2019

# Pump it Up: Data Mining the Water Table

NagaSanthosh Adaveni, Brian Keafer, Srinivasan, Anand

DSA 5103-Intelligent data analytics-Fall2019

12/9/2019

# Table of Contents

# Executive Summary

This project aims at applying data analytics knowledge and skills acquired in duration of this course to real life situation and determine solution to the problem at hand. A dataset with details of waterpoints across locations in Tanzania has been selected. The task at hand is to determine correlation between various parameters in this data and predict the state of each water point as one of three predefined categories. This prediction would help classify the wells as functional, functional but needs repairs, and non-functional.

As part of the data understanding, histograms were plotted to understand the data spread as well as the missingness. Box plots were also created to understand the skewness/outliers. Longitude/Latitude based plots provided an understanding on how the waterpoints are spread and if this related to functionality of each well. Various plots and matrices also help deduce the correlation between the data points.

Data munging, duplication identification and removal was the first step. A very simple missing data imputation strategy was applied. A summary of missing values was obtained and then all the predictors with a high percentage of missingness were identified and imputed with NAVL values. Remaining predictors were further lumped into buckets for better interpretation. Symbox and Boxcox transformations, Principal components analysis, and Linear discriminant analysis of variables were carried out in order to reduce dimensionality and re-shape abnormal distributions, but these did not yield any significant results or help identify key variables to consider for model training and analysis. Hence, models were run to get a summary that let us know the important variables to consider.

The NaiveBayes model was then run to identify relevant predictors. The refined set of predictors were then used for training other models like penalized linear regression(glm), support vector machines (SVM) and random forest(rf). Other various models have been trained (penalized regression model, decision tree, SVM, random forests) have been trained to predict the status_group variable. The random forest performed above the rest, producing an accuracy measurement of about 75%. However, certain factor levels in the dependent variable were not predicted with the same consistent accuracy, namely “non-functioning” and “needs repair”. A uniform size random sample was utilized to increase the distribution of the variables of interest in the sample size to train the model. This increased accuracy of the factor levels of interest. A summary of accuracy for various models have been listed in results analysis section. Future work on this project would be generalizing this algorithm for various kind of equipment failures and being able to predict failure before it happens, not after the fact.

# Problem Description:

The Tanzanian Ministry of Water aims at providing potable water at all water points throughout the country. Not all water points are functioning so it is necessary to identify the water points that are functional and the ones that are not functional so that non-functional one’s can be replaced. In addition, identification of water points that are functional but needs repair will help meet objective of providing potable water at all water points with lesser effort and cost compared to installing a new waterpoint.

Achieving a precise and accurate understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania. A data set has been provided by Taarifa and the Tanzanian Ministry of Water. Taarifa aggregates data from the Tanzania

Ministry of Water. The data set provided has variables like kind of pump, year it was installed, and how it is managed etc. There are 39 variables in all. They are numeric, characters as well as factors.
The objective is to understand the correlation between these variables and develop various regression models with hyper tuning parameters in order to predict the operating condition of a waterpoint for each record in the dataset. The predicted operating condition needs to be one of the three possible values.

- functional - the waterpoint is operational and there are no repairs needed.
- functional needs repair - the waterpoint is operational but needs repairs.
- nonfunctional - the waterpoint is not operational.
-

# Predictor Details

The following set of information has been provided about the waterpoints. Names of all the variables present in the data set and a description of what they stand is documented below. There are many variables that provide the same type of information and in some cases have zero variance between them. Therefore, it was only necessary to include a few of the most helpful variables from each category.

**Location Variables**

- longitude – GPS coordinate
- latitude – GPS coordinate
- basin – Geographic water basin
- subvillage – Geographic location
- region – Geographic location
- region_code – Geographic location (coded)
- district_code – Geographic location (coded)
- lga – Geographic location
- ward – Geographic location
- gps_height – Altitude of the well

**Operation**

- scheme_management – Who operates the waterpoint
- scheme_name – Who operates the waterpoint
- permit – If the waterpoint is permitted

**Extraction**

- extraction_type – The kind of extraction the waterpoint uses
- extraction_type_group – The kind of extraction the waterpoint uses
- extraction_type_class – The kind of extraction the waterpoint uses

**Management & Payment**

- management – How the waterpoint is managed
- management_group – How the waterpoint is managed
- payment – What the water costs
- payment_type – What the water costs

**Quality & Quantity**

- water_quality – The quality of the water
- quality_group - The quality of the water
- quantity - The quantity of water
- quantity_group - The quantity of water

**Source & Kind**

- source - The source of the water
- source_type - The source of the water
- source_class - The source of the water
- waterpoint_type - The kind of waterpoint
- waterpoint_type_group - The kind of waterpoint

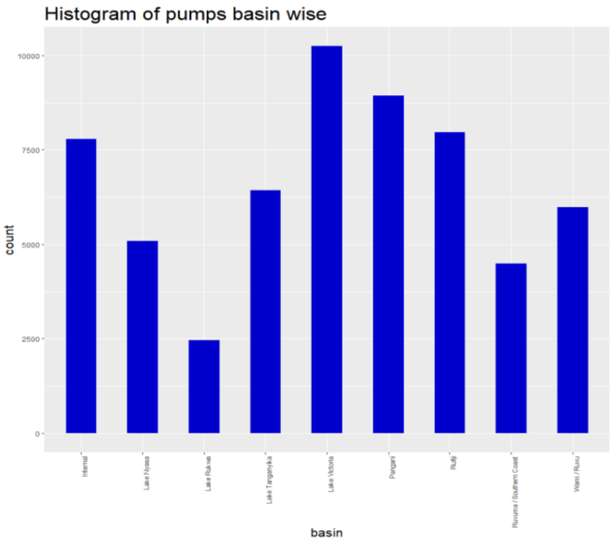
**Various**

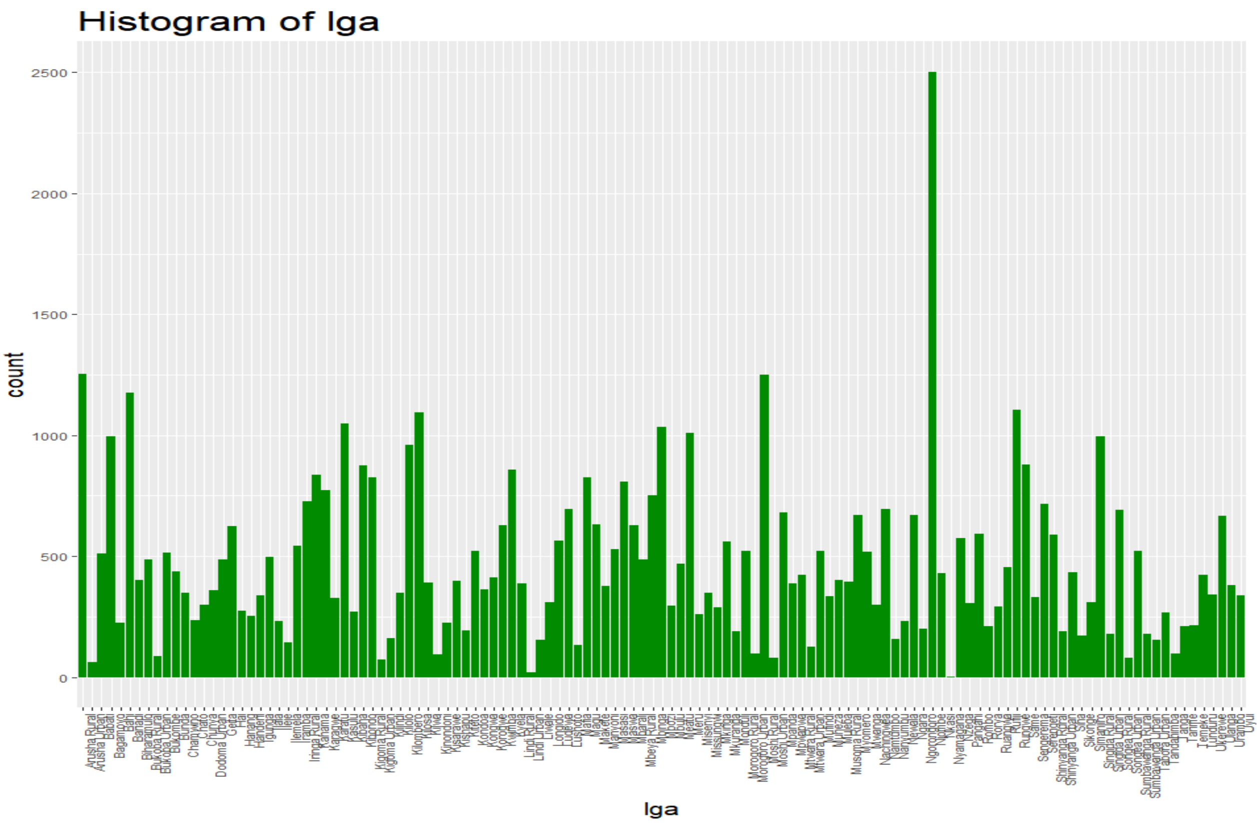
- amount_tsh – Total static head (amount water available to waterpoint)
- date_recorded – The date the row was entered
- funder – Who funded the well
- installer – Organization that installed the well
- wpt_name – Name of the waterpoint if there is one
- num_private –
- population – Population around the well
- public_meeting – True/False
- recorded_by – Group entering this row of data
- construction_year – Year the waterpoint was constructed

# Exploratory Data Analysis

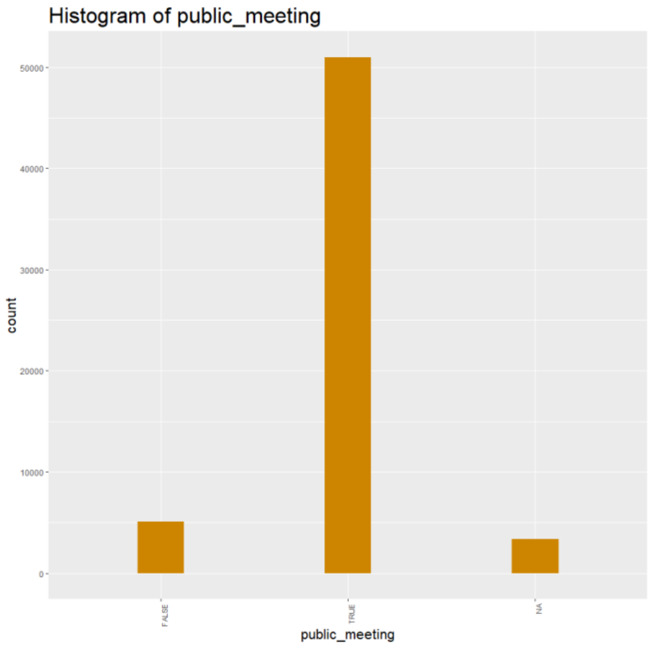
Various plots including scatter plots, box plots, bar plots, histograms, pivot tables have been created to get a good understanding of the data. This is an important precursor to the data wrangling step. A few important plots and all associated inferences are listed below.

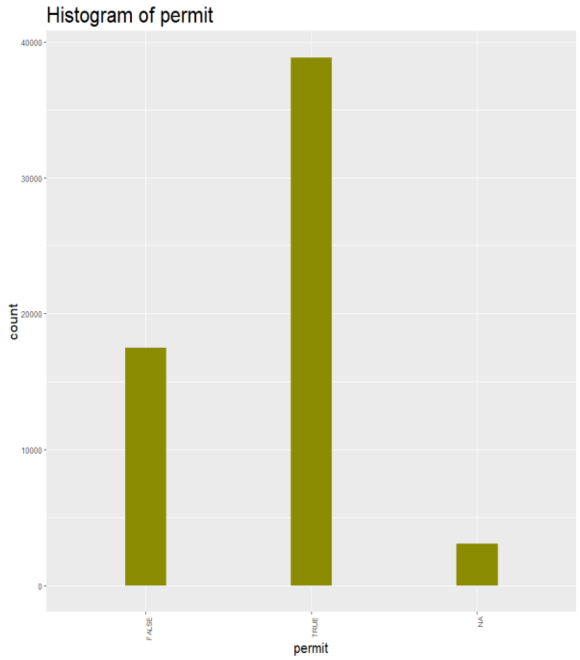
A histogram of pumps basin wise indicates that the pumps are spread across all the basin's and distribution of pumps varies among the basin's hence predictors basin and pump are important to consider during analysis. Similarly, a histogram of pumps geographic location (lga) wise indicates that the pumps are spread across geographic locations and vary in their count across locations. These indicate the need to consider the variables for analysis.



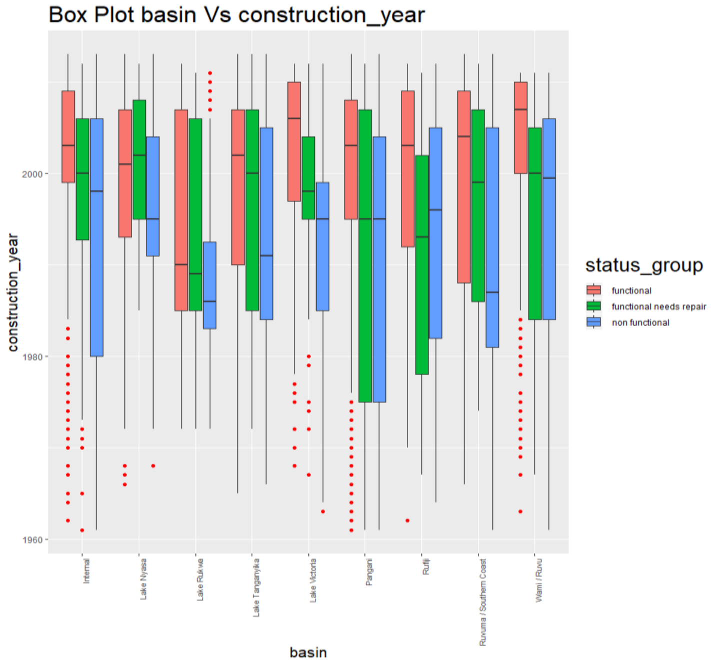


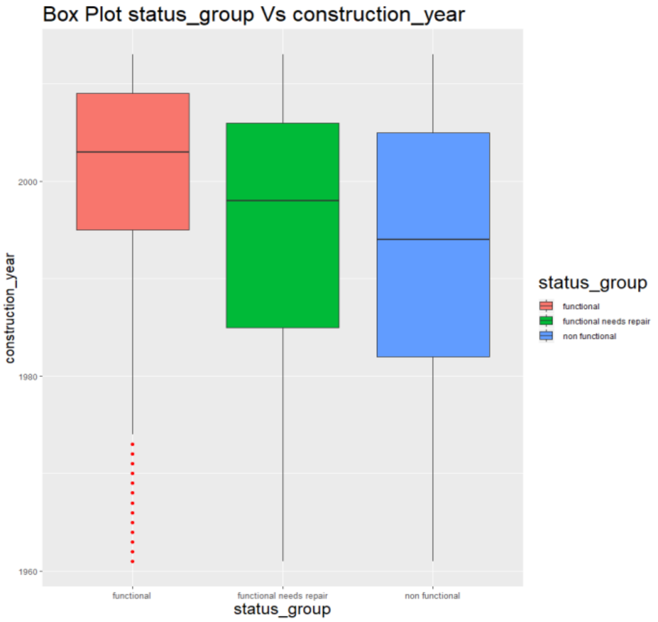
Histograms plots of variables public meeting and permit shown below not only indicate that they have significant amount of true  and false values in addition both have significant missing values (~ 4000 records each with missing values). This indicates necessity of handling these missing values either by imputation or plugging in requisite values that are appropriate for further analysis.

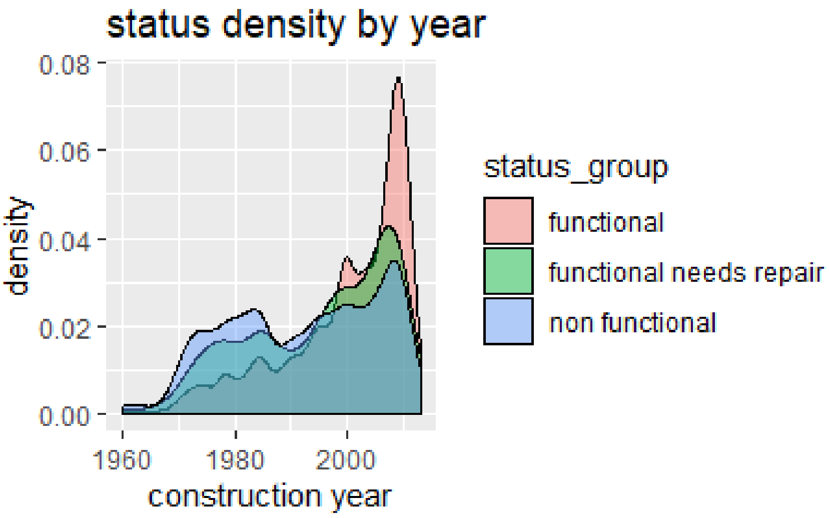




Box plots of basin versus construction_year and status_group versus construction_year status_group wise is shown below. These plots show vital information that all categories of status are prevelant in all the basins and majority of pumps that need repair and that are non functional are of construction year lesser than 1995.

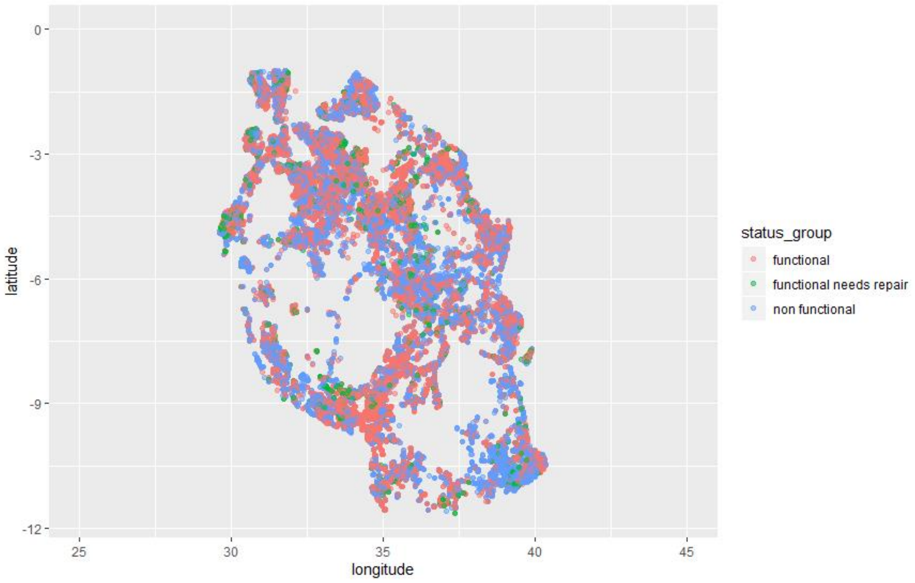




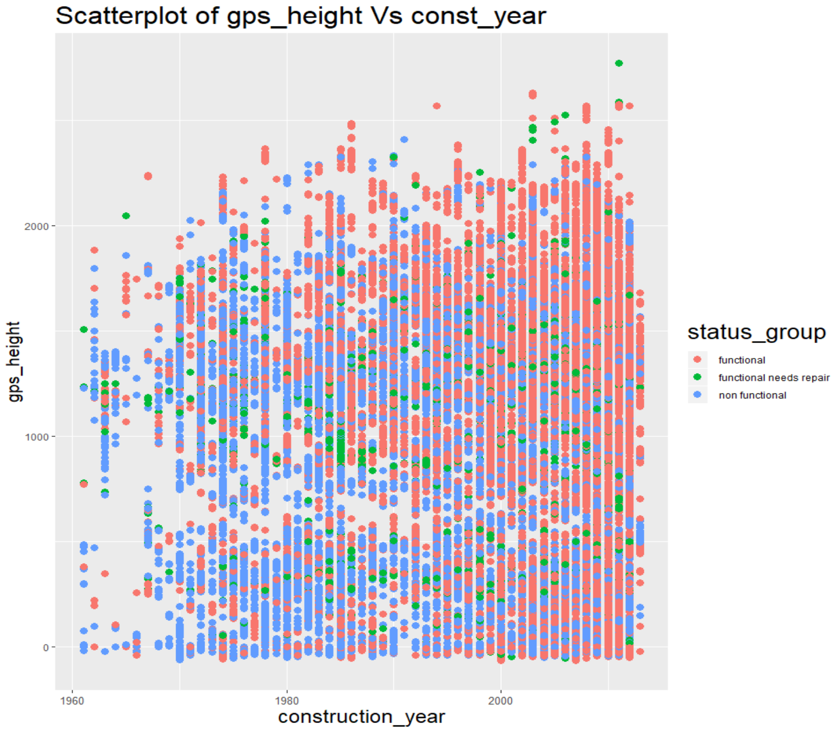


A density plot of construction year status group wise shows that the ratio of functional to non-functional pumps gets to be > 1 only after construction year 1998 while data set has values from the year 1960.

A plot of location of each water point with classification of status_group as shown below indicates that these are spread across and are significant variables to consider for analysis.



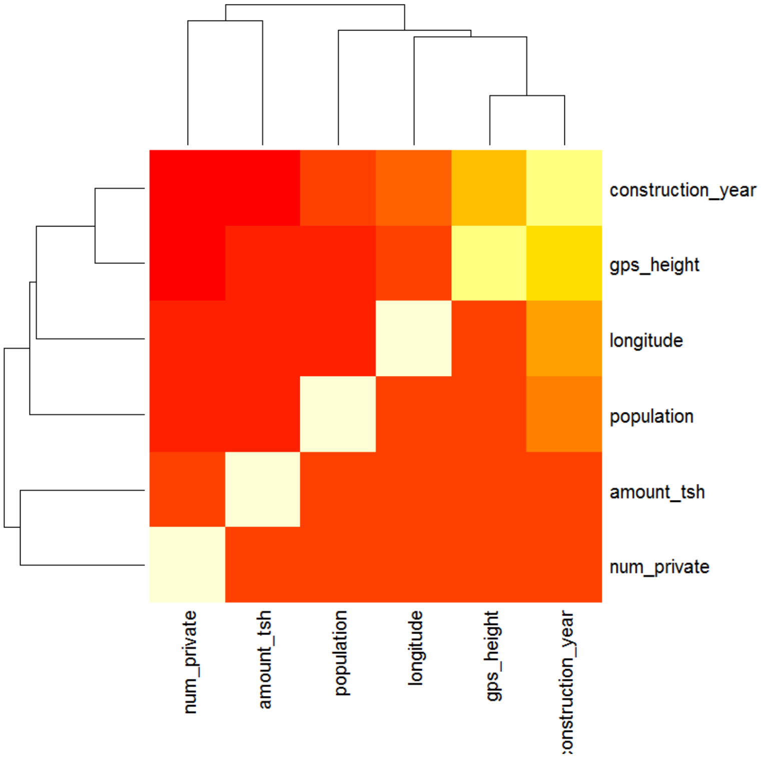
The visualization to the left represents location as compared to status_group. This shows the location of each well and gives an understanding of the density, relationship, scarcity, and density of each well location within the country. There are clearly locations that correlate to a particular status group.



A scatter plot of gps_height(altitude of the well) versus water point construction_year by status_group gives a clear picture that pumps with construction year greater than 2000 and with altitude of well greater than 1000 were more likely to be functional. This indicates that variables gps_height, construction year and pump status_group are correlated and need to be considered for analysis.

## Heatmap of Numeric Values.

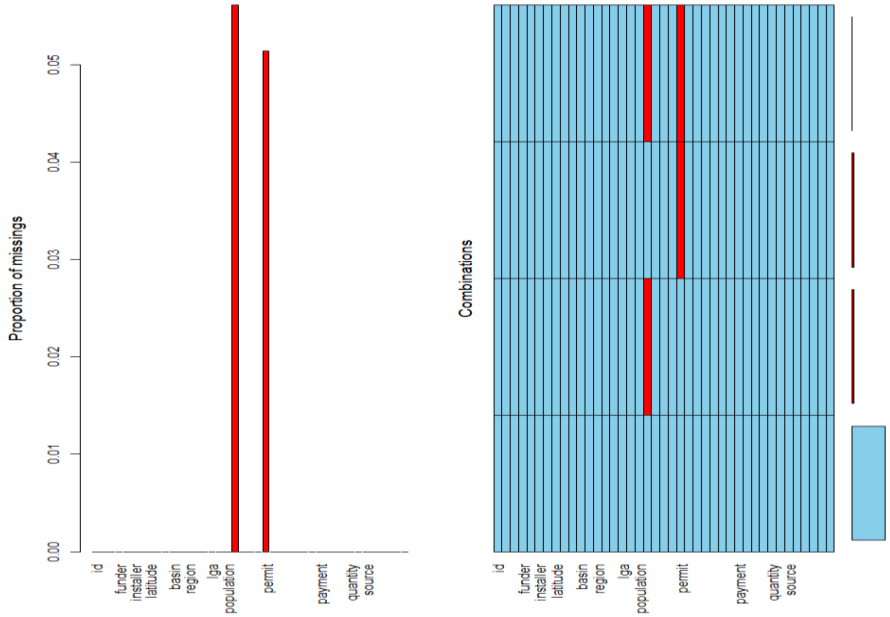
A heat map was generated for the variables of the data set to understand correlation between variables of data set. Significant correlations were observed only between construction year and GPS height. Slight correlations with construction year, longitude and population (probs not significant). A plot of generated heatmap is show below.



## Key Inferences

1) As construction year decreases the proportion of nonfunctional pumps increases.
2) Funder has an impact on the classification (Ex) HiFab has less that is non-functional.
3) Installer and funder are sparse attributes.
4) Instead of lga we can consider region because region is more clustered.
5) Water quality, Basin, Payment, Permit, Public meeting, Waterpoint_type_group are important attributes to consider.
6) Extraction type is already grouped as extraction type class and is useful.
7) Since payment and payment type are having similar data, we will consider payment
8) Quantity seems to be very promising attribute.
9) Scheme management seems to be relevant compared to scheme name.
10) Source type seems to be more relevant option compare to source class.
11) Population is an attribute to consider since this is a skewed measure, population range feature will need to be created instead of exact population.
12) Lump amount_tsh attribute to get a correlation.

The next step was to determine missingness of data. Details of missing data is as shown. Predictors public_meeting and permit had missing count and percentage as shown below.



A) Table indicating total missing values

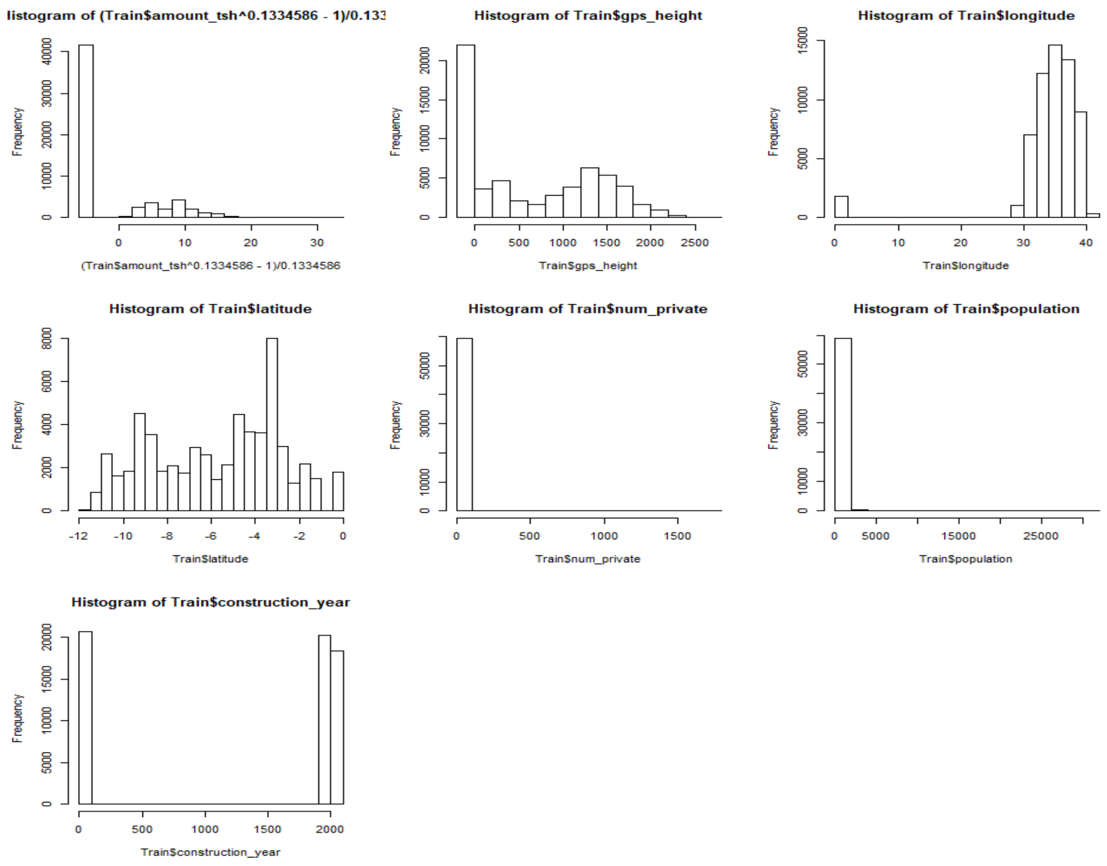
| Variable name | Count missing |
|---|---:|
| permit | 3056 |
| public_meeting | 3334 |

B) Table indicating missing values variable combination wise.

| Variable name | Count missing | Missing percentage |
|---|:---:|:---:|
| public_meeting | 3063 | 4.60% |
| permit | 2785 | 5.10% |
| public_meeting, permit | 271 | 0.45% |

Boxcox transformation of predictors and plots were plotted below. However, this did not drastically change the distribution of the numeric data.

| **Results of Box-Cox Transformation** | |
|---|---|
| Objective Name | PPCC |
| Data | Train$amount_tsh |
| Sample Size | 59400 |
| Bounds for Optimization | lower = -1,upper =  1 |
| Optimal Value | lambda = 0.1334586 |
| Value of Objective | PPCC = 0.8065708 |

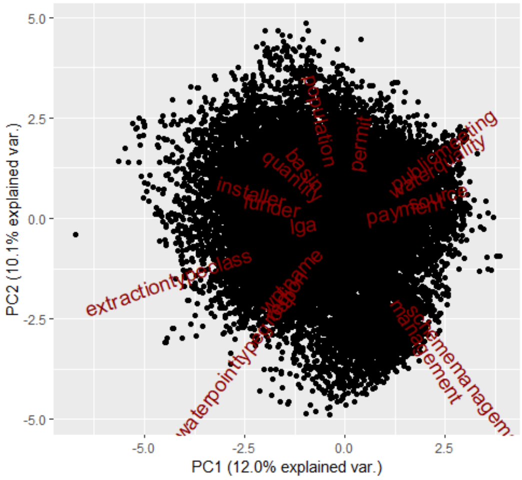


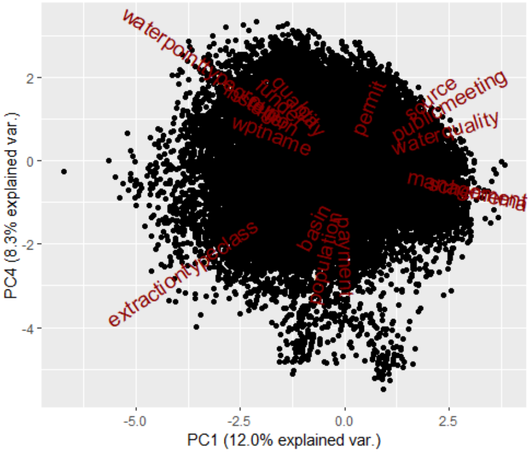
Principal Component Analysis was carried out and details of this are as mentioned below.

| **Description** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** | **PC8** | **PC9** | **PC10** |
|---|---|---|---|---|---|---|---|---|---|---|
| Standard deviation | 1.4294 | 1.3087 | 1.22838 | 1.1886 | 1.0893 | 1.04883 | 1.03168 | 1.01775 | 0.96513 | 0.93058 |
| Proportion of Variance | 0.1202 | 0.1008 | 0.08876 | 0.0831 | 0.0698 | 0.06471 | 0.06261 | 0.06093 | 0.05479 | 0.05094 |
| Cumulative Proportion | 0.1202 | 0.2209 | 0.30969 | 0.3928 | 0.4626 | 0.52731 | 0.58992 | 0.65085 | 0.70564 | 0.75658 |

| **Description** | **PC11** | **PC12** | **PC13** | **PC14** | **PC15** | **PC16** | **PC17** |
|---|---|---|---|---|---|---|---|
| Standard deviation | 0.91608 | 0.86902 | 0.83776 | 0.76642 | 0.70033 | 0.6624 | 0.57034 |
| Proportion of Variance | 0.04936 | 0.04442 | 0.04128 | 0.03455 | 0.02885 | 0.02581 | 0.01913 |
| Cumulative Proportion | 0.80594 | 0.85037 | 0.89165 | 0.9262 | 0.95505 | 0.98087 | 1 |

A biplot of principal components is as mentioned below. The below biplot shows PCA. It does not indicate any further condensation of features.





# Data Preparation

The data set provided has 40 variables, hence it was necessary to determine and select important and relevant variables for further analysis.

Every variable in the data set was reviewed. The first step was to check the variable names and replace '_' with '.' in the column names so that the models work without error messages. The next aim was to review all the variables to identify and eliminate sparse variables, variables with zero variance, identify duplicated variables i.e. those that had same information but in multiple variables. Such variables were identified and grouped into single variable.

Variables public_meeting, permit had true/false values these were converted to character variables. Variables ID and date_recorded were deemed to be not relevant predictors since they change for the test data. Variable amount_tsh was observed to have value of 0 for 80% of the data, hence it was decided not to consider it for further analysis.

Since variables funder, installer, wpt_name, had many levels "forcats" was used to lumping these to 20 levels. Variable gps_height which is a height measure was also grouped in steps of 300 variance. Variable population was lumped to 10 levels using "forcats".

Variables longitude, latitude was observed to be sparse, hence these were converted to integer, it was evident that precision would be compromised but the variables were kept intact.

The removal of irrelevant variables from the dataset is another important aspect of data preparation. Predictor variables that create noise within a model or create less efficient computation without providing insight are unnecessary. They have very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large. This type of variable considered "zero variace" was removed from consideration. The variables "gpsheight", "numprivate", "ward", and "recordedby" were removed from the dataset used to model train via this technique

Numerous variables in the data set were correlated, since these had similar information it was decided to consider a single variable from each of these sets for further analysis. Table below lists these similar variables and final variable that was selected from these to proceed further.

| Sl.no. | Similar variables | Final variable considered |
|---|---|---|
| 1 | extraction_type, extraction_type_class, extraction_type_group, extraction_type_class | extraction_type_class |
| 2 | management, management_group | management |
| 3 | waterpointtype, waterpointtypegroup | waterpointtypegroup |
| 4 | scheme_name, scheme_management | scheme_management |
| 5 | waterquality, waterqualitygroup | waterquality |
| 6 | payment and payment_type | payment |

Since the project's objective is to identify water points with pumps that need repair and those with non-functional pumps it was decided that the training set should be sampled in such a way that it has equal representation of the all the possible factor values. This would then train the model with all factors taken into consideration. Hence steps were put in place to select 4000 records for each status group factor value of functional, functional needs repair and non-functional. This training set was then used to train various model. The models tried and accuracy observed have been mentioned in the summary below.

An initial logistic glm model was used to understand the predictor significance, then identify the right set of predictors before training a more complex model. Predictors with low p-values were selected as predictors for random forest, decision tree, and SVM models.

**The final set of predictors are below:**

amount_tsh, gps_height, longitude, latitude, basin, population, construction_year, extraction_type_class, management_group, payment_type, quality_group, quantity_group, source_class, and waterpoint_type_group.

# Analysis Plan

The high degree of complexity and dimensionality of the water table dataset uncovered through initial exploratory data analysis was a great point of concern, when attempting to predict the class of the "status_group" variable. Attempting to train a model, even a simple penalized logistic regression model, would require immense computing power and would undoubtedly yield a model full of unnecessary noise. Utilizing the analytical base table that limited this noise through various data wrangling steps mentioned above, allowed computation more efficient and accurate.

The waterpump predictor variables were mostly of the character class. Both simple and complex modeling techniques would require factor class predictor variables to properly train the models. Additionally, identifying variables that have a manageable number of factor levels, and reducing the factor levels for variables that do not, will help reduce model complexity. The use of the "factor lump" function enabled the removal of any factor level that did not exceed 1% of that variable. The structure of the remaining dataset still had variables with a high number of levels. Due to the complexity a high number of levels would create in a model, when compared to the relative importance these variables would be in prediction. A significant number of variables were eliminated from the modeling process due to their complexity.

The possibility of three outcomes including "functional, "nonfunctional" and "repair" eliminated the possibility of a simple logistic regression model. However, a penalized logistic regression model includes the functionality to train a model to predict more than two outcomes. More complex models like Decision Trees and Random Forrest were selected to be used in order to take advantage the relatively small number of variables identified as extraordinarily significant. The low computational costs of limiting predictor variables and the factor levels within these variables increases model efficiency and allowed multiple iterations to optimize each model.

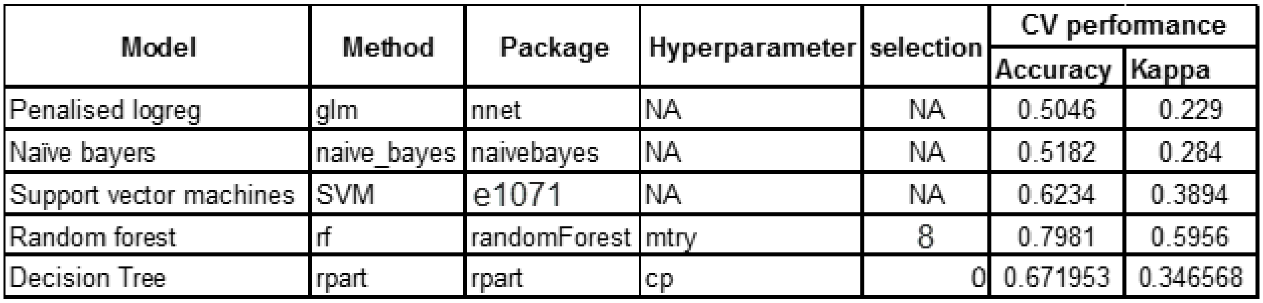
The penalized logistic regression model was established as the initial base model due to the penalty applied to the logistic model for having too many variables, essentially shrinking the coefficients of the less contributive variables toward zero (regularization). The hyperparameters used were kept constant at alpha = 1 and lambda = 0 in order to create a baseline in which to build upon.

The SVM algorithm was applied keeping in mind that it is one of the key algorithms to solve classification problems. Given this is not a problem that could be easily be classified using a linear hyper plane, polynomial and radial kernel types gave the best results. Both turned out close to 62% accuracy with the uniform size random sampling. Given the random forests was returning better results, there was no further hyper-parameter tuning that was performed on this model apart from the various kernel selection. Linear kernel type gave an accuracy of 51% and Sigmoid kernel type gave an accuracy of 35%

The Decision Tree and Random Forrest Model were selected to take advantage of the very stringent feature selection criteria for predictor variables to be included in the model training function. Each of these models typically take significant computing power and time, but due to up front data manipulation and analysis, each model ran quite efficiently. This allowed time to properly optimize the decision tree model using the cp hyperparameter to regulate the complexity of the model. The mtry hyperparameter, number of variables randomly sampled as candidates at each split, was optimized in the Random Forrest Model. It was optimized at 8, but given more time or higher computing power, additional random splits could have been made to continue to increase the accuracy of the model.

# Results and Validation

Accuracy, the measurement of percentage correct classifications out of all instances is the performance evaluation technique used to optimize, improve, and compare each model. The Accuracy and Kappa measurements validate the usefulness of each model respectively. As seen below, the Random Forrest model was the most accurate, most likely due to data preparation and limiting the complexity of the predictor variables that the random forest model was based upon.



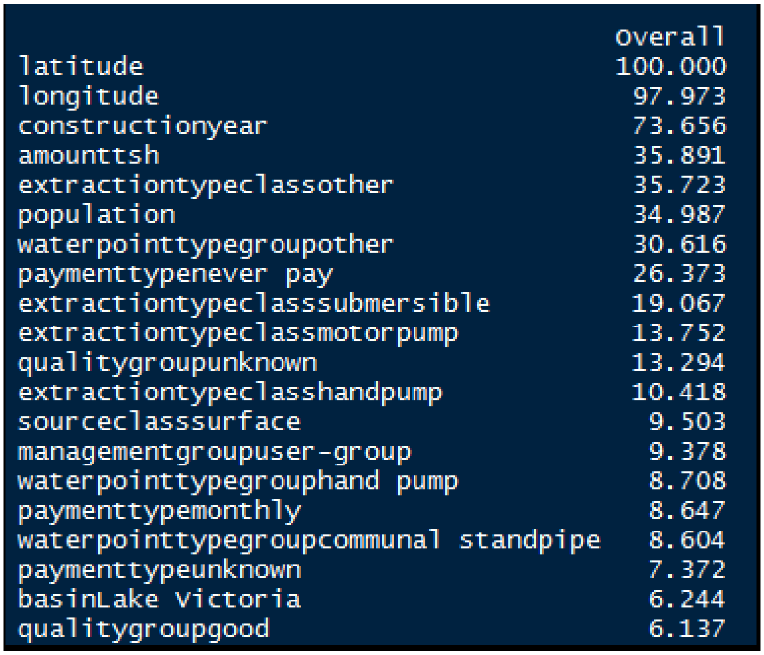
| Model | Method | Package | Hyperparameter | selection | CV performance | |
|---|---|---|---|---|---|---|
| | | | | | Accuracy | Kappa |
| Penalised logreg | glm | nnet | NA | NA | 0.5046 | 0.229 |
| Naïve bayers | naive_bayes | naivebayes | NA | NA | 0.5182 | 0.284 |
| Support vector machines | SVM | e1071 | NA | NA | 0.6234 | 0.3894 |
| Random forest | rf | randomForest | mtry | 8 | 0.7981 | 0.5956 |
| Decision Tree | rpart | rpart | cp | 0 | 0.671953 | 0.346568 |

The logistic regression and classification trees performed in a very similar range. There is a bit more separation on the Kappa metric between the Tree models and the Logistic Models. The Tree model's faired quite a bit better when Kappa was measured, which considers the possibility that a correctness could happen by chance.

The Random Forrest Predictive model produced the most accurate model by far by utilizing the below predictive formula:

Statusgroup ~ amount_tsh + gps_height + longitude + latitude + basin + population + construction_year + extraction_type_class + management_group + payment_type + quality_group + quantity_group + source_class + and waterpoint_type_group

The coinciding importance of these variables is seen below, calculated via the “VarImp” function in the caret library. This graphic displays the top 20 most “important” variables. Variable importance evaluation functions can be separated into two groups: those that use the model information and those that do not. This list contains the top 20 most model affecting variables. The advantage of using a model-based approach is that is more closely tied to the model performance and that it may be able to incorporate the correlation structure between the predictors into the importance calculation

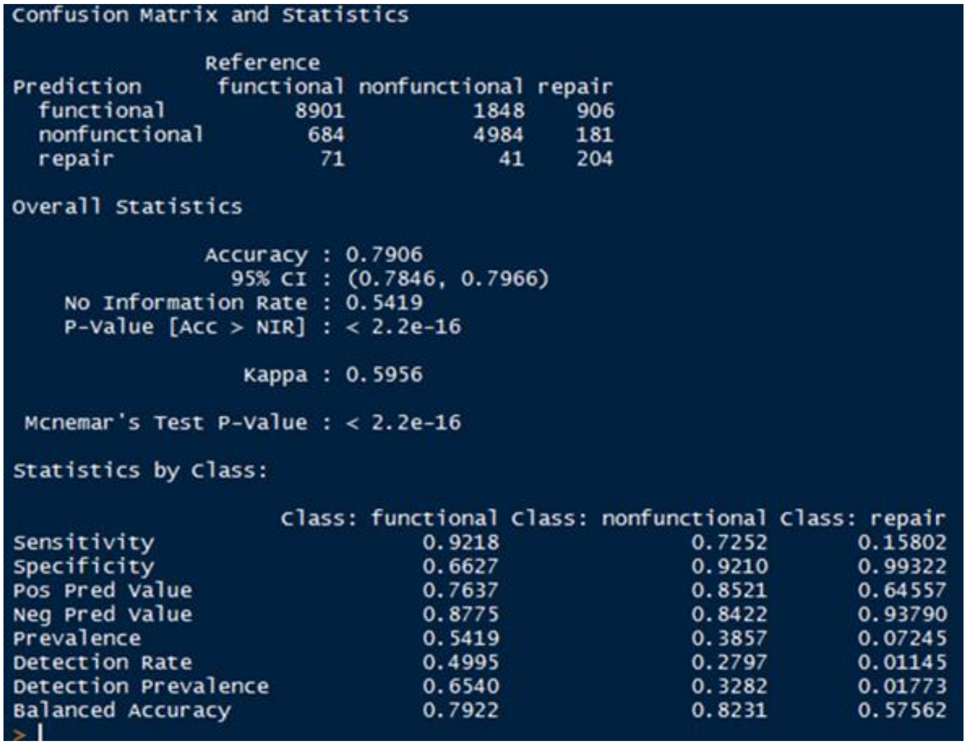


```
                                    Overall
latitude                            100.000
longitude                            97.973
constructionyear                     73.656
amounttsh                            35.891
extractiontypeclassother             35.723
population                           34.987
waterpointtypegroupother             30.616
paymenttypenever pay                 26.373
extractiontypeclasssubmersible       19.067
extractiontypeclassmotorpump         13.752
qualitygroupunknown                  13.294
extractiontypeclasshandpump          10.418
sourceclasssurface                    9.503
managementgroupuser-group             9.378
waterpointtypegrouphand pump          8.708
paymenttypemonthly                    8.647
waterpointtypegroupcommunal standpipe 8.604
paymenttypeunknown                    7.372
basinLake Victoria                    6.244
qualitygroupgood                      6.137
```

Takeaways from this graphic are that geographic location (latitude and longitude) were of increasing importance. The exact and precise location allowed models to pinpoint locations, rather than training based upon categorization. This freed the model from being “hand-cuffed” by factor levels. Construction year, specifically after 1960 was a leading indicator of functional waterpoints. It is believed that construction methods improved overtime, and older water points decayed overtime. Further, amount_tsh is an indicator of the amount of water flowing through each point.

Additionally, group specific variables played a key role in model training. The quantity group factor level “enough” indicated a functional water point, but “seasonal” factor level was an indication of needing repair, while the “insufficient” factor level indicated a non-functional point. Each group played a pivotal and unique role in classifying the water point as one of the group status variables, this lent itself to the success of random forest modeling technique quite well.

The results of random forest model can be seen below. This model was created from 70% of the total dataset and validated against 30%. 79.06% accuracy is quite good, especially compared to the 54.19% No Information Rate. This represents a nearly 25% increase in correct classification as compared to only knowing the distribution of the dataset. The kappa value at 59.56%, which is observed value with expected random chance, is well over 50%.

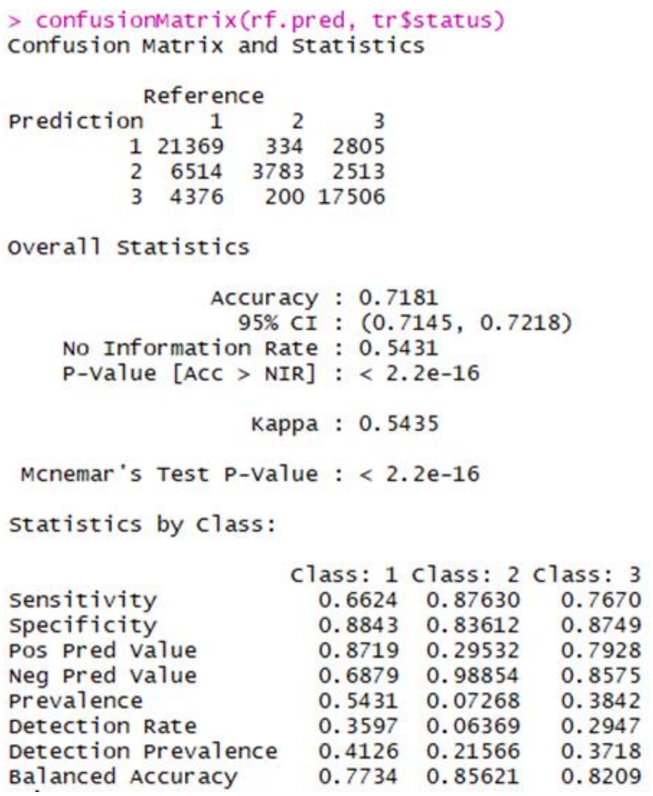


```
Confusion Matrix and Statistics

               Reference
Prediction      functional nonfunctional repair
  functional          8901          1848    906
  nonfunctional        684          4984    181
  repair                71            41    204

Overall Statistics

               Accuracy : 0.7906
                 95% CI : (0.7846, 0.7966)
    No Information Rate : 0.5419
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5956

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: functional Class: nonfunctional Class: repair
Sensitivity                     0.9218               0.7252       0.15802
Specificity                     0.6627               0.9210       0.99322
Pos Pred Value                  0.7637               0.8521       0.64557
Neg Pred Value                  0.8775               0.8422       0.93790
Prevalence                      0.5419               0.3857       0.07245
Detection Rate                  0.4995               0.2797       0.01145
Detection Prevalence            0.6540               0.3282       0.01773
Balanced Accuracy               0.7922               0.8231       0.57562
> 
```

However, the sensitivity, detection prevalence, and balanced accuracy of the class “repair” and to a lesser extent “functional” is lacking.  Sensitivity is the true positive rate or recall. It represents the percentage of instances from the first positive class that were predicted correctly. This is a very poor rate for the nonfunctional and repair classes. Specificity which is essentially a measurement of negative instance that were predicted correctly. This is better, but seemingly provides stronger evidence that the model created, while highly accurate overall, lacks predictive power on the status group classes that matter the most. The repair and non-functional class identification are more important so that they can be fixed. A false negative in the functional calls is not a big deal as it simply means the water point is working. However, misclassifying the non-functional or repair class could mean we think water points are producing water to people in need when it is not. A solution to this failure in the model is to increase the proportion of repair and nonfunctional instances within the training data.

In order to prepare a uniform size random sample training data with capturing set of very few records for each water pump status would mean decreasing the overall model accuracy. The confusion matrix results can be seen below. Cleary a decrease in accuracy is not ideal, but the drastic increase in sensitivity and specificity of the "non-functional" and "repair" classes make this trade-off worth it.

This uniform size random sample model was necessary because the total percentage of data where the water point status is 'functional needs repair' is less than 10%. To be able to uniformly sample and then train the model with data from all the statuses would mean that, there is only 15% of the total data that is training worthy (Since the remaining 5% of functional needs repair is needed for testing the model). This impacted the overall model efficiency by a steep 10% and kind of undermined the predictive power with respect to the functional and non-functional wells.



```
> confusionMatrix(rf.pred, tr$status)
Confusion Matrix and Statistics

          Reference
Prediction     1     2     3
         1 21369   334  2805
         2  6514  3783  2513
         3  4376   200 17506

Overall Statistics

               Accuracy : 0.7181
                 95% CI : (0.7145, 0.7218)
    No Information Rate : 0.5431
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5435

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.6624  0.87630   0.7670
Specificity            0.8843  0.83612   0.8749
Pos Pred Value         0.8719  0.29532   0.7928
Neg Pred Value         0.6879  0.98854   0.8575
Prevalence             0.5431  0.07268   0.3842
Detection Rate         0.3597  0.06369   0.2947
Detection Prevalence   0.4126  0.21566   0.3718
Balanced Accuracy      0.7734  0.85621   0.8209
```

Though, sub-optimal, it was necessary to keep the promise of predicting the non-functional and repair necessary water points intact. It would highly help if the 'functional needs repair' water points are also considered 'non-functional' making this a binary classification problem. Though, it means higher severity reporting condition for certain water points, it would create an ideal mix of data for training, testing as well as makes modeling and analysis much simpler. To not meddle with the original problem statement, that option to make 'functional needs repair' as 'non-functional' was not considered.

All things considered, with skewed training data, numerous features, data that ranges a wide number of years going back five and a half decades, the accuracy of prediction, though needs further work, of 70% is a great start.

# Conclusion

The models that were initially developed were bettered considerably which is evident with the spike in accuracy by close to 12%. This was achieved with better feature selection and feature engineering. Removal of predictors with too much or too little variance greatly helped improving model accuracy. Construction year, Location attributes, various classifiers have created an ideal data set for any classification algorithm. The source data set was heavily skewed in terms of status because it mirrors the situation in that country that there are more functional water points than the non-functional ones. Hence, the initial models had the utmost predictive accuracy when it came to functional water points.

It is convincing to see how flexible the models are, with the two different sampling techniques that have been used to achieve overall better accuracy and balanced accuracy of the various result outcomes. Random forests modeling provided the most accurate classification. With a simple 70% for training and 30% for testing strategy, model has had a very good accuracy but did not predict the non-functional and the functional needs repair wells accurately (In fact, functional needs repair had an accuracy of 18%). False positive are far worse than false negatives and considering this, a uniform sampling method to take the equal number of pumps across statuses helped improving the accuracy rates across statuses but decreased the overall accuracy. Since the intent is to the get to the non-functional as well as the ones that need repairs, the uniform sampling data trained model is the best one is the conclusion.

# Future Work

Equipment failure/maintenance prediction is one of the challenging tasks in all industries. Future work objective is to further generalize this work to better predict equipment failures, equipment of different types and from different conditions before it happens. For example, from the experience of one of the project developers, there is a challenge for detecting equipment failure in Oil and gas industry. Those are high cost equipment and at the same time will cause a halt of production and result in immediate expensive work orders to fix them. This would entail needing different set of routines written to modularize the work for different kinds of equipment.

The current study is a prediction of current status and lacks the insight of future status of the water pumps. The at-the moment status prediction can be augmented to model the future predictability of the water point status by incorporating analysis methods associated with time series data. To further improvise modeling strategy, keeping in mind all the forward-looking thoughts discussed above, it would be beneficial to use boosted random forests and neural networks.

# References

1. Problem Statement and the Training/Test Data - https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/
2. Max Kuhn, Kjell Johnson (2013). “Applied Predictive Modeling” (Springer)