An Algorithm for Predicting the Functionality of Water Mines in Tanzania

Final report prepared for the
Data Analysis and Interpretation Specialization

June 11, 2017

Introduction

This capstone research project is about "data mining the water table." It is a competition from the website *DrivenData*. The purpose of this research is to use all the given information about the water mines and predict their functionality, among functional, not functional and functional but needs repair.

It is important, for me, to identify how each predictor plays its role in predicting the results. Having a better understanding of factors that are most likely to make an impact on the functionality of a water mine will allow me to identify which factors to focus on in order to predict the functionality of water mines.

A more precise prediction of the functionality of water mines could help people to improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Methods

Sample

The sample was drawn from the data provided by Taarifa and the Tanzanian Ministry of Water. This sample measures aspects of a water point and the goal is to use the data to predict the functionality of each water point. The initial sample size is 59,400. I choose all the quantitative variables, and one categorical variable public_meeting. Because these variables have more connection to the functionality of water points. The result training set gives 36,385 samples. This sample measures aspects of a water point and the goal is to use the data to predict the functionality of each water mine.

Measure

All the variables used to predict status_group is shown below:

amount_tsh - Total static head (amount water available to waterpoint)
gps_height - Altitude of the well
longitude - GPS coordinate
latitude - GPS coordinate
region_code - Geographic location (coded)
district_code - Geographic location (coded)
population - Population around the well
public_meeting - True/False
construction_year - Year the waterpoint was constructed

After choosing these variables, I drop all the data with nan value or have no record on the column construction year. Besides, I computed the number of years from the water points were established to replace the original data in construction year. Also, I assign a number to each leve in the response variable "status_group": 2 means that the water mines are functional, 1 means that the water mines are functional but need repair and 0 means that the water mines are not functional. This could help fit into the model later. In addition, I transform the True and False label in the variable "public_meeting" to 1 and 0, respectively, as well. The above changes give me a dataset full of real values, and I can directly start to do analysis and perform model on it.

Analysis

I first examine the data. It is easy to get an idea about the data by examining frequency tables for categorical variables and calculating the mean, standard deviation and minimum and maximum values for quantitative vari-

ables.

Then I run the Analysis of variance (ANOVA) for each predictor in order to get a distribution of values of each predictor for each level of status_group.

Since the response variable is a three-level categorical variable, I decided to use three methods to predict: Decision Tree, Random Forest and Adaboost. Each method is suitable in this situation. I split the data to get a train set as 60% (n = 21831) of the whole data and the rest 40% (n = 14554) are test set. After fitting the model, I use the test set to evaluate the accuracy of each model. The accuracy of the model is determined by the proportion of test observations predicted right compared to the true label. Comparing the accuracy scores of all three models, I can determine which model to use for the given test set, which is what the competition evaluates on.

Results

Descriptive Analysis

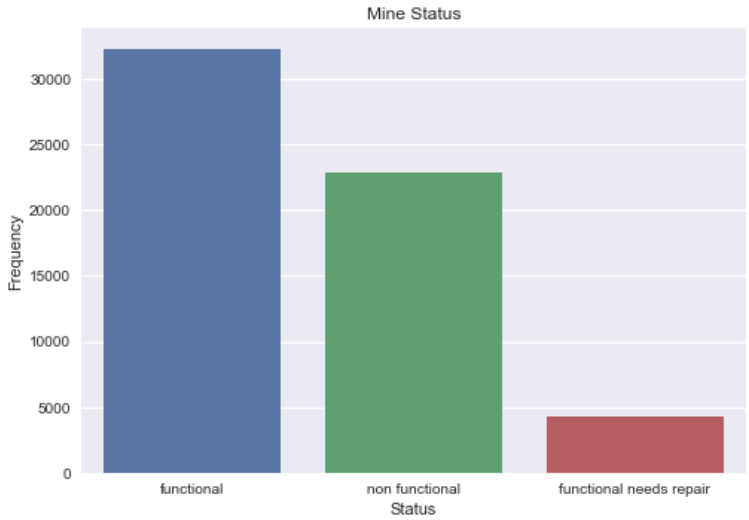
The following output shows the descriptive statistics for all predictor and response variables.

	amount_tsh	gps_height	longitude	latitude	region_code
count	36385.000000	36385.000000	36385.000000	36385.000000	36385.00000
mean	489.134731	1006.627731	35.995127	-6.247756	15.46181
std	3629.425988	619.139926	2.576249	2.690373	21.02849
min	0.000000	-63.000000	29.607122	-11.649440	2.00000
25%	0.000000	372.000000	34.703445	-8.673177	4.00000
50%	0.000000	1147.000000	36.680745	-6.015980	10.00000
75%	200.000000	1495.000000	37.778425	-3.716346	16.00000
max	350000.000000	2770.000000	40.345193	-1.042375	99.00000

	district_code	population	public_meeting	construction_year
count	36385.000000	36385.000000	36385.000000	36385.000000
mean	6.021575	269.430672	0.915212	20.293720
std	10.766659	559.950280	0.278569	12.497418
min	1.000000	0.000000	0.000000	4.000000
25%	2.000000	25.000000	1.000000	9.000000
50%	3.000000	150.000000	1.000000	17.000000
75%	5.000000	300.000000	1.000000	30.000000
max	63.000000	30500.000000	1.000000	57.000000

	status_group
count	36385.000000
mean	1.200357
std	0.945226
min	0.000000
25%	0.000000
50%	2.000000
75%	2.000000
max	2.000000

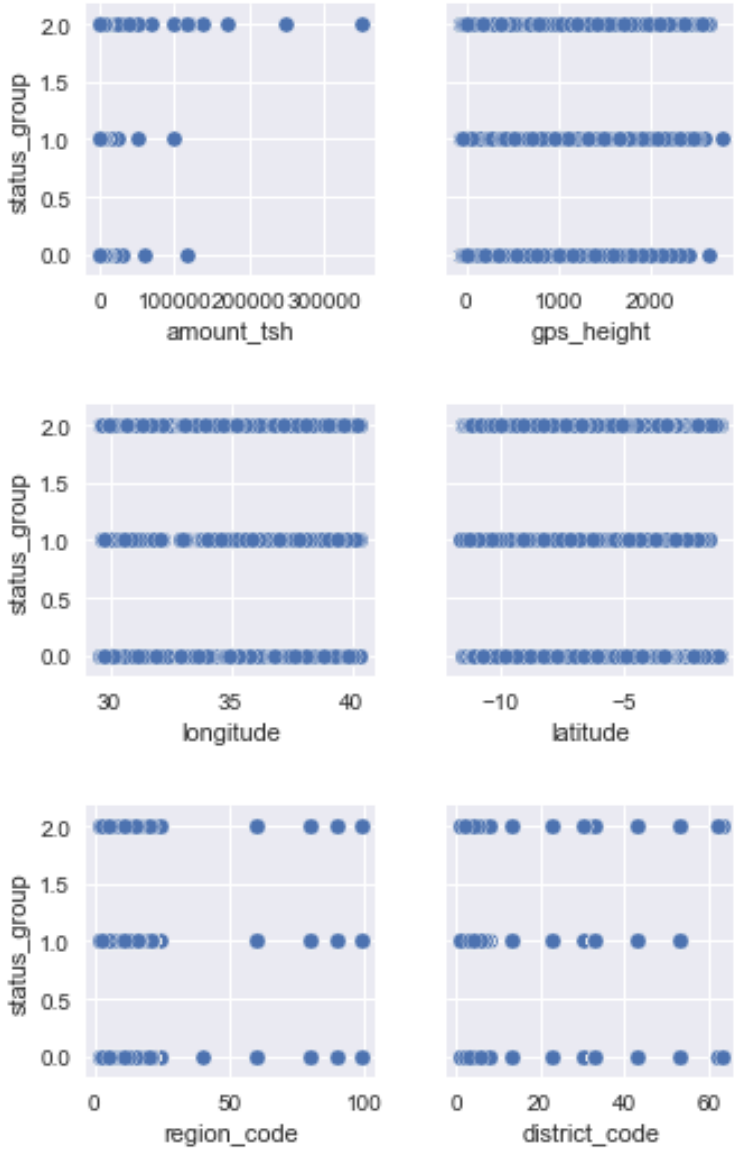
In the column for response variable "status_group," it is easy to notice that most of the mines are apparently functional in the sample. So it is necessary to plot the frequency table for it.

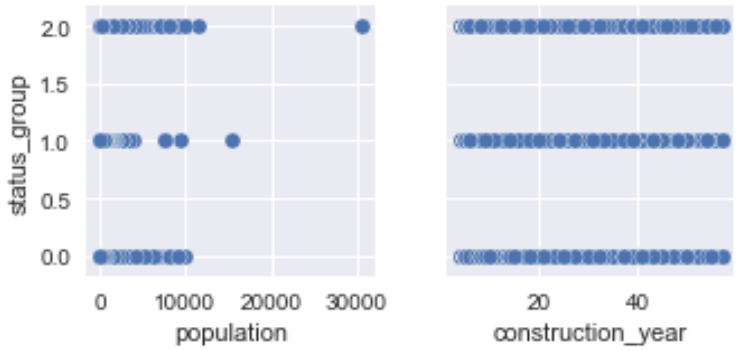


This plot is based on all 59,400 training observations.

Bivariate Analysis

Since the response variable is a categorical variable, I first plot the distribution of each quantitative variable on each level of functionality.





From the plots, it is extremely hard to see which variable is more significant in predicting the functionality of a water mine. Each variables have its values distributed pretty equivalent across all three levels of functionality.

Model Fitting Analysis

This prediction problem is to predicting a categorical variable. After the analysis for all variables, I decide to use Decision Tree, Random Forest and Boosting. All of these three models are suitable to predict categorical variable with more than two levels. The general precess is to first fit the model using training set, then predict for the data of test set. For each model, I can get a confusion matrix and a score of accuracy. In the end, comparing the results on each model, I am able to choose the best fitted model among these three.

I first fit the Decision Tree model. The resulting confusion matrix and accuracy score is:

```
confusion matrix:
[[3598  251 1401]
 [ 237  308  436]
 [1561  471 6291]]
accuracy score
0.700632128624
```

Then I fit the Random Forest model. The resulting confusion matrix and accuracy score is:

```
confusion matrix:
[[3755   91 1404]
 [ 204  281  496]
 [1131  237 6955]]
accuracy score
0.755187577298
```

At last, I fit the boosting model. The resulting confusion matrix and accuracy score is:

```
confusion matrix:
[[2981   25 2244]
 [ 301   55  625]
 [1460   33 6830]]
accuracy score
0.677889240071
```

Comparing these three results from three chosen models, we can find that the random forest model has the best prediction accuracy with score of 0.755. All of these three models are hard to interpret. It is hard to plot trees with such a large amount of data. So the accuracy is the only determinant to choose the best model. In the end, I fit the random forest model to the whole training data, which generates the following confusion matrix and accuracy score:

```
confusion matrix:
[[13283     3    53]
 [   11  2387    19]
 [   31     5 20593]]
accuracy score
0.996646969905
```

The Random Forest model on the whole training set gives an accuracy of 0.99! This is an very impressive score, although we cannot use this score to determine how this model can perform on unknown testing data. In the end, I use this Random Forest model generated by all training data to make prediction on the test data with the true labels unknown. The prediction results can be used in the competition, but I decided against it, because I

think this testing error can be pretty large to deal with and cannot compete against other model using possibly deep learning structures.

Conclusions/Limitations

In this project, I use Random Forest, Decision Tree and Boosting to predict the functionality of water mines in Tanzania. The span of the number of years constructed by now ranges from 4 years old to 57 years old, while both latitude and longitude records spread lots of area in or near Tanzania. Therefore, the sample has lots of variability in terms of the condition of water mines.

All of these three models can predict the functionality at an accuracy score of larger than 0.6. Among them, Random Forest model has the best performance. Using this model on the whole training set, the training accuracy is over 0.99, while the estimated testing accuracy is 0.75. Therefore, I can expect that the Random Forest model, using all nine predictors, is able to predict the test data at an accuracy of 0.75. The bivariate analysis above shows that it is hard to choose the best variables to predict, since the values for each predictor evenly spread across all three levels of the response variable, as shown in figures. Also, Random Forest is not a good model to interpret which variables play an important role in predicting. Therefore, minimizing the error is the whole purpose fo this project. Since the Random Forest model fits the data to a convincing level, we can be confident in using this model to make prediction on other unknown data.

The results of this project can be used to the industry, which can help in predicting the functionality of a water mine using some information that can be gained easily. The algorithm developed can also be used further in other analysis. However, since the given data only includes water mines in Tanzania, it is unknown whether the same procedure can be used on water mines on other area. Maybe using boosting model can be better when dealing with another data of water mines. Besides, I exclude some features

that are hard to process in predicting or, in my opinion, are irrelevant in this project, but it does not mean that they do not have any impact on the prediction. Using some precisely developed method or model, these predictors might be critical in helping prediction. Therefore, what these predictors mean in predicting the functionality of a water mine requires some more analysis. More work needs to be done in order to make this algorithm useful for a wider range of data sample.