Capstone Project 1 - Milestone Report

Overview

Taarifa is an open source platform for crowd sourced reporting and triaging infrastructure problems in Tanzania. It aggregates information from sms, webforums, emails, and twitter from citizens, placed it in a workflow that triaged the problems and notify local authorities that can deal with the problems. Its goal is to function as a communication channel between the citizens who have first-hand perspectives to local problems and decision makers who have the power to fix those problems. Using the data from Taarifa and the Tanzanian Ministry of Water, we are asked to predict whether a waterpump, or waterpoint, is functional, needs repair, or non-functional. Having the ability to predict which waterpoint will fail or need repair can improve maintenance operations of waterpoints and ensure clean and potable water is available to the people across Tanzania.

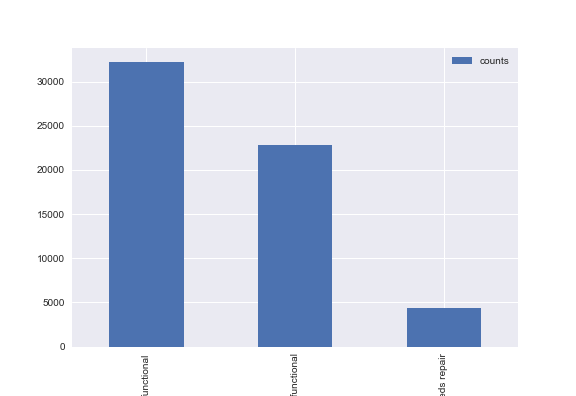
The dataset includes the following 39 features of 59400 water pumps in Tanzania:

* amount\_tsh - Total static head (amount water available to waterpoint)
* date\_recorded - The date the row was entered
* funder - Who funded the well
* gps\_height - Altitude of the well
* installer - Organization that installed the well
* longitude - GPS coordinate
* latitude - GPS coordinate
* wpt\_name - Name of the waterpoint if there is one
* num\_private -
* 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
* population - Population around the well
* public\_meeting - True/False
* recorded\_by - Group entering this row of data
* scheme\_management - Who operates the waterpoint
* scheme\_name - Who operates the waterpoint
* permit - If the waterpoint is permitted
* construction\_year - Year the waterpoint was constructed
* 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 - How the waterpoint is managed
* management\_group - How the waterpoint is managed
* payment - What the water costs
* payment\_type - What the water costs
* 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 - 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

By using the predicting variables above, the goal is to assign one of the three operating conditions of each waterpoints:

* functional - the waterpoint is operational and there are no repairs needed
* functional needs repair - the waterpoint is operational, but needs repairs
* non functional - the waterpoint is not operational

The operationing conditions of the dataset broke down as followed with 54% of the waterpoints being “functional”, 38% “non functional”, and 7% “functional but needs repair”.



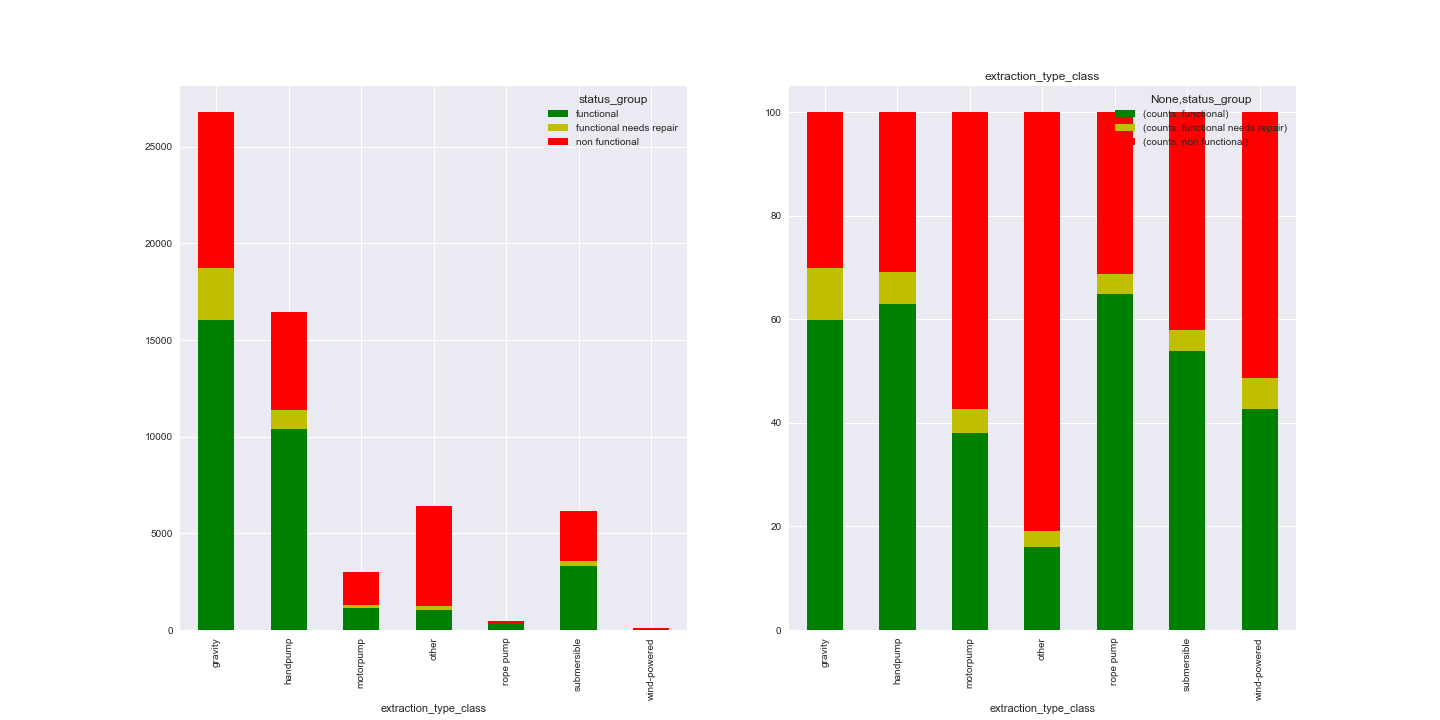
Missing Data.

There were missing data in the following columns within the dataset:

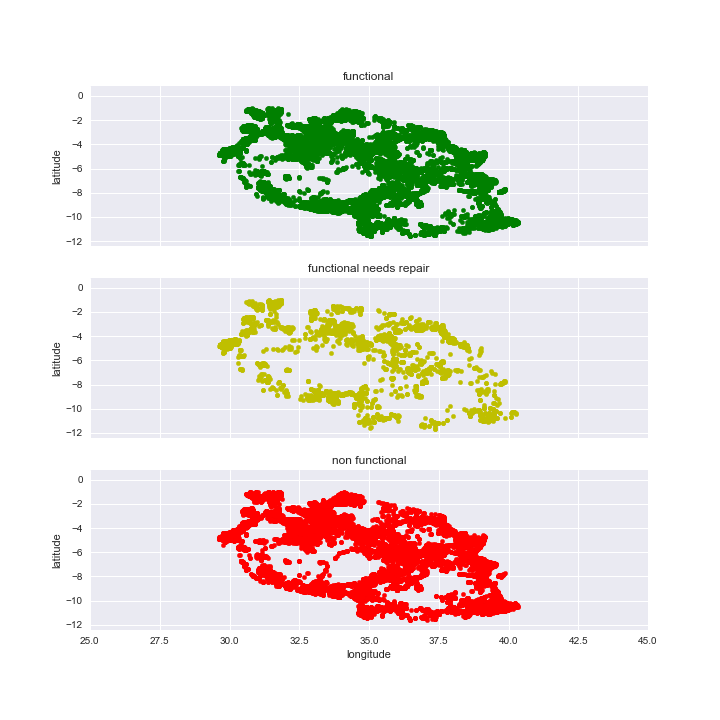
|  |  |  |
| --- | --- | --- |
| **Features** | **# of missing values** | |
| funder | 3635 |
| installer | 3655 |
| subvillage | 371 |
| public\_meeting | 3334 |
| scheme\_management | 3877 |
| scheme\_name | 28166 |
| permit | 3056 |

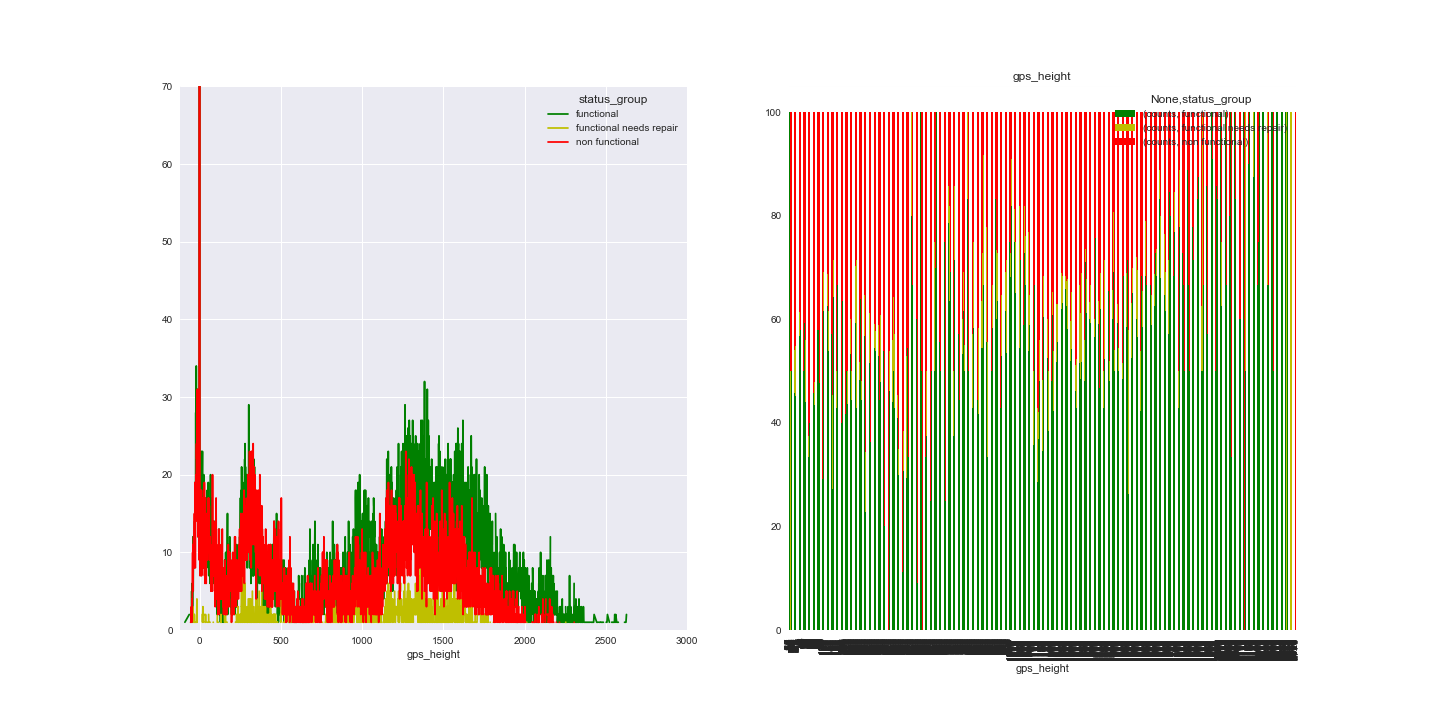
The missing data was imputed by replacing with the value “missing”, this addresses if the “missing” data contributes to the condition of the waterpoints.

To identify if and how each of the features correlates to their operational conditions, each feature was plotted with stacked bar chart illustrating the proportion of functional, non functional, and functional needs repair in count number and percentages.

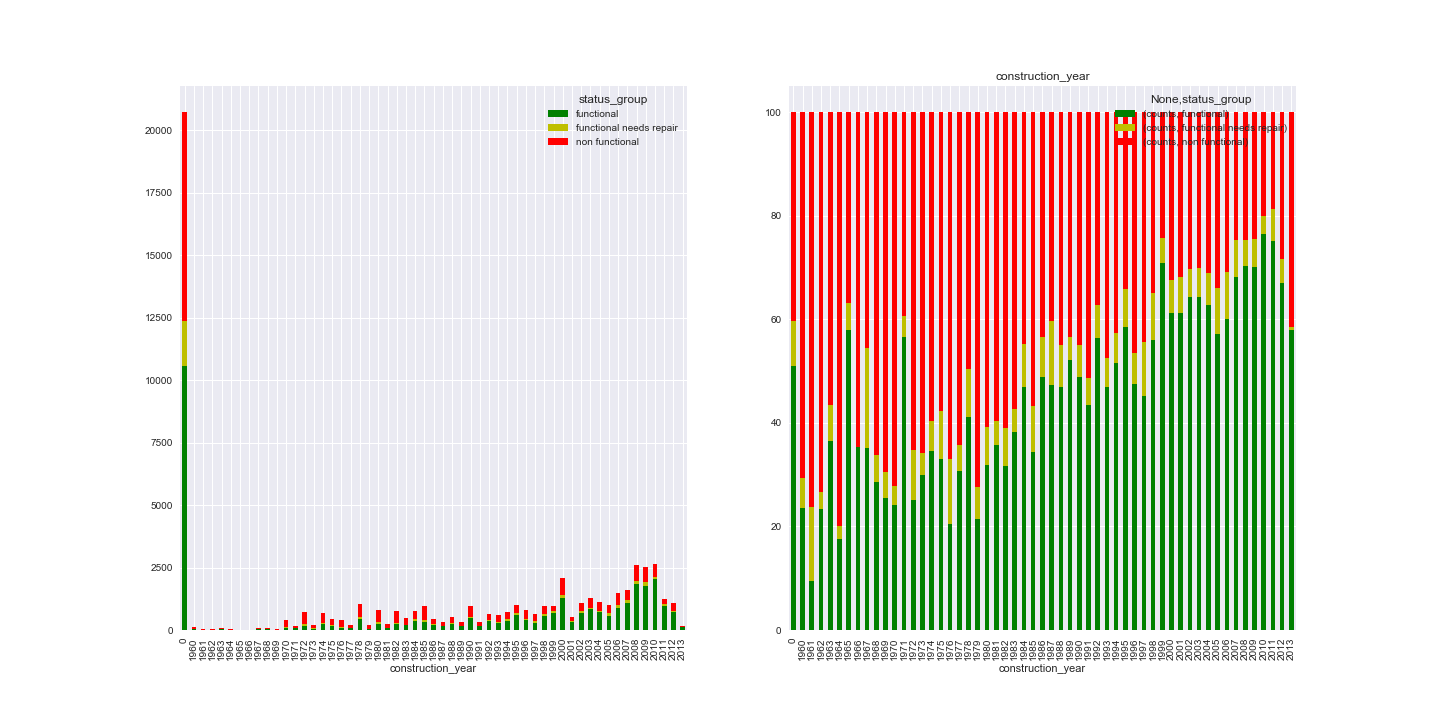


There doesn’t seemed to be any correlation between the longitude /latitude data of the waterpoints to the condition of the waterpoints, as shown below. The scatter for the three categories seemed to be very similar. However, there may be some effects on gps height data, as there seemed to be a lower proportion of non functional waterpoints in higher elevations. But this could be effected by the fact that higher elevations maybe be harder to get to and thus there would be less report of status in higher elevation.



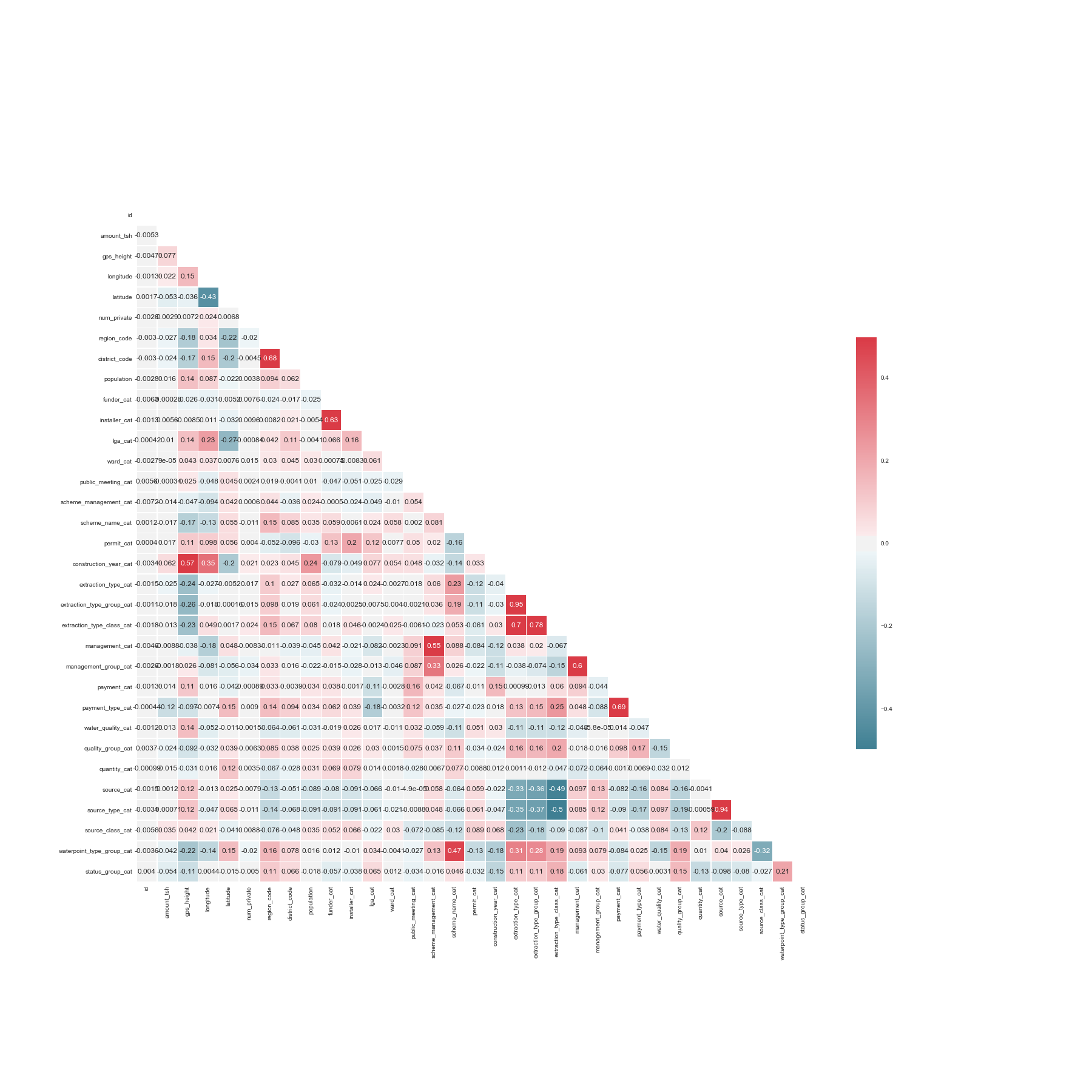


The only feature that show obvious correlation to the condition of the waterpoints is the construction date. The following figure shows that the older waterpoints clearly have a larger proportion of the non functional well compare to the newer ones.



Correlation Matrix

In order to have an overview to compare how all the features correlate to the operational status. I used label encoding to convert each of the categorical features to numerical values. This allowed for constructing a correlation matrix as followed:



Eliminating features.

The following features were eliminated from the dataframe for analysis:

1. recorded\_by and date\_recorded, because there’s only 1 unique value throughout the dataset.
2. Wpt\_name and subvillage had 37400 and 19288 unique values, respectively, for a dataset of 59000 records. They do not have enough predictive power.
3. Waterpoint\_type and quantity\_group have equivalent columns that include the same information in waterpoint\_type\_group and quantity, respectively.

The majority of the features within this dataset are text based categorical data. In order to use the various model fitting methods, I used one hot encoding to numerate the dataset and proceeded to use various modeling methods. The metrics I used to compare the different methods were the accuracy score and confusion matrix.

Accuracy Score:

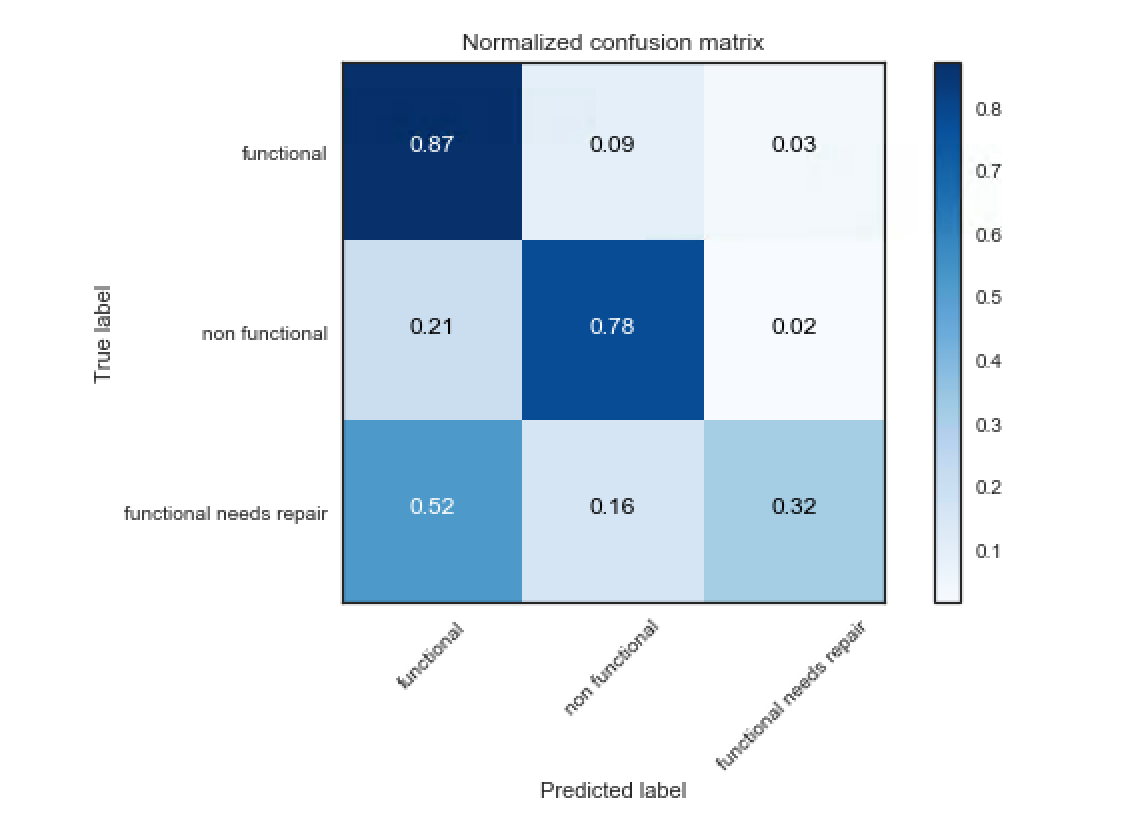
|  |  |
| --- | --- |
| **Logistic Regression** | |
| **C** | **Accuracy** |
| 0.001 | 0.707 |
| 0.1 | 0.709 |
| 1 | 0.718 |
| 10 | 0.706 |
| 100 | 0.708 |
| 1000 | 0.708 |
| 10000 | 0.706 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random Forest** | | | | |
| **Leaf** | **n\_estimators** | **Accuracy Score** | | |
| 1 | 100 | 0.798316498 | | |
| 50 | 100 | 0.697979798 | | |
| 500 | 100 | 0.618350168 | | |
| 1500 | 100 | 0.558417508 | | |
| 1 | 500 | 0.799242424 | | |
| 1 | 2500 | 0.801515152 | | |
| 1 | 5000 | 0.800925926 | | |
|  |  | **Accuracy** | | |
| **Gaussian Naïve Bayes** | | | 0.44 |
| **Multinomial Naïve Bayes** | | | 0.45 |

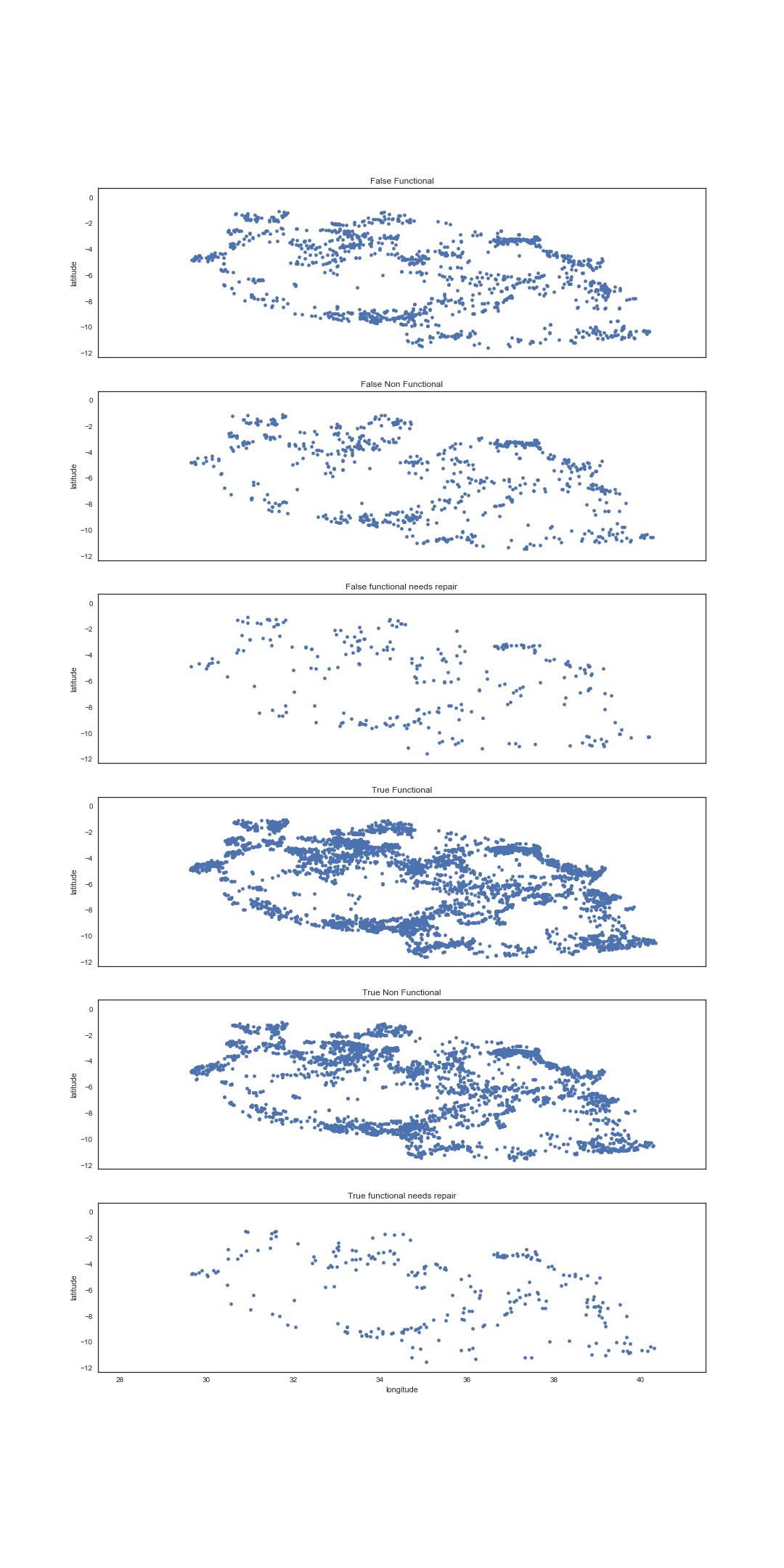
By looking at other performance measures, it is obvious that random forest was the best model for predicting operation condition of the waterwell.

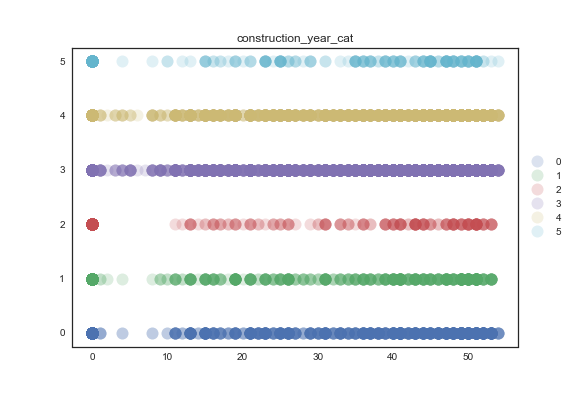
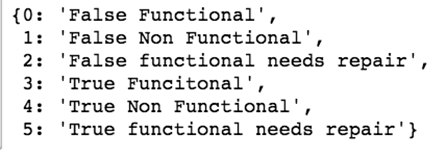
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Accuracy** | **Precision** | **Recall** | **F1** |
| **Logistic Regression** | | 0.71 | 0.68 | 0.71 | 0.68 |
| **Random Forest** | | 0.80 | 0.79 | 0.80 | 0.79 |
| **Gaussian Naïve Bayes** | | 0.44 | 0.73 | 0.44 | 0.50 |
| **Multinomial Naïve Bayes** | | 0.45 | 0.60 | 0.45 | 0.38 |

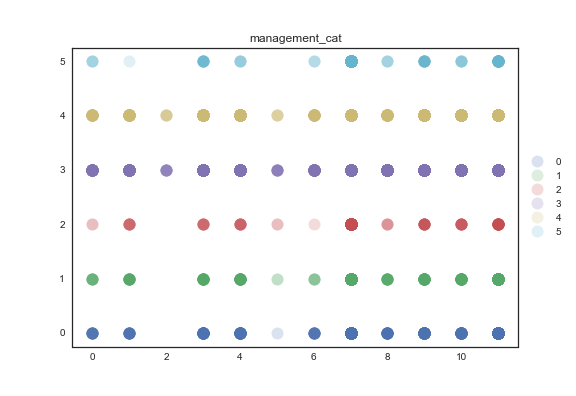
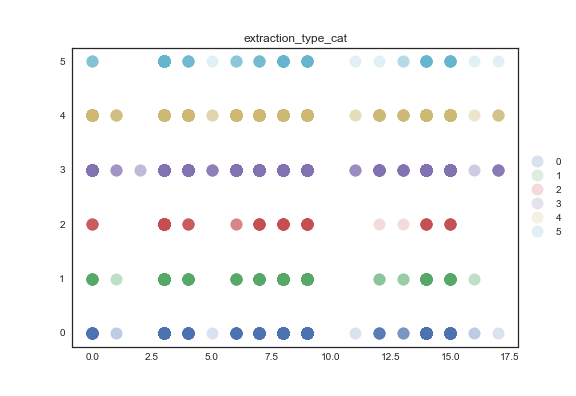
When using a confusion matrix to look into the prediction accuracy, it shows it did the worst at predicting those wells that are “functional but needs repair”.

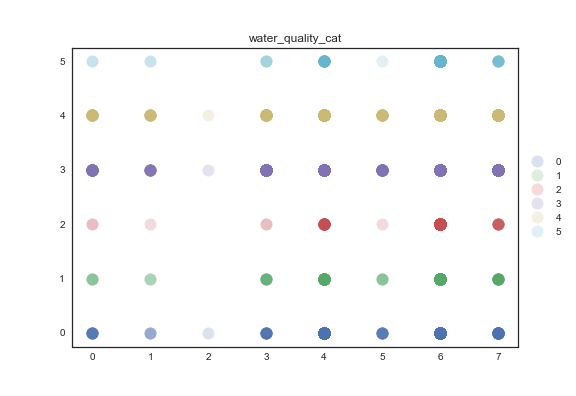


So in trying to see if there’s something unique or noticeable about the misclassified “functional needs repair”, I tried to plot the dataset by the classification score of the predicted value: true or false, functional/non functional/funcation needs repair.







a