Problem Description

The Taarifa platform is an open source web application that tries to bring people from around the world together to build and shape “tools for a better world.” In this particular instance, they have made a challenge to predict whether or not a given waterpoint in Tanzania needs to be repaired. This is important to the Tanzanian Ministry of Water because it is their job to ensure that citizens have access to safe drinking water in order to maintain an acceptable standard of living throughout this country. There are three “classes” of waterpoint that we are given- ‘functional’, ‘functional-needs repair;, and ‘not functional’. Both not functional and functional-needs repair indicate waterpoints that need to be repaired.

Important Data

When examining the data set, I found that there are quite a few fields that could be important in the final model that is used to predict whether or not the waterpoint needs repair. These include the following

|  |  |
| --- | --- |
| Variable | Definition |
| gps\_height | Altitude of the well |
| age | Age of the well |
| management\_group | How the well is managed |
| basin | Geographic water basin |
| permit | If the waterpoint is permitted |
| extraction\_type\_class | The kind of extraction the waterpoint uses |
| payment\_type | What the water costs |
| Waterpoint\_type\_group | The kind of waterpoint |
| source | The source of the water |
| quantity | Quantity of the water |

I determined these variables likely to be important based on the following logic. The prevailing weather conditions at a waterpoint could differ significantly based on the altitude of the well. In low areas where the weather is hotter, maybe the well is damaged more easily. Or vice versa. Why the age of the well matters is pretty self-explanatory- the older the well, the more likely it is to have worn down or been damaged. By using management\_group in our model, we will be able to determine if the way a well is being managed has any effect on its functional status. Checking the basic where the waterpoint is located can let us know if the type of water that is coming to a particular well has any effect on its status. Permit is used because there may be a difference in the quality of a well based on if it is actually allowed to be there or not. For example, wells that are permitted might be in better condition because it is easier to get regular maintenance performed on them. Non-permitted wells may not have access to these resources. Extraction\_type\_class is used to see if different methods of extracting water can be useful in predicting functional status. Payment\_type is considered, because wells with more costly water might receive better care than less costly water because these particular wells are deemed more profitable. Water\_point\_type\_group is considered to see if a particular type of water pump is more likely to need repair than another. Source of the water can let us know if specific “types” of water are more likely to damage a well than others. And finally, quantity can let us know if the amount of water that flows through the well is significant in determining.

Limitations of Data

This data does have limitations, there is a lot of other potential information that is not provided in the dataset. For example, we do not know what materials the wells are made out of, which could definitely impact how quickly they wear down. We do not know what the climate conditions surrounding a particular well are, which could again cause a well to damage more easily. These are some questions that we cannot answer with the given data.

Data Cleaning and Wrangling

Significant cleaning and wrangling of the data was necessary. To begin with, the dependent variable, “status\_group” is a factor with three levels. This variable had to be converted into a binomial where the value 0 indicated that the well was not in need of repair and 1 means that it does need repair. This was accomplished using the following code:

**#Assign 0 or 1 to Status\_Binom. 0 == does not need repair, 1 == needs repair**

**data$Status\_Binom <- with(data, ifelse(status\_group == "functional", Status\_Binom <- 0, Status\_Binom <-1))**

Once this was done, the age of the wells had to be calculated because we were not given this information directly. To do this, I subtracted the year in which the well was constructed from the current year, 2016.

**current\_year <- 2016**

**data$age <- current\_year - data$construction\_year**

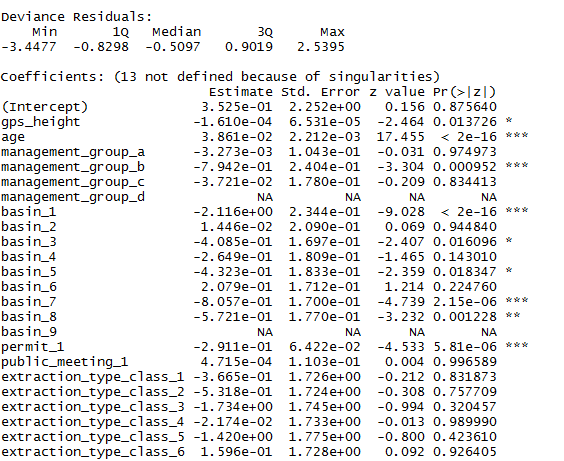
**#Remove all rows that do not have construction\_year**

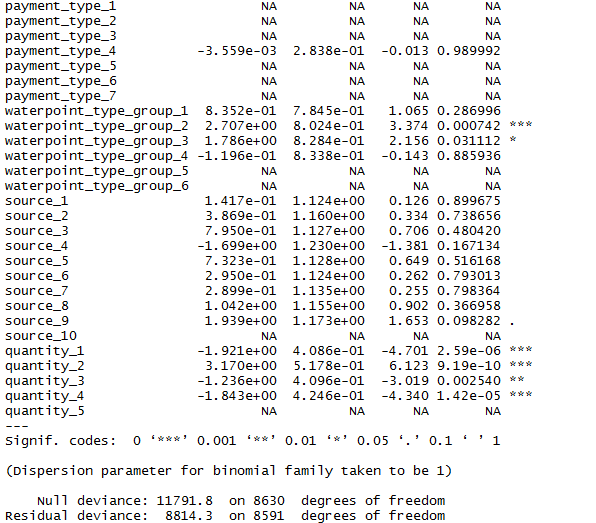
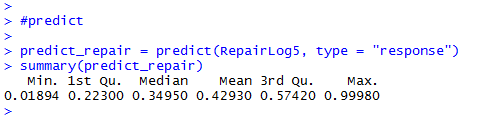
**data = data[data$age != 2016,]**

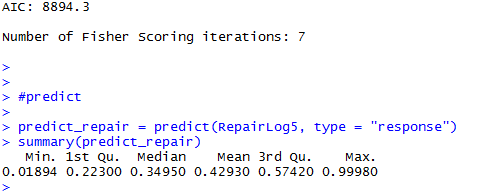
Some wells had construction year=0. These rows were removed. The variable management\_group was also categorical and thus had to be converted into dummy data. The same applied to basin, permit, public\_meeting, extraction\_type\_class, payment\_type, waterpoint\_type\_group, source, and quantity. The code for this is available on the R script. All NA values also had to be removed. The Training data also had to be split into two separate sets, one for the purposes of building the model, and another for testing the model. This was done using the function sample.split in package caTools

Since the dependent variable in this problem turns out to be binomial (needs repair, does not need repair), a logistic regression model will be used to model the data and create a prediction equation for the dataset.

Model

Once we built the model and ran it, we come up with the following summary of the model, when it has been applied to the test data. We can see that age, management\_group\_b, basin\_1 basin\_7,basin\_8, permit\_1, water\_point\_type\_2, quantity\_1, quantity\_@, quantity\_3, and quantity\_4 are all highly significant variables in this model:





I then build a function called predict\_repair after actually modelling the data to help us predict whether or not a given well needs fixing. After testing out various threshold values, it turned out that t= 0.5 had the most accurate predictions, so this is the threshold value that should be used.

Recommendations

Based on this initial analysis, the following are the recommendations that the client should pursue further:

1. Further examine the impact that age of the well has on the functional status. This was the most significant variable and learning more about the relationship between age and functional status can help us better predict whether a given well will need repairing
2. Further examine the effect that the basin which the well is drawing water from has on the functional status. Specifically, Basin 7 and Basin 8 should be investigated for common characteristics that might contribute to poor well function.
3. Finally, further examine the relationship that the quantity of water at each waterpoint has on the need for it to be repaired. These variables all turned out to be very significant.