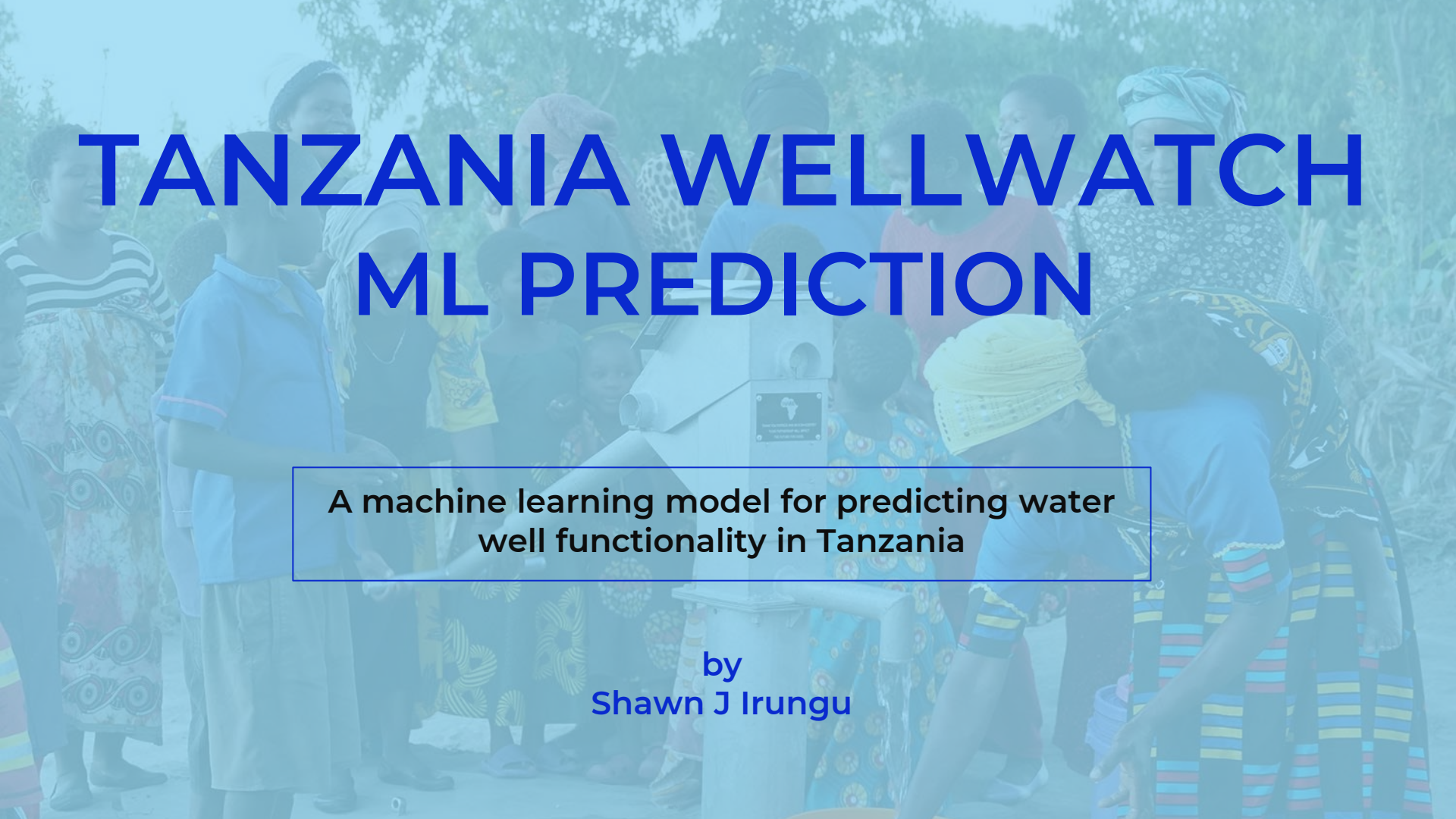


TANZANIA WELLWATCH ML PREDICTION

A machine learning model for predicting water well functionality in Tanzania

by
Shawn J Irungu

Introduction

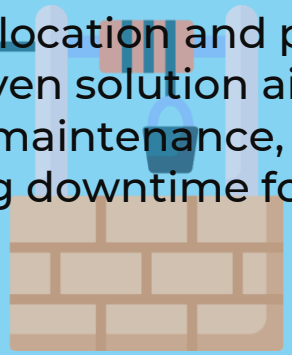
Like many other sub-Saharan African countries, Tanzania is a developing country struggling to provide adequate clean water for its bulging population that is growing at 3% per annum.

This project seeks to build a Machine Learning classifier algorithm that can predict the condition of a water well (functional, functional-but-needs-repair, and non-functional), using data such as the kind of pump, when it was installed, the installer, the region, and so on.

I am committed to partnering with the Tanzanian government to tackle this challenge - ensuring clean, safe, and reliable water for millions.

01 Business Understanding

The WellWatch ML project addresses Tanzania's critical challenge of maintaining reliable water wells by leveraging machine learning to predict failures before they occur. With many wells becoming non-functional due to mechanical issues, low water tables, or poor maintenance, the Tanzanian government and water authorities face inefficient resource allocation and prolonged water shortages. This AI-driven solution aims to shift from reactive to proactive maintenance, optimizing repair budgets and reducing downtime for communities.



Main Objective

To build a Machine Learning classifier that will predict the condition of a water well (functional, functional-but-needs-repair, and non-functional).

Specific Objective

1..... To help the Government of Tanzania find patterns in functional and non-functional wells, to help influence how new wells are built.

2..... To find the most important factors that influence whether a pump is functional, functional-but-needs-repair, or non-functional. This can guide the management of new and existing water wells.

Data Understanding

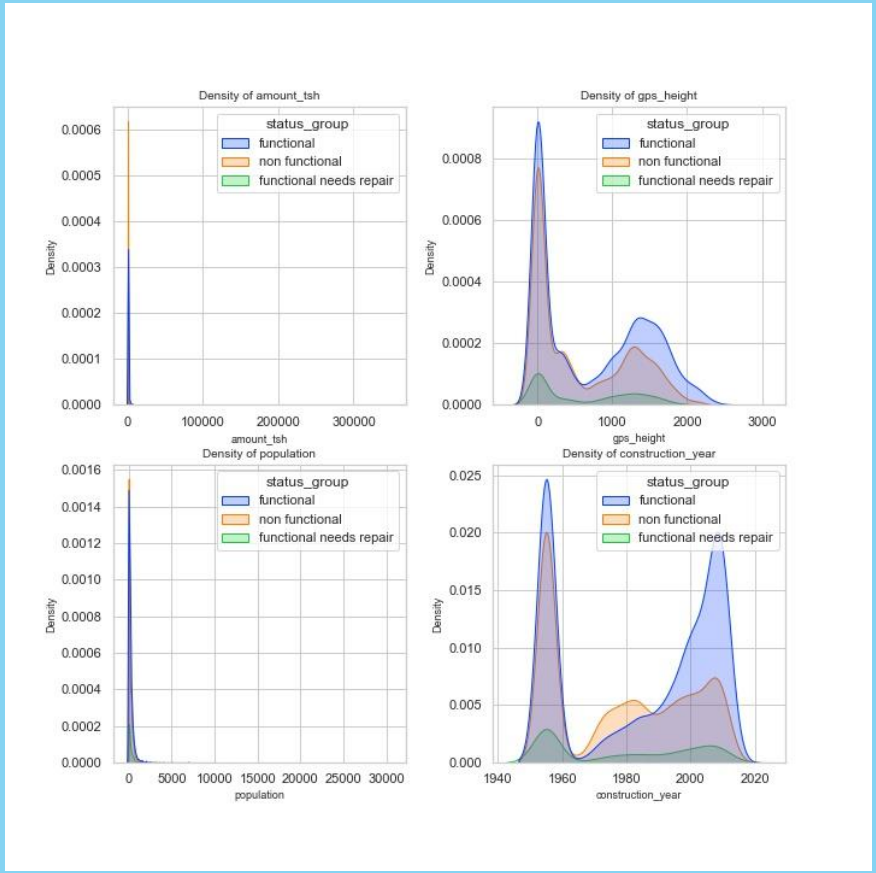
02

Datasets for this project have been provided on the Driven Data website which is hosting a competition for this project. Data for the dependent variable in the test dataset has not been provided, therefore, I will make use of Training set values and Training set labels datasets. The training dataset contains 59,400 waterpoints in Tanzania and 39 features.

By cleaning, analyzing, and visualizing this data, we extract insights that help Tanzania Government make informed decisions when constructing water wells.

1. Relationship Between Pump Functionality and Continuous Variables

Data Analysis 03



1. Relationship Between Pump Functionality and Continuous Variables

Data Analysis 03

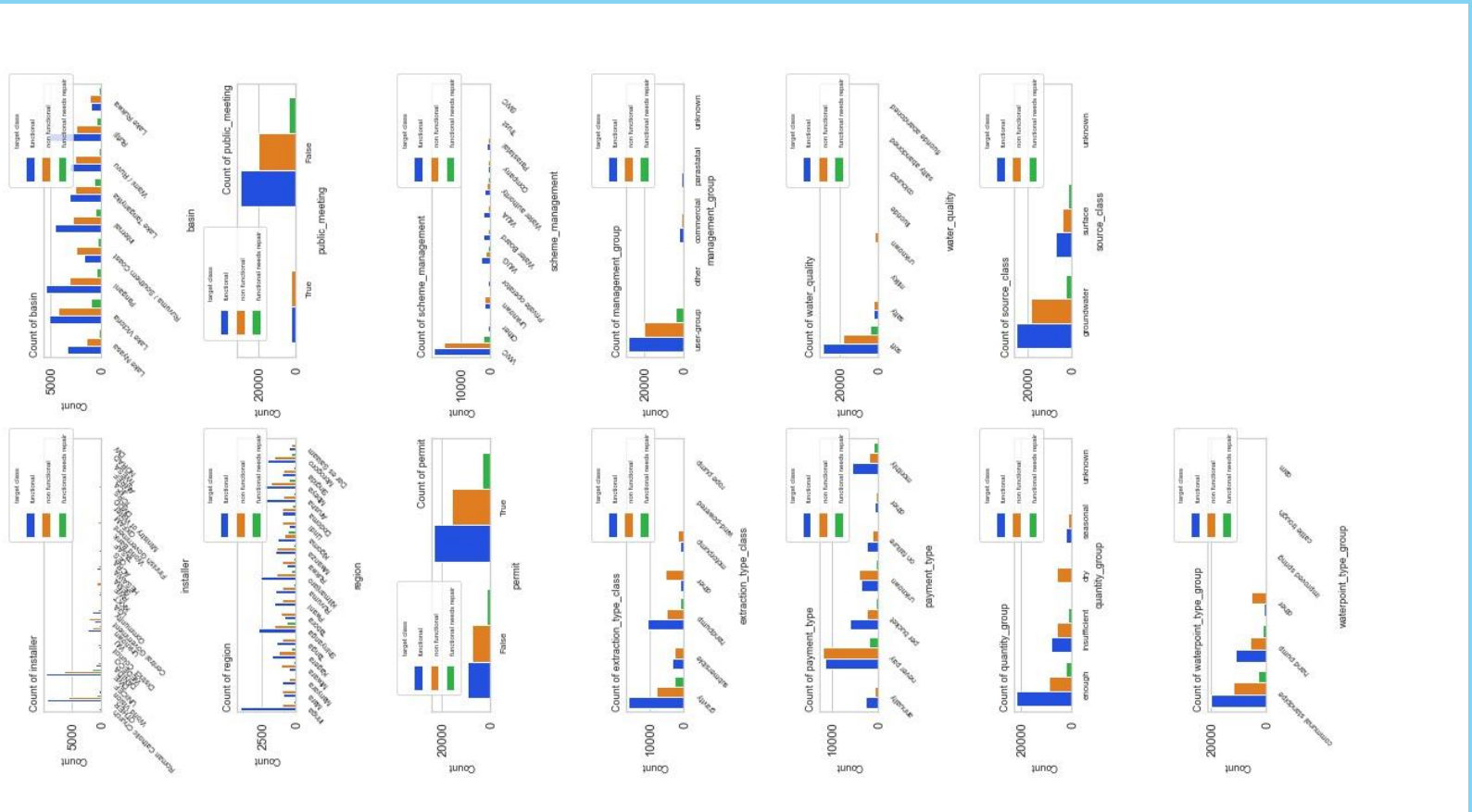
- For the total static head feature (amount_tsh), waterpoints with zero static head have the highest density of pumps overall. Also, among the three pump classes at this point, non-functional pumps have the highest density followed by functional pumps. Functional-needs-repair pumps are the least.

- For the population feature, waterpoints located in areas with zero population have the highest density of pumps overall. Also, among the three pump classes at this point, non-functional pumps have the highest density followed by functional pumps. Functional-needs-repair pumps are the least.

- For the construction year feature, the year 1955 has a high density of pumps, but these are the year 0 rows which I imputed with 1955. The KDE plot also shows that the density of functional pumps is higher among the newest pumps while non-functional pumps are higher among the older pumps, from around 1965 to 1990.

2. Relationship Between Pump Functionality and Categorical Variables

Data Analysis



2. Relationship Between Pump Functionality and Categorical Variables

Data Analysis

From the different visualizations of categorical variables, we notice that some classes of categories are more popular than others. For example, Iringa and Kilimanjaro regions have the highest number of pumps. The never-pay payment scheme is most popular and over 40,000 out of 59,400 wells have soft water quality.

From the distribution of pump functionality class for each class of a categorical variable, we notice that the functional pumps are more frequent than functional-needs-repair and non-functional pumps.

A notable deviation from this trend is the never-pay class of the payment-type category, where non-functional pumps are more than the other classes of pumps.

3. To build a Machine Learning classifier that will predict the condition of a water well.

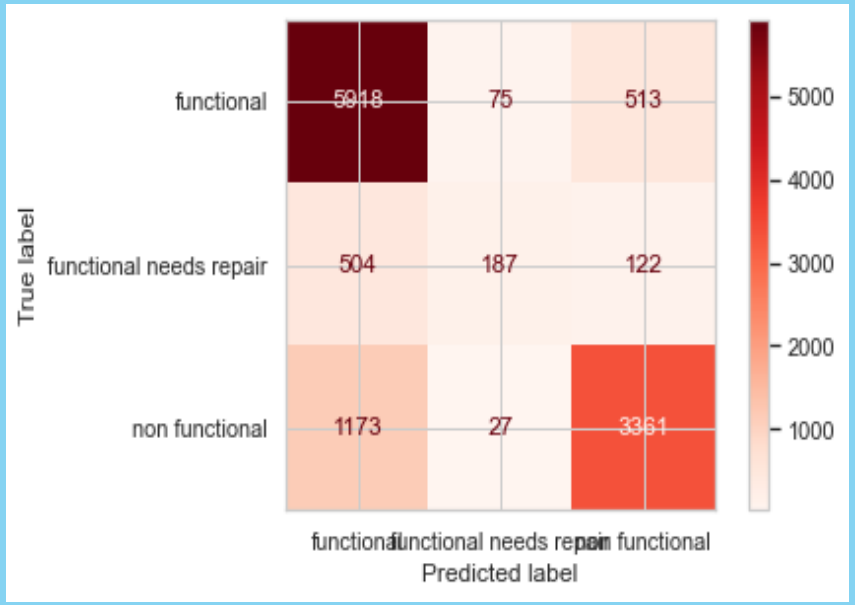
Data Analysis

```
======================================================================
MODEL REPORT
======================================================================
                         precision    recall  f1-score   support

             functional       0.78      0.91      0.84      6506
functional needs repair       0.65      0.23      0.34       813
         non functional       0.84      0.74      0.79      4561

               accuracy                           0.80     11880
              macro avg       0.76      0.63      0.65     11880
           weighted avg       0.79      0.80      0.78     11880


======================================================================
MODEL METRICS - TRAIN SET
======================================================================
Overall accuracy score 0.8293350168350169
Overall precision score 0.8319071109633489
Overall recall score 0.8293350168350169
Overall F1-score 0.8186159596319987
MODEL METRICS - TEST SET
======================================================================
Overall accuracy score 0.7968013468013468
Overall precision score 0.793916660876027
Overall recall score 0.7968013468013468
Overall F1-score 0.7844948789304663
```

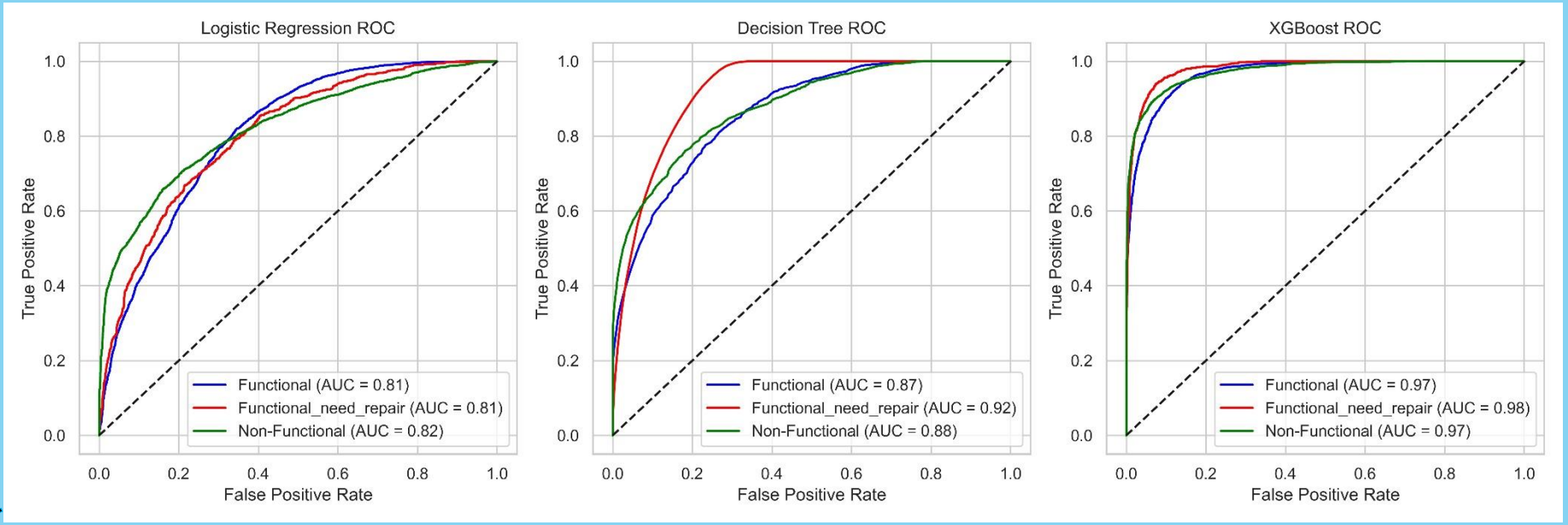


3. To build a Machine Learning classifier that will predict the condition of a water well.

Data Analysis

We conclude that XGBoost will be final model since the difference in metrics between train set and test set is 0.04 which is a tiny effect to say that our model is overfitting. It performed well with high metrics than other models, has an accuracy score of 0.7968 and a f1 score of 0.7845 on test set, an accuracy score of 0.829 and a f1 score of 0.819 on train set, we are going to use F1 score as our metric since our target classes were imbalanced and thus 0.7845 is a high scoring from all models we created

4. AUC – ROC Analysis



Above is the AUC ROC curve comparison between the 3 models.
XGBoost Performs Best in the 3 classes

5. AUC – ROC Analysis

XGBoost shows to be the best model since the 3 class shows a fpr-tpr tradeoff close to 1.0 at TPR(true positive rate) compared to other models.

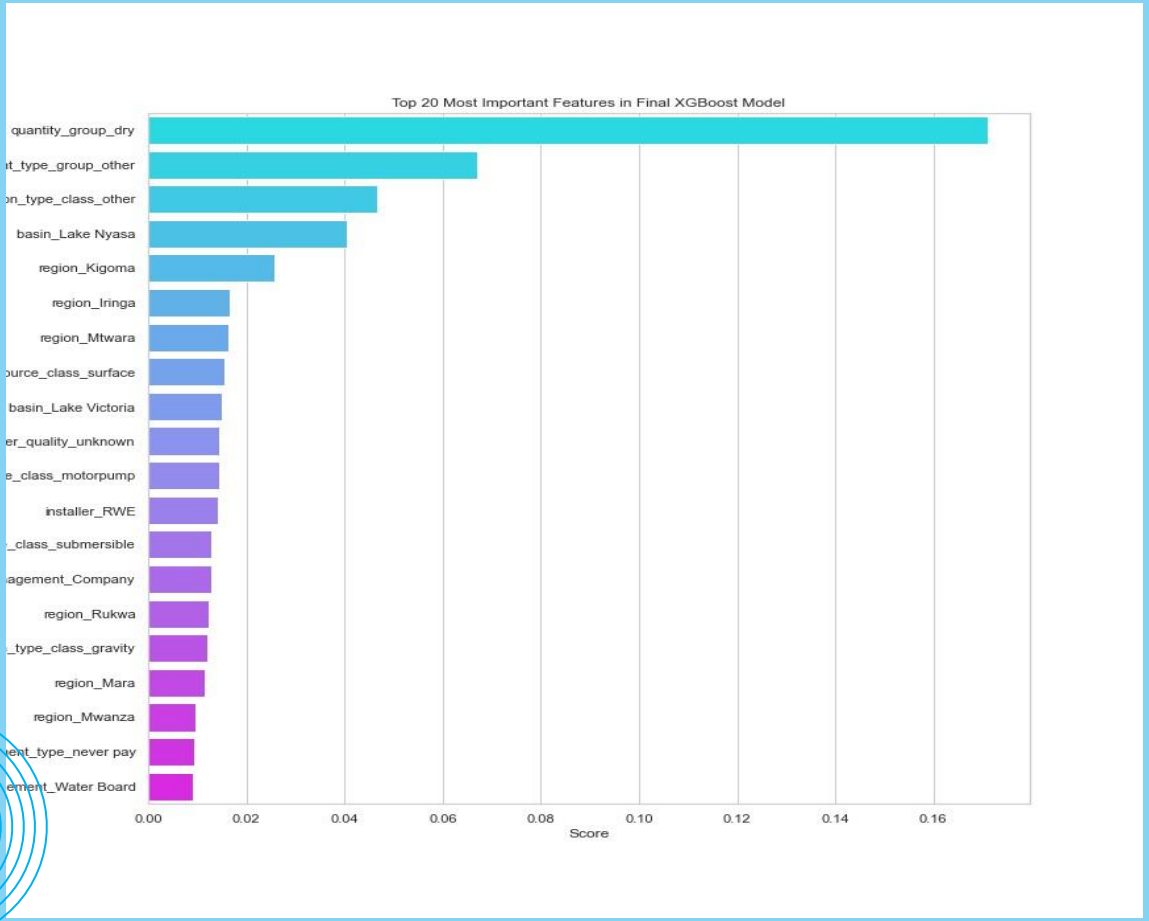
XGBoost will be final model since the difference in metrics between train set and test set is 0.04 which is a tiny effect to say that our model is overfitting. It performed well with high metrics than other models, has an accuracy score of 0.7968 and a f1 score of 0.7845 on test set, an accuracy score of 0.829 and a f1 score of 0.819 on train set, we are going to use F1 score as our metric since our target classes were imbalanced and thus 0.7845 is a high scoring from all models we created.

Below are XGBoost metrics comparison

MODEL METRICS	TRAIN SET
Accuracy	0.8293350168350169
Precision	0.8319071109633489
Recall	0.8293350168350169
F1 Score	0.8186159596319987

MODEL METRICS	TEST SET
Accuracy	0.7968013468013468
Precision	0.793916660876027
Recall	0.7968013468013468
F1 Score	0.7844948789304663

6. 20 Important Features



Some of the top features influencing a prediction include:

i.) quantity-group (the quantity of water)

ii.) The water point type

iii.) The extraction type class

iv.) The basin

v.) scheme management

vi.) The installer

vii.) payment type

04 Conclusion

For the total static head feature (amount_tsh), waterpoints with zero static head have the highest density of pumps overall. Also, among the three pump classes at this point, non-functional pumps have the highest density followed by functional pumps. Functional-needs-repair pumps are the least.

From the box plot of total_static_head vs. pump condition, we can see that the pumps having tsh above approx.125,000 are all functional. High static head may be an important feature because the higher the tsh the higher the probabilty of a pump being functional.

For the population feature, waterpoints located in areas with zero population have the highest density of pumps overall. Also, among the three pump classes at this point, non-functional pumps have the highest density followed by functional pumps. Functional-needs-repair pumps are the least.

For the construction year feature, the year 1955 has a high density of pumps, but these are the year 0 rows which I imputed with 1955.

A KDE (kernel density estimation) plot shows that the density of functional pumps is higher among the newest pumps while non-functional pumps are higher among the older pumps, at 1955, 1980 and around 2010.

RECOMMENDATIONS 05

01

I advise the Government of Tanzania to apply my final model in predicting the condition of well pumps across Tanzania. It will help them to correctly predict the actual condition of each pump at at least 78% success rate.

02

The government will need to implement and operationalize a payment scheme for the water points, having observed that the sites where people never pay for water had the highest frequency of non-functional pumps.

03

Finding out if there is more data that can balance the target classes. The current classes are imbalanced with the most frequent class comprising 37.2% of the data while the least class comprises only 4.42%. This affected the prediction score of the least class compared to the other classes. Availability of more data that can balance the classes would realize much better prediction scores.

Thank you!

Do you have any questions?

For collaboration click on the GitHub Icon

Irungu Shawn