Predictive Model for water pumps functionality in Tanzania

Thesis submitted in partial fulfillment of the
requirements for the

Post Graduate Diploma in Data Science

By

Kabilarasan

18125760035

Under the guidance of

Divya Kamath
Associate Faculty
MAHE
Bangalore

MANIPAL ACADEMY OF HIGHER EDUCATION, MANIPAL

Predictive Model for water pumps functionality in Tanzania

Thesis submitted in partial fulfillment of the
requirements for

Post Graduate Diploma in Data Science

By

(Signature)

Kabilarasan

18125760035

Under the guidance of

Divya Kamath
Associate Faculty
MAHE
Bangalore

INSPIRED BY LIFE

MANIPAL ACADEMY OF HIGHER EDUCATION, MANIPAL

Predictive Model for water pumps functionality in Tanzania

Thesis submitted in partial fulfillment of the
requirements for

Post Graduate Diploma in Data Science

By

Kabilarasan

18125760035

Examiner 1	**Examiner 2**
Signature:	Signature:
Name:	Name:

MANIPAL ACADEMY OF HIGHER EDUCATION, MANIPAL

CERTIFICATE

This is to certify that the project work titled

Predictive Model for water pumps functionality in Tanzania

is a bonafide record of the work done by

Kabilarasan
18125760035

In partial fulfillment of the requirements for the award of **Post Graduate Diploma in Data Science** under Manipal Academy of Higher Education, Manipal, Manipal and the same has not been submitted elsewhere for any kind of certification/recognition.

(Signature)
Divya Kamath
Associate Faculty
MAHE
Bangalore

Table of Contents

1.Acknowledgements

I would like to acknowledge my supervisor Prof. Divya Kamath for providing me the necessary guidance and valuable support throughout this research project. I value the assistance of Prof. Subhabaha Pal. Learning from their knowledge helped me to become passionate about my project.

2.Abstract

Water is critical to a country's development, as it is not only used in agriculture but also for industrial development. Though Tanzania has access to a lot of water, the country still faces the dilemmas of many African countries where many areas have no reliable access to water. In a household where money is scarce, families have to often spend several hours each day walking to get water from water pumps. We are looking at the dataset of water pumps in Tanzania to predict the operating condition of a water point. By finding which water pumps are functional, functional needs repairs, and non functional, the Tanzanian Ministry of Water can improve the maintenance operations of the water pumps and make sure that clean, potable water is available to communities across Tanzania. While we weren't able to identify all the pumps that need repair, our confidence in the ones we did is high and we expect this to aid the maintenance process.

3.Introduction

3.1 Motivation

Our motivation for the project is to identify the water pumps that are functional but need repair. To date, a total of $1.42 billion has been donated and spent to fix the water access crisis in Tanzania [3]. Even though there were many water pumps constructed with these donations, these water pumps are not well maintained. Many of the water pumps that were built with the donations are now in danger of failing across communities [3]. We want to help the Tanzanian Ministry of Water in identifying these water pumps that are functional but need repair so that an immediate action can be taken to keep them running in a healthy state. By fixing these water pumps early, the people of Tanzania could have improved and continuous access to running water.

3.2 Project Goal

Our project goal is to reduce the maintenance cost by half. To date, a total of $1.42 billion has been donated and spent to fix the water access crisis in Tanzania [3]. Even though there were many water pumps constructed with these donations, these water pumps are not well maintained. Many of the water pumps that were built with the donations are now in danger of failing across communities [3]. We want to help the Tanzanian Ministry of Water in identifying these water pumps that are functional but need repair so that an immediate action can be taken to keep them running in a healthy state. By fixing these water pumps early, the people of Tanzania could have improved and continuous access to running water.

4. Project Description

4.1 Data Understanding

On eyeballing the data, we identified few features that seemed discriminative based on our human intuition. According to us, amount_tsh (amount of water available to water point), gps_height, basin, installer, population, scheme_management, construction year, extraction_type, management_group, water_quality, payment type, source, and waterpoint_type seemed like they could be extremely important in identifying the pump status.

4.2 Data Limitations

1.Our dataset contains many missing values in the features associated with the water pump.

2. Our dataset has severe class imbalance, with 32,259 data points for functional water pumps, 4,317 data points for functional water pumps but needs repair, and 22,824 data points for non functional water pumps.

3. Our dataset contains many categorical features. There are 31 categorical features in our data and many of them have high arities.

4. Our dataset contains many features that contain similar representation of data presented in different grains.

4.3 Benefits of project

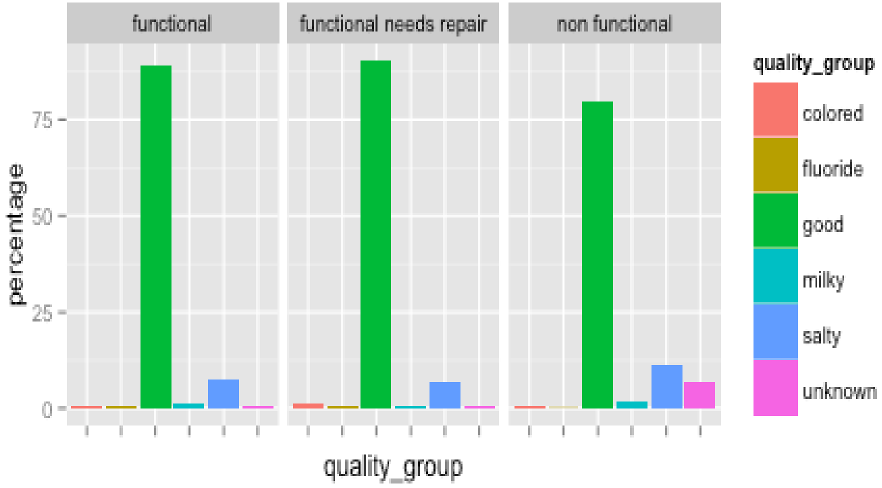
We Can reduce the maintenance cost by half and By fixing these water pumps early, the people of Tanzania could have improved and continuous access to running water.

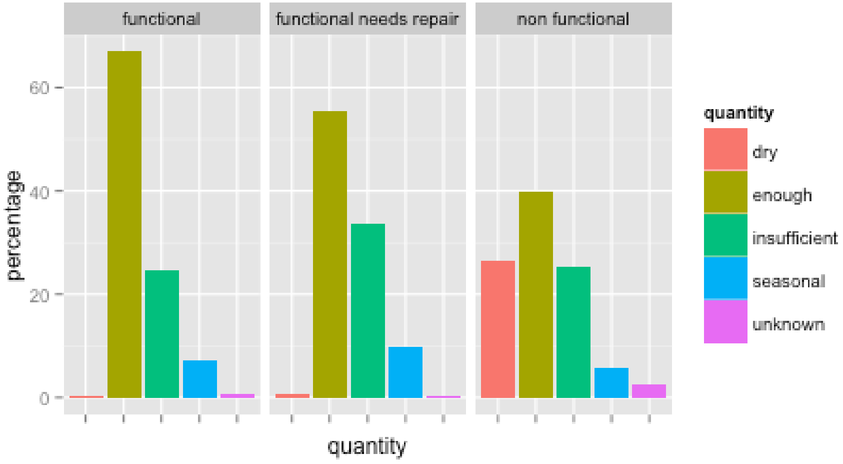
5. Exploratory Data Analysis

5.1 Data collection

The data of the water pump was collected using handheld sensor, paper reports, and user feedback via cellular phone.

5.2 Data Analysis





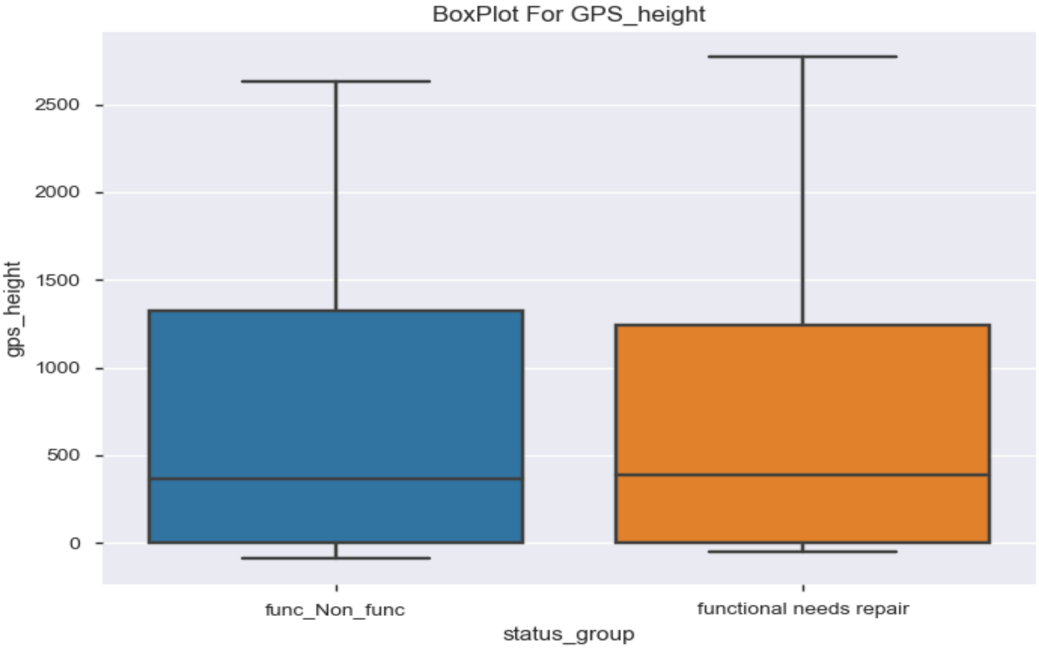
6. DATA PREPARATION FOR MODELLING

6.1 Feature Description

1. amount_tsh - Total static head (amount water available to waterpoint)
2. date_recorded - The date the row was entered
3. funder - Who funded the well
4. gps_height - Altitude of the well
5. installer - Organization that installed the well
6. longitude - GPS coordinate
7. latitude - GPS coordinate
8. wpt_name - Name of the waterpoint if there is one
9. num_private -
10. basin - Geographic water basin
11. subvillage - Geographic location
12. region - Geographic location
13. region_code - Geographic location (coded)
14. district_code - Geographic location (coded)
15. lga - Geographic location
16. ward - Geographic location
17. population - Population around the well
18. public_meeting - True/False
19. recorded_by - Group entering this row of data
20. scheme_management - Who operates the waterpoint
21. scheme_name - Who operates the waterpoint
22. permit - If the waterpoint is permitted
23. construction_year - Year the waterpoint was constructed
24. extraction_type - The kind of extraction the waterpoint uses
25. extraction_type_group - The kind of extraction the waterpoint uses
26. extraction_type_class - The kind of extraction the waterpoint uses
27. management - How the waterpoint is managed
28. management_group - How the waterpoint is managed
29. payment - What the water costs
30. payment_type - What the water costs
31. water_quality - The quality of the water
32. quality_group - The quality of the water
33. quantity - The quantity of water
34. quantity_group - The quantity of water
35. source - The source of the water
36. source_type - The source of the water
37. source_class - The source of the water
38. waterpoint_type - The kind of waterpoint
39. waterpoint_type_group - The kind of waterpoint
40. status_group - functional status of the pump

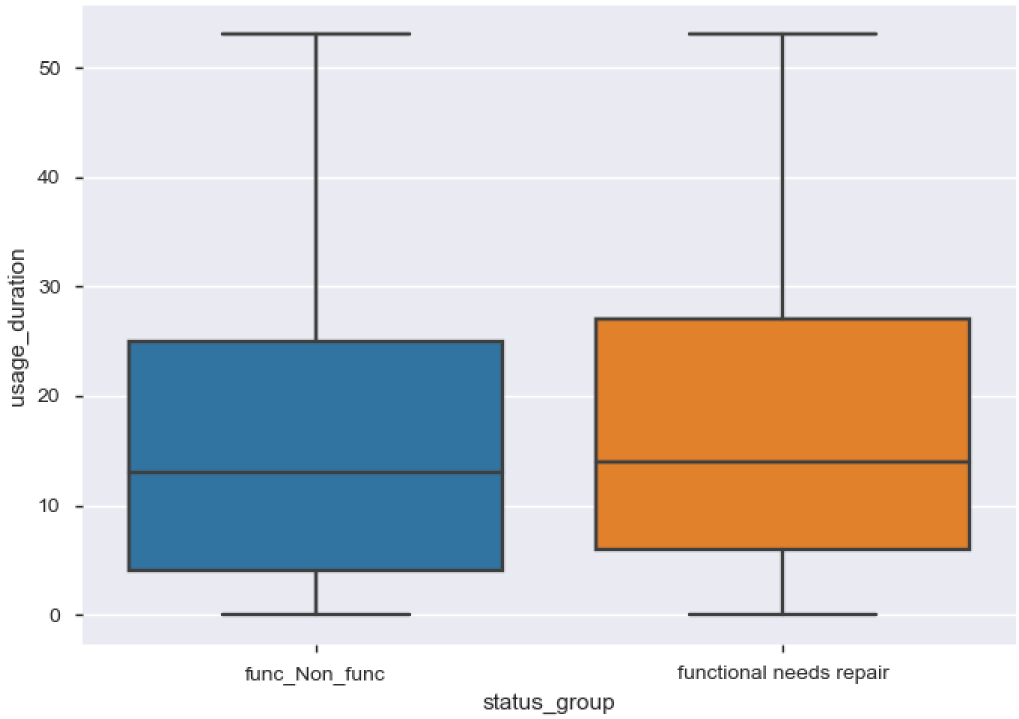
6.2 Feature Engineering

Our dataset contains many features that contain similar representation of data presented in different grains. The group of features of (extraction_type, extraction_type_group, extraction_type_class), (payment, payment_type), (water_quality, quality_group), (source, source_class), (subvillage, region, region_code, district_code, lga, ward), and (waterpoint_type, waterpoint_type_group) all contain similar representation of data in different grains. Hence, we risk overfitting our data during training by including all the features in our analysis. We tried to avoid this risk by identifying features in each group that contained the finer grain which held more information in the analysis or looked at the correlation analysis across the features to see which one is a better fit.



since the gps_height is having almost simliar distribution for both class in status group, this variable will not be helpful for us to classify.so, i dropped this variable.

```
data['usage_duration'] = pd.to_datetime(data['date_recorded']).dt.year - data['construction_year']
```



since the usage duration is also having the same disribution for both classes, I dropped the usage duration varaible,construction_year,recoreded_date

Since the wpt_name , installer has more than 1000 unique values. I catagorised it into three different groups.

```
dic = defaultdict(list)

temp =
pd.DataFrame(data.groupby("wpt_name")['wpt_name'].count().sort_values(ascending  =
False),index= None,columns = ['wpt_name'])

temp['wpt_name'].items()

for i,j in temp['wpt_name'].items():

    if i== "none":
        dic['wpt_name'].append(i)
        dic['wpt_label'].append("wpt_none")

    elif j>=5 and i!= "none":
        dic['wpt_name'].append(i)
        dic['wpt_label'].append("wpt_high")

    elif j<5 and i!= "none" :
        dic['wpt_name'].append(i)
        dic['wpt_label'].append("wpt_less")
```

7. Modelling

7.1 Selection of model/technique

Since our goal is to predict the pump which needs repair,To make the classification simple I converted the three label problems to binary classification problem.
I tried to build the model using Logistic Regression, Decision Tree, Random Forest, ExtraTreeClassifier, Naïve bayes classification model. Because of imbalanced Data f_1 score , recall, precision would be the appropriate performance metrics.
Naïve bayes performed well for this data than other classification model.

7.2 Challenges faced

There were quite a lot of challenges faced while working on this project, some of them are mentioned below:

a.) Even though the accuracy of the model is very high, the data is highly imbalanced so, we can't relay on accuracy of the model.

b.) We used Random Forest on our raw data, but faced issues due to low memory. While the heap size was extended to 4GB, because few features had high arity e.g. for funders the arity was 1897, we could not build a Random Forest model with 50 trees and maximum depth of 10.

7.3 MODEL

Random Forest Model :

```
 X = model_data.drop(['status_group'],axis =1)
 Y = model_data['status_group']
 X = pd.get_dummies(X)
new_model_1 = RandomForestClassifier()
train_x,test_x,train_y,test_y=train_test_split(X,Y,test_size=0.3,random
_state = 100)
new_model_1.fit(train_x,train_y)
```

Classification_Report:

```
                          precision  recall  f1-score  support
func_Non_func                  0.94      0.97    0.95     8143
functionalneeds repair    0.33      0.20    0.25     671
avg / total    0.89 0.91
```

Extra Tree Classifier:

```
model_1 = ExtraTreesClassifier(class_weight= {"functional needs repair"
: 0.92,"func_Non_func" : 0.08})

model_1.fit(train_x,train_y)

accuracy_score(test_y,model_1.predict(test_x))
```

Classification Report:

```
                         precision    recall  f1-score   support

           func_Non_func       0.95      0.94      0.94      8243

functional needs repair       0.23      0.26      0.25       571

             avg / total       0.90      0.90      0.90      8814
```

Naïve bayes :

```
gaussian_model = GaussianNB()

gaussian_model.fit(train_x,train_y)

print(classification_report(test_y,gaussian_model.predict(test_x)))
```

Classification Report:

```
                         precision    recall  f1-score   support

           func_Non_func       0.97      0.57      0.72      8243

functional needs repair       0.11      0.74      0.18       571

             avg / total       0.91      0.58      0.68      8814
```

7.4 Model Selection:

Out of three models , Naïve bayes has good recall value. Using Naïve bayes model 60 percentage of cost can be reduced. Even though precision is very low in this model , with good recall value we can reduce the cost of maintenance .

Precision = True Positive /(True Positive + False positive)
Recall = True Positive / (True positive + False negative)

Since we can find the 74 percentage of positive classes using this naïve bayes and while comparing other classification models this model performed good and able to give some useful outputs. So I conclude this naïve bayes model as a appropriate model for this data to classify.

8. Key Results

8.1 Final Results

Our analysis leads us to believe that the dataset was difficult to classify. We focused our results based on the costs associated with fixing a pump and effort wasted in false identification. We prioritized the reduction of false positives, since these were either functional pumps that needed no repair or non---functional where maintenance cost was significantly higher being misclassified into functional needs repair. The cost of a standard pump ranges from $100 ---$2000 . Installing this pump requires drilling which can be anything between $1000--- $3000. On the other hand maintaining the pump would only cost tens of dollars .Here we have not considered the transportation costs. While we want to improve our true positive rates which would eventually be beneficial to the society, our major concern was to reduce effort in travelling to a non functional water pump or even worse to functional water pump. That is time which could have be effectively used elsewhere.

Though we could not identify all the pumps that needed repair, we have high confidence in the ones which we did recognize. By figuring out the 74% of water pumps that are functional which needs repair, the Tanzanian Ministry of Water can set appropriate priority for their maintenance operations and help relieve water crisis for the communities in Tanzania.

9. Future Work

9.1 Future work

1. Identify the lifetime of a pump
   We would like to identify the decay rate of a pump using the construction year and see how it correlates across different population and number of pumps across the area. Our hypothesis is that the pump will be more prone to being non-functional and functional needs repairs in areas where there is dense population and not enough coverage by the functional pumps. We would further work with Tanzanian Ministry of Water and Taarifa to provide us the appropriate construction year for the missing values in the 20,709 water pump data points.

2. Sparse Classification Methods
   We would like to fit a sparse classification model such as logistic regression with L1 penalty on our dataset to create a list of features that are important in determining the status of the water pumps as being functional, functional needs repair, and non functional. Since our dataset contains 40 features with 31 categorical features, we believe that using sparse classification models will help determining important features that will significantly speed up our training time for models such as Random Forest.Additionally, we would like to try out few other methods to reduce the arity of the features such as clustering.

10. References

1. http://www.humanosphere.org/world-politics/2014/12/tanzania-failed-fix-water-access-problem/
2. https://www.kfw-entwicklungsbank.de/migration/Entwicklungsbank-Startseite/Development-Finance/About-Us/Local-Offices/Sub-Saharan-Africa/Office-Tanzania/Activities-in-Tanzania/Water-Sector-Status-Report-2009.pdf
3. http://www.simplepump.com/APPLICATIONS/Developing-Nations/True-Dollar-Costs.html