a critical role in determining the status of that water pump. There are some features with missing values and we need to decide whether to drop that feature, fill the missing values with the mean and median, or to assign them to the 'unknown' category.

2.1 Dataset Description:

The data is provided to us by Taarifa and the Tanzanian Ministry of Water. This dataset contains 59,400 data points and 41 features.
The dataset has two parts. Train values and Train labels are loaded on URLs below. We merge them together to have the full dataset.

- Train values [1]
- Train labels [2]

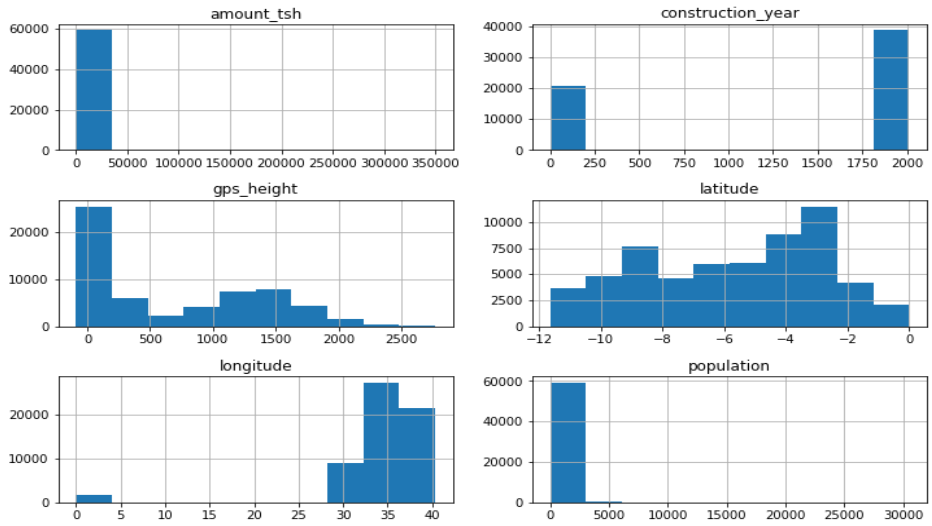
2.2 Data Cleaning:

[1] http://s3.amazonaws.com/drivendata/data/7/public/4910797b-ee55-40a7-8668-10efd5c1b960.csv

[2] http://s3.amazonaws.com/drivendata/data/7/public/0bf8bc6e-30d0-4c50-956a-603fc693d966.csv

We assume that the data of the water pumps were collected using handheld sensors, paper reports, and via cell phones so many of the feature values were not accurate or have missing values.
There are some features with missing values. The most noticeable features with missing values are ‘construction_year’, ‘gps_height’, ’population’, ‘amount_tsh’, ’latitude’ and ‘longitude’. For example, you can see that approximately 35% of the ‘construction_year’ feature contains 0 (20709 rows) [3]. We also have 36% (21,381 rows) of ‘population’ with the value of 0 and 12% of this feature with the value of 1. For other features like ‘amount_tsh’, we have about 70% of rows with the value of 0 and ‘gps_height’ feature with 34.5% of rows with the value of 0.
We need to be deciding whether to drop them, fill the missing values with the mean and median based on the region they are in, or to assign them to a ‘Missing’ or ‘Unknown’ category.

The histograms of the numerical data were plotted to reveal some of the outliers or invalid data.

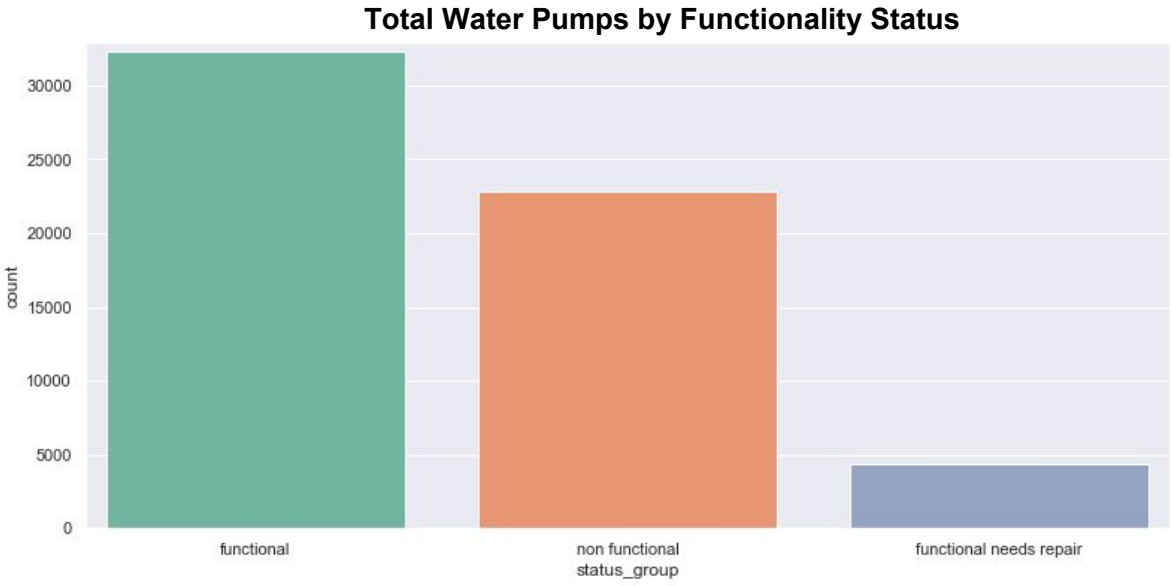


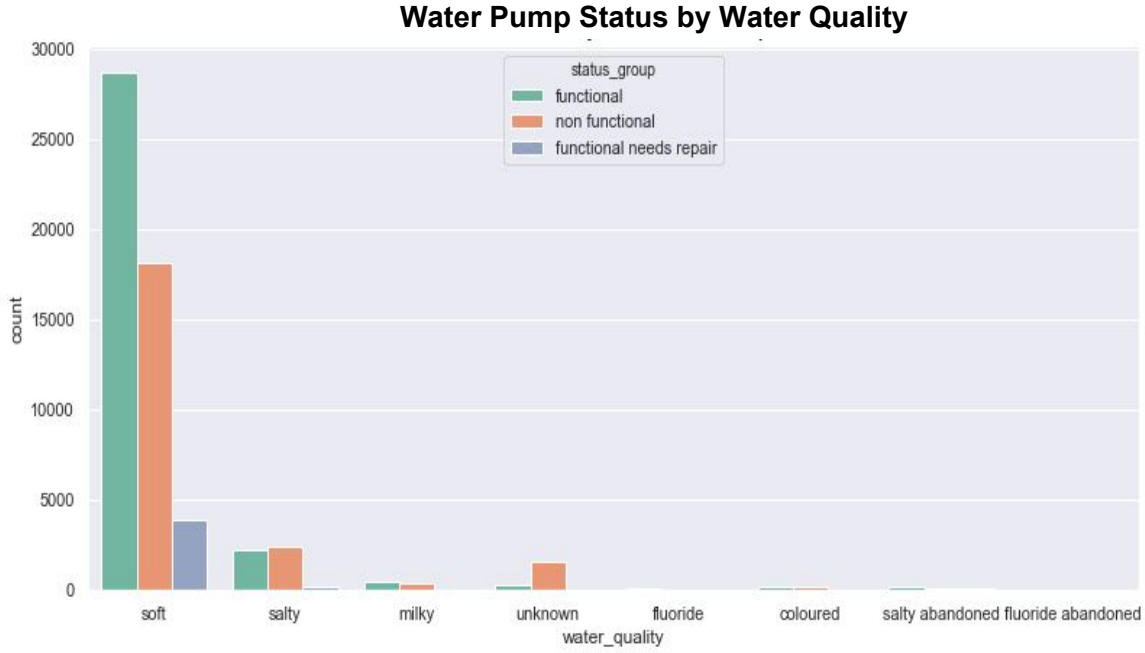
Some of the categorical features have duplicate data points. To fix this issue, we added one of these duplicated set of features like (region and region_code), (waterpoint_type_group and waterpoint_type), (source and source_type), (quantity and quantity_group), (water_quality and quality_group), (payment and payment_type), (extraction_type and extraction_type_group), (installer and funder) which are almost representing the same data, and then dropped the ‘garbage_features’ from our df. [4]

After dropping all the duplicate categorical features and taking care of the missing values of numerical features (NaNs and 0s), now we have 25 features. We created a new feature called ‘pump_age’ which is ‘date_recorded’ minus ‘construction_year’ to keep track if older pumps are more likely to fail. We still have a lot of features and there are only a few of them that would be sensitive to predicting the status of the water pumps. Ideally, we would have a couple of features and later on, we will add more features to see if any improvements can be made in our model scores.

[3] Lines 15-21 of the code
[4] Line 65 of the code

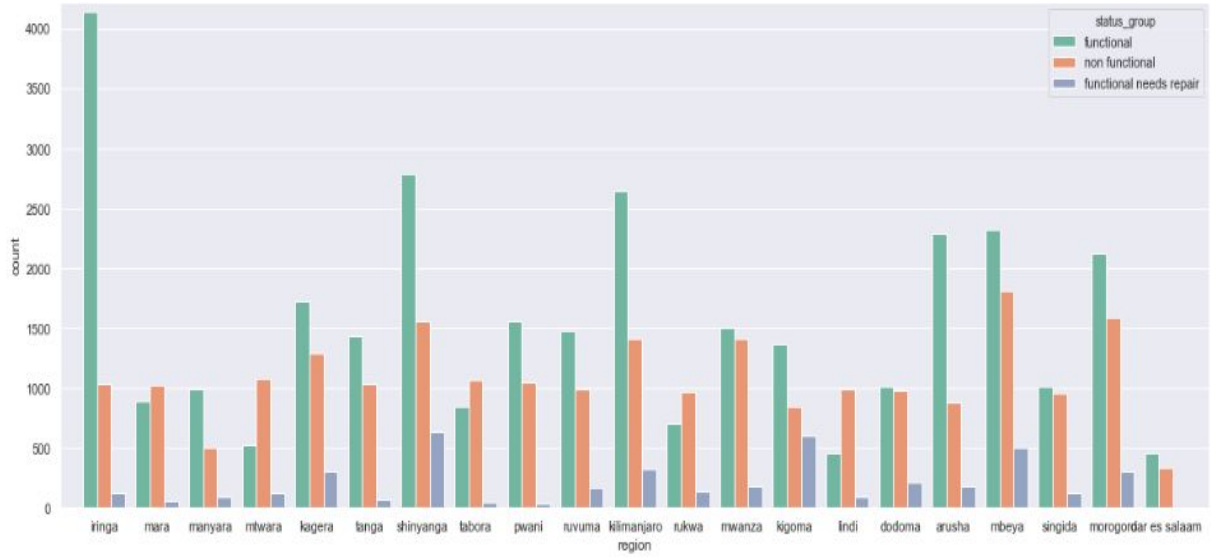
2.3 Data Visualization:





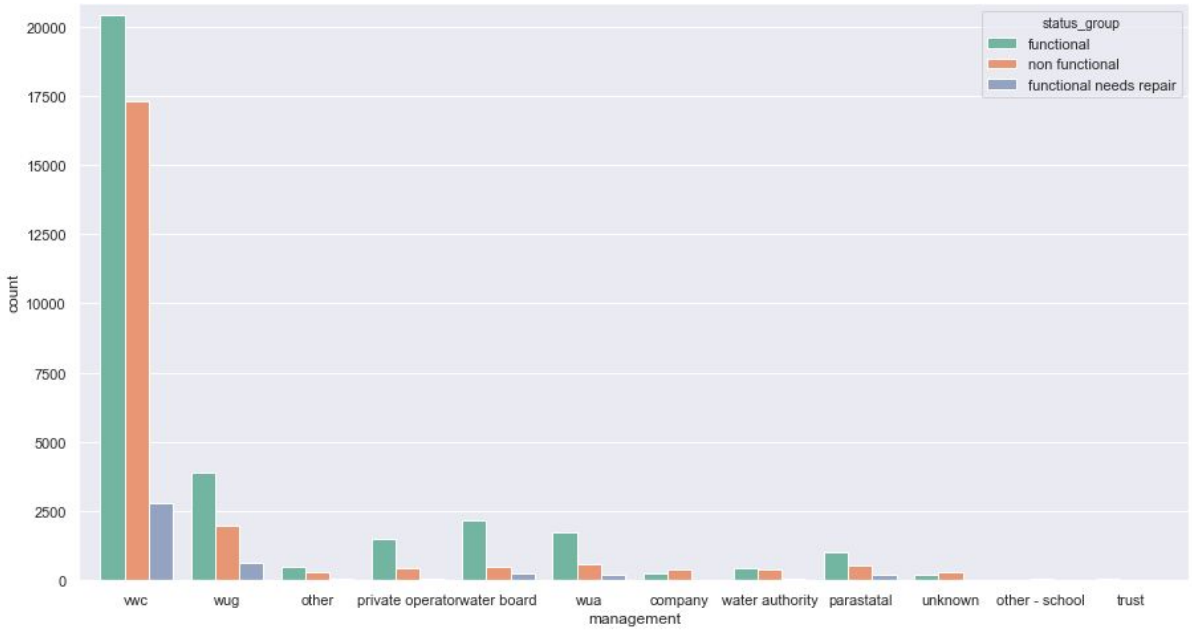
When the water quality is soft, there is a very high probability of the water pump being functional. Also, when the water quality is salty, there is an equal probability of functional and non-functional status!

Water Pump Status by Region



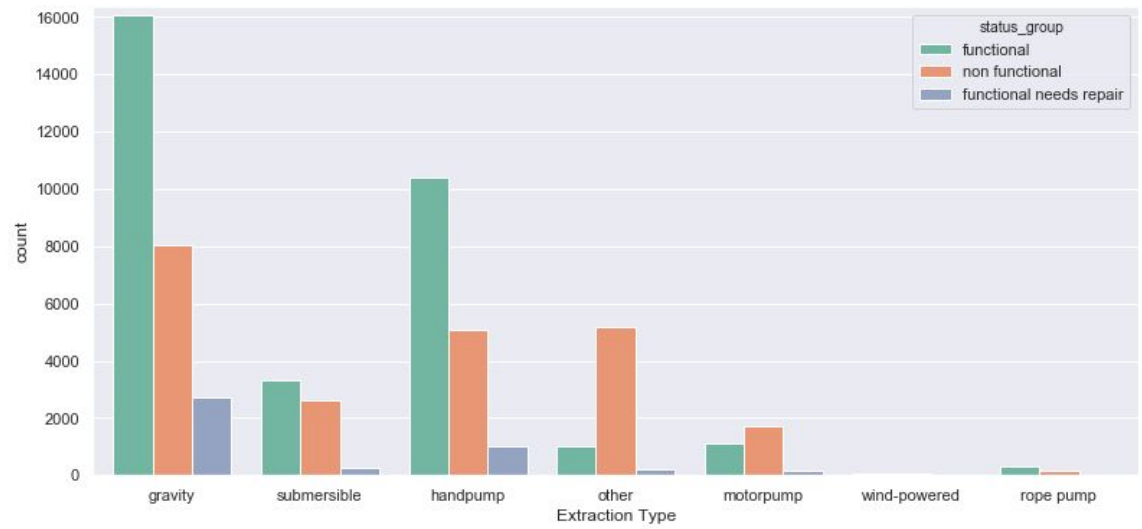
In some regions, there is a very high probability of a water pump being functional vs non-functional.

Water Pump Status by Management Types



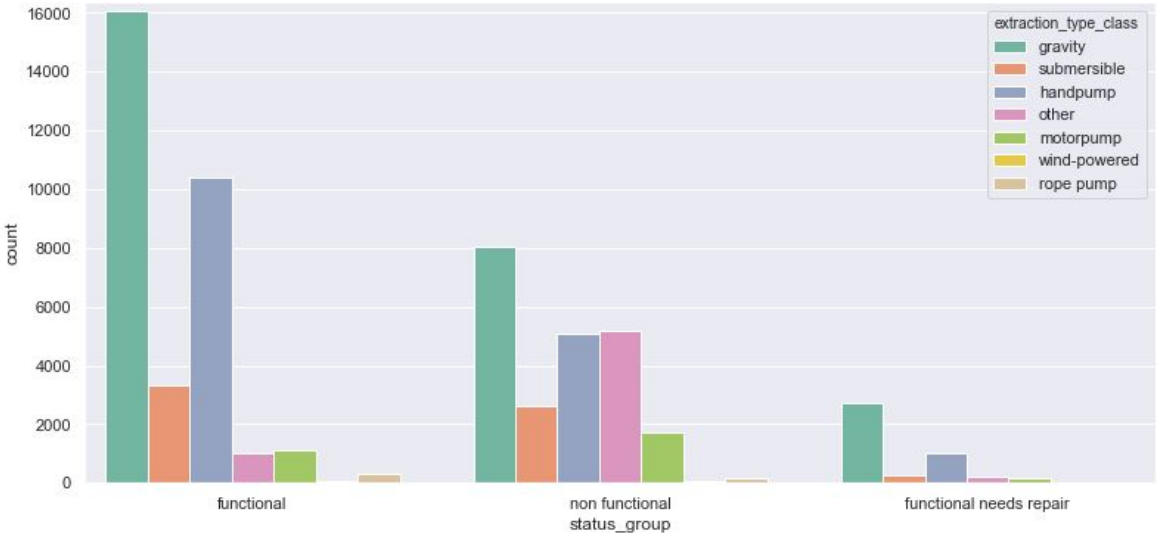
The (vwc) manages approximately 70% of the pumps! They also have the highest numbers of functional and nonfunctional pumps.

Water Pump Status by Pump Types

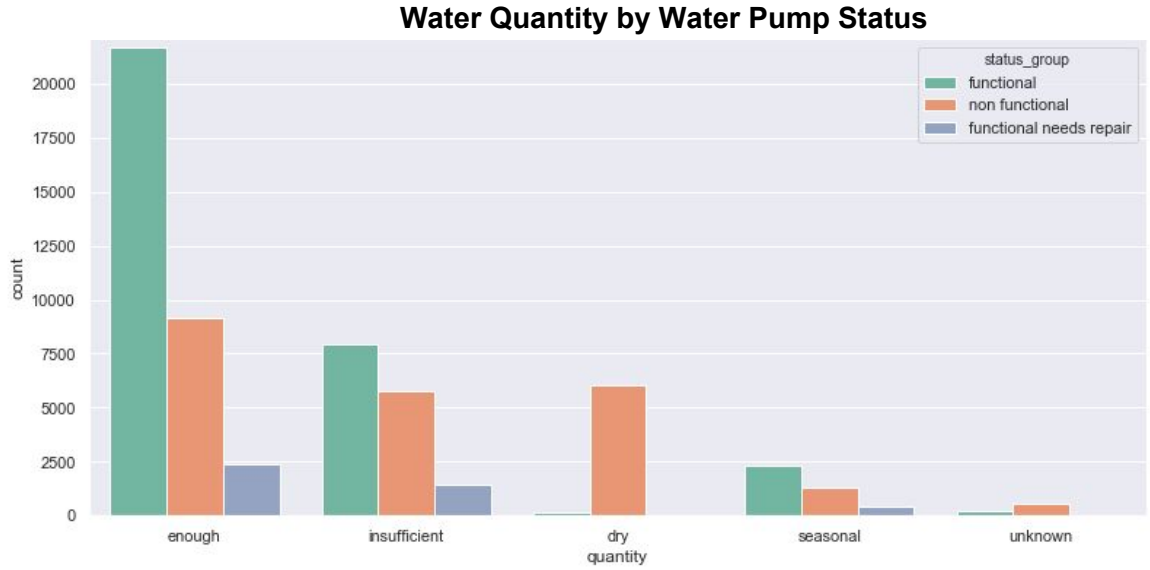


70% of the pumps are type gravity and hand pumps. The gravity and hand pump types have the most functional ones. The motor pumps have a higher probability to be nonfunctional than functional.

Pump Types by Water Pump Status



The gravity and hand pump types have the highest percentage in all of the status groups.



If the water pump gets enough water quantity, the probability of it being functional is high. On the other hand, if the water quantity is dry, a water pump is most likely nonfunctional.

3. Modeling

Now we need to divide our dataset into training and test sets. The target variable (status_group) consists of three different groups of water pumps: Functional, Non-function, and the ones that are still functional but require repair work. This is a classic classification problem of how to accurately predict the classes. The challenge is if we can predict which pumps are Non-functional.

For this classification problem, Decision Tree and Random Forest models were used. GridSearchCV was also used to find the best parameters. The ROC Curve and the AUC score were added to check the accuracy for each model. Then, we plotted each model score on a histogram to compare these scores easily.

You can see that we have used three lists that hold names of our categorical, numerical and date features.[5] This helps us later on to add/remove any of these features from our training and test sets.

The date_recorded was divided into three features which will be used later to see if any of them help with any of our models.

In order to help model performance and make the model fitting process faster, the top 10 values in each categorical feature were retained while the rest were dropped. Then, the rest of the values were set to “OTHER”. This helps with having less transformed columns when we do the ColumnTransformer in later steps.

To transform all the categorical features at once, compose.ColumnTransformer was used. We have defined a pipeline (transformers) that gets the categorical feature names from the list we defined on line 3, and then performs the OneHotEncoder on them. OneHotEncoder, encodes categorical features as a one-hot numeric array. By default, the encoder derives the categories based on the unique values in each feature and this is the reason we reduced the feature sizes to the top 10.

The categorical features were converted to ordinal digits as part of the transformation and the column names were changed to show which ones were representing the values from each original column. Research showed that a useful function can attach the name of the original columns to the values and add them as the new column names. We also concatenate all the transformed categorical features with our numerical features and the date features.
A binary target was chosen for (status_group) values. To do this, we assigned all of the functional and needed repair pumps to 1 and all of the broken ones to 0.

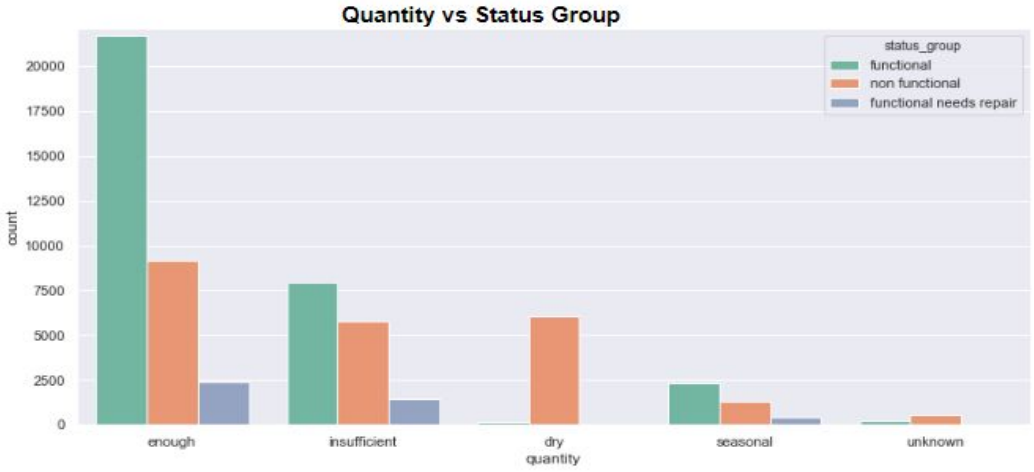
The data was split into training and test sets. The split was done using 70% for training and 30% for testing. Random_state was used to make sure our split is always the same.

[5] Line 3 of the ML code

After splitting the data, we used our first model which was the Decision Tree Classifier. We set a few hyperparameters for this model but later on we will use the GridSearchCV to find the best_params_ for this model. The accuracy score for our Decision Tree model was 82% which shows the percentage of correct predictions to the total number of input samples.

In order to make sure that the model was not guessing the output, the ROC Curve as well as the AUC score were used. The AUC score was 0.85 which is much greater than 0.5 and this means the model is better than random guessing.

We used our model's feature_importances_ parameter and plotted the most important features of the model. It looks like quantity-dry plays a very important role in the Decision Tree model. This is something that we have discovered previously that if the quantity is dry, then the pump is most likely non-functional. You can see this result by looking at the chart below.

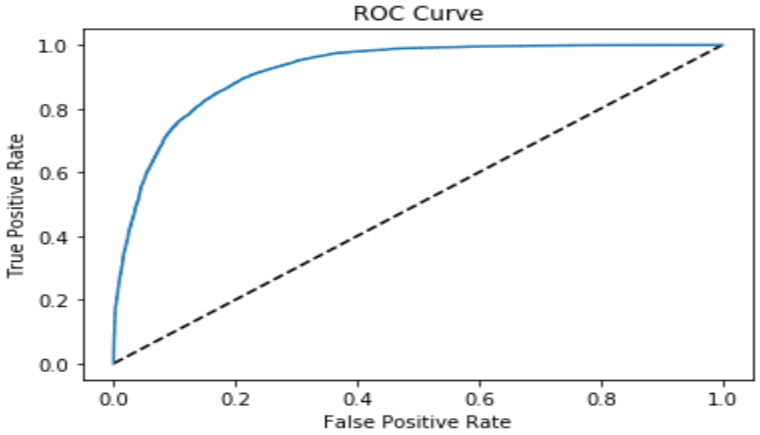


As previously mentioned, this is a classification problem and the Random Forest Classifier is the best model to make predictions on because it has the highest accuracy score between all of our models. A few more models were used and then they were compared by their scores at the end to see which one has the highest score and can make the best predictions.

After selecting the Random Forest model as our best model, we started training it using the n_estimators=200 as well as setting the criterion='gini'. The accuracy score of the RF model is 85% which shows the percentage of correct predictions to the total number of input samples. The formula below was used to evaluate the accuracy of our model. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

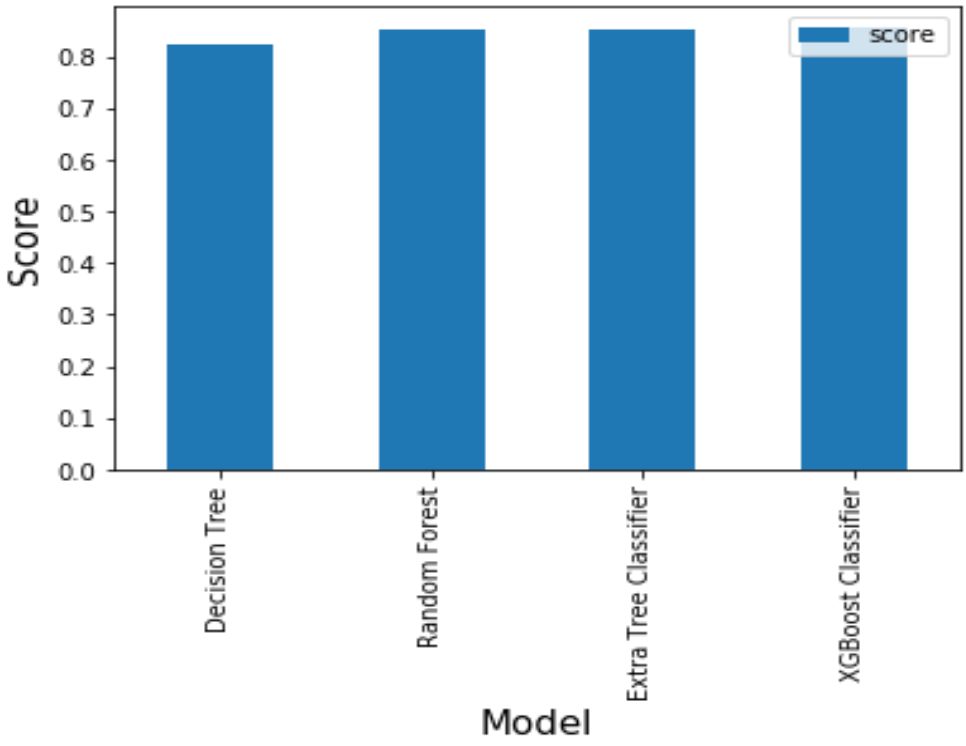
In order to find the best_params_ for this model, the GridSearchCV was used.



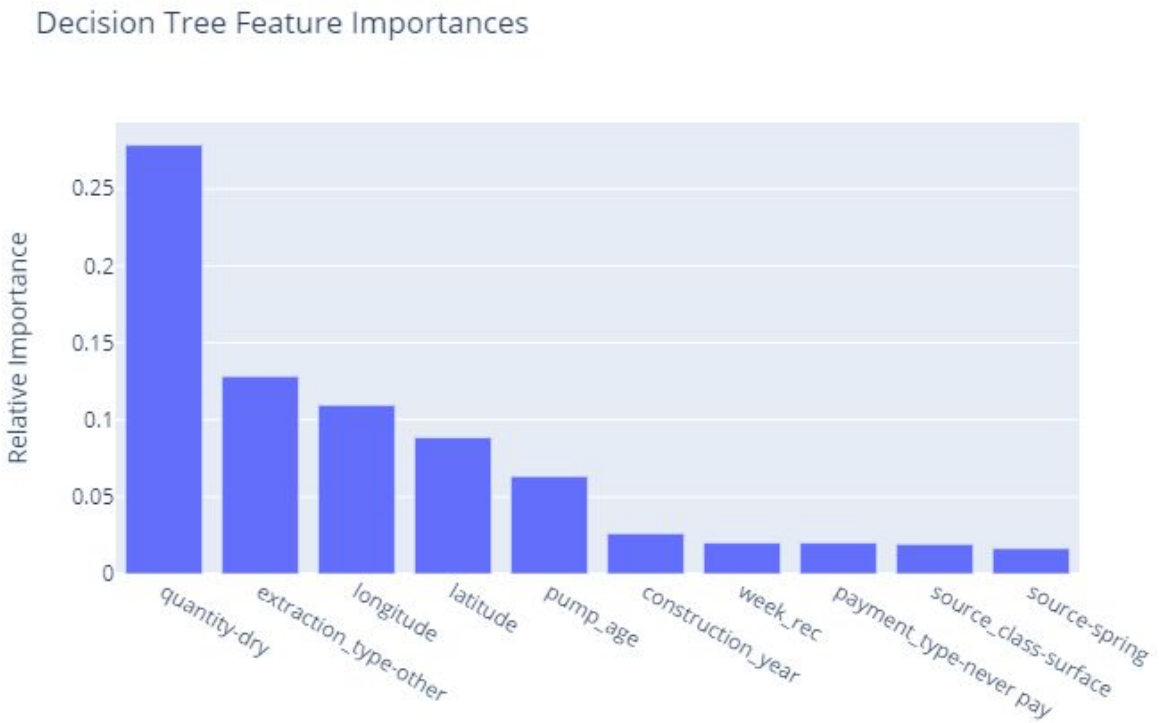
Just like the Decision Tree model, we used the ROC Curve shown above as well as the AUC score. The AUC score is 0.92. The ROC Curve and AUC score for Random Forest Classifier are higher and the higher AUC score means a more sensitive and better performing model.

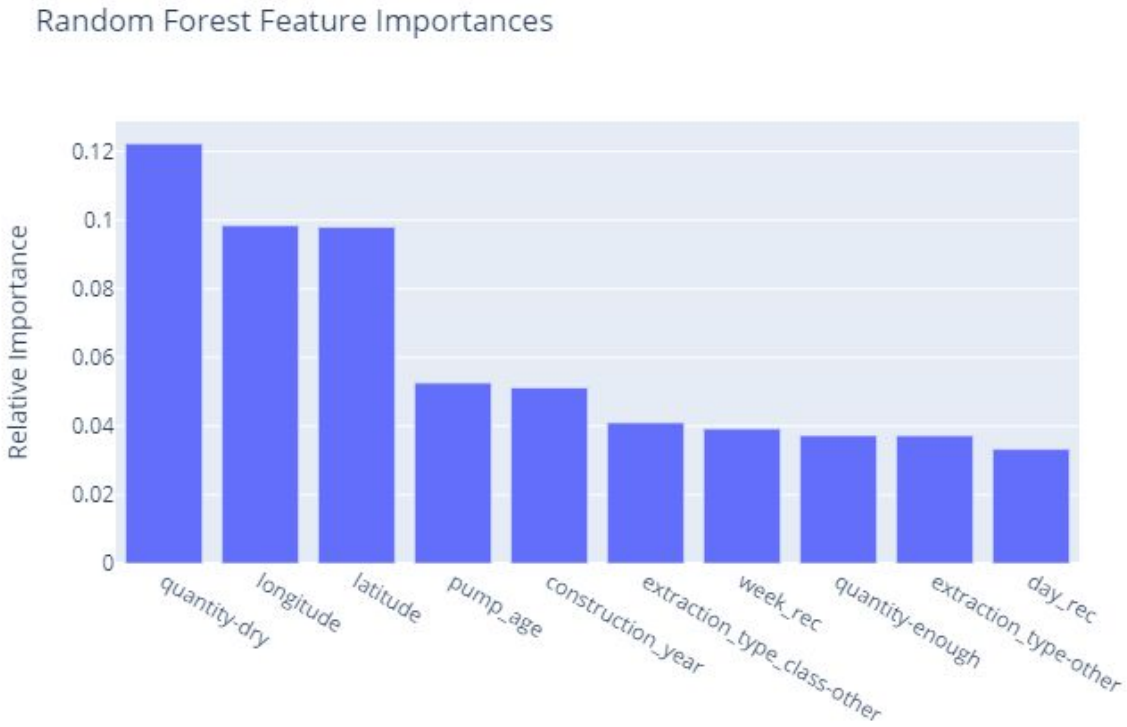
Plotting the most important features shows that quantity-dry, longitude and latitude are playing important roles in this model.

As for other models, we used the Extra Tree Classifier and the XGBoost Classifier. The results are pretty close to the Random Forest model. You can see the score comparison on the chart below.



3.1 Model Performance:





Our Decision Tree modeling efforts indicated that the quantity-dry, extraction_type-other, longitude, latitude, the derived pump_age and construction_year variables appeared to have the highest levels of predictive power. The week_rec and quantity-enough variables appeared to be of little or even no value to the Decision Tree model but had very high predictive value for the Random Forest and Extra Tree models.

By contrast, the funder and many of the district_code and lga variables were found to have relatively low predictive strength across these four types of models, and the region variable was found to add no value if it was already included within a model.

The overall accuracy of the four models when applied to the test set ranged from a low of approximately 82% (Decision Tree) to a high of just over 85% (Random Forest).

The Random Forest model scored highest achieving an overall accuracy of .85. Therefore, the Random Forest model is recommended for use by Tanzania's Ministry of Water if overall accuracy across each of the possible pump statuses is of most importance.

4. Conclusions

Access to water fuels health, hope, education, optimism, and prosperity. It boosts opportunities for everyone. Children can stay in school longer. Families have more time to work, play, and simply spend time together rather than wasting that time by walking to a pump that is not functional! In this project, we attempted to predict the functional status of each water pump from one of the two possible status values (functional or non functional) based on a variety of qualitative and quantitative water pump attributes and geographical variables.
Insights gained from this project can help the Ministry of Water to save money by predicting which pumps have failed. Currently, in order to check these pumps, it is required to send out a technician to every pump across Tanzania. This also means travelling to functional water pumps.

4.1 A Case Study:

Assume we want to maintain a functional rate of 99% throughout the year. This means that for every 100 functional pumps, we will keep 99 of them functional by either replacing or checking them to make sure they are functional. If the cost of each pump check is $50 and the cost of each pump replacement is $2,000, we will evaluate the savings below using No Model and Model results for a 99% sample and 1% non-sample. Pumps predicted as functional, will not be checked.
After making predictions using the Random Forest model on the training and test sets, the results are combined to have a full dataset again for this case study. Then we take a 99% sample from the merged dataset and we notice that there are 22,571 non functional pumps out of a total of 59,400 pumps. No model is being used in this step so all the pumps in the 99% sample are predicted as non-functional and all the pumps in the 1% non-sample are functional in order to maintain the 99% functional rate. We need to keep in mind that in the 1% non-sample, we have non-functional pumps that we need to replace throughout the year so even though the check pump cost in the non-sample is $0, we have to set $506,000 aside to fix these pumps when they break.

Functional Rate	no-model cost of checks	no-model cost of replacement	model cost of checks	model cost of replacement	Savings
99%	$2,940,300	$45,648,000	$720,900	$45,648,000	$2,219,400

The figure above shows the Saving amounts in 99% Functional Rate

Savings between No Model and using Model is ($48,588,300 - $46,368,900 = $2,219,400). This is how much we save by using the model with a 99% functional rate.

Now, let's assume that we want to change the functional rate to 90% and see how much more we can save! This time, we have 20,514 non functional pumps out of a total of 59,400 pumps. Just like the 99% sample, we predict that all the pumps in the new sample are non-functional and all the pumps in the non-sample are functional in order to maintain the 90% functional rate. We have to set $4,620,000 aside to fix the non-functional pumps that are in the 10% non-sample and need to be replaced throughout the year. We also have to check all the pumps

in the 90% sample and this will cost $2,673,000. The total cost of pump checks and replacements in sample and non-sample pumps without a model is **$48,321,000.**

The cost of checking all these pumps is $1,226,700. Out of 24,534 predicted non-functional pumps, we have 21,166 pumps that are actually non-functional and need to be replaced. The cost of non-functional pump replacement is $42,332,000 and this would bring the total cost of pump checks and replacements using Model to **$46,874,700.**

Functional Rate	no-model cost of checks	no-model cost of replacement	model cost of checks	model cost of replacement	Savings
99%	$2,940,300	$45,648,000	$720,900	$45,648,000	$2,219,400
90%	$2,673,000	$45,648,000	$1,226,700	$45,648,000	$1,446,300

The figure above shows the Saving amounts in different Functional Rate %

Savings between No Model and using Model is ($48,321,000 - $46,874,700 = $1,446,300). This is how much we save by using the model with a 90% functional rate.

For simplicity, we can see the different Savings amounts between the 99% and 90% Functional rates in the figure above. Ultimately, the Tanzanian Ministry of Water and local organizations employing this research will have to decide what percent functional rate they want to use.

4.2 Future Work:

While the conclusions described herein apply only to the water pumps described in the aforementioned DriventData.org data set, the extensible/reusable design of the data transformation pipeline allows much of what we've implemented to potentially be applied to other types of predictive/classification modeling efforts. Future work could include creating a visualization dashboard to incorporate graph theory techniques for purposes of identifying clusters of pumps that are in need of repair. Such functionality could provide the Ministry of Water with a way to improve the efficiency of its repair efforts, e.g., repair teams could be deployed to clusters of problematic pumps, etc. Moving forward, we will also focus on other models like the SVM model and further improve our success metrics for the project by defining how many lives are we improving by fixing a water pump. Since there is a water crisis in Tanzania, a dashboard for Tanzanian Ministry of Water that specifies the functional but needs repair water pumps and the number of lives it will help improve by fixing the water pump will be important.