WATER IS FOR EVERYBODY

Bringing clean water to Tanzania.
Presented by Shefat Moral, Gabriel Santorelli, David Jimenez

We believe in delivering clean water to everyone.

Tanzanian Waterwell Project

Providing functional waterwells��

Our goal for this project was to build an effective ML model that can predict water pump functionality for the purpose of bringing clean water to the people of Tanzania .

EDA Process:

- Analyze the data.
- Select determining factors for our model.
- Cleaning the data for our model.

Model Selection:

- Preprocessing for our baseline model.
- Decision Tree - baseline model.

A Model the People Can Trust:

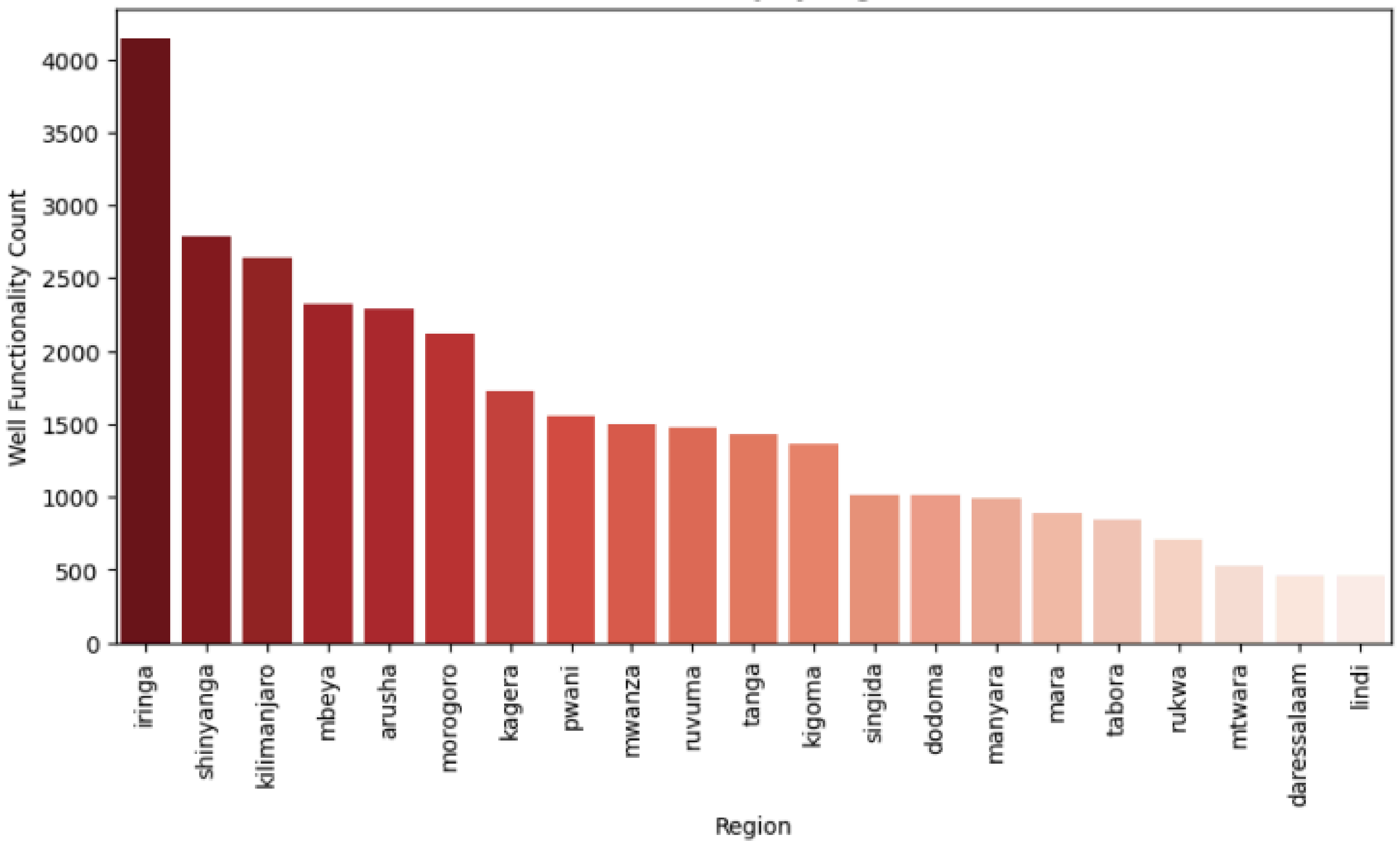
- Feature Engineered
- Hyperparameter Tuning
- Class Balancing

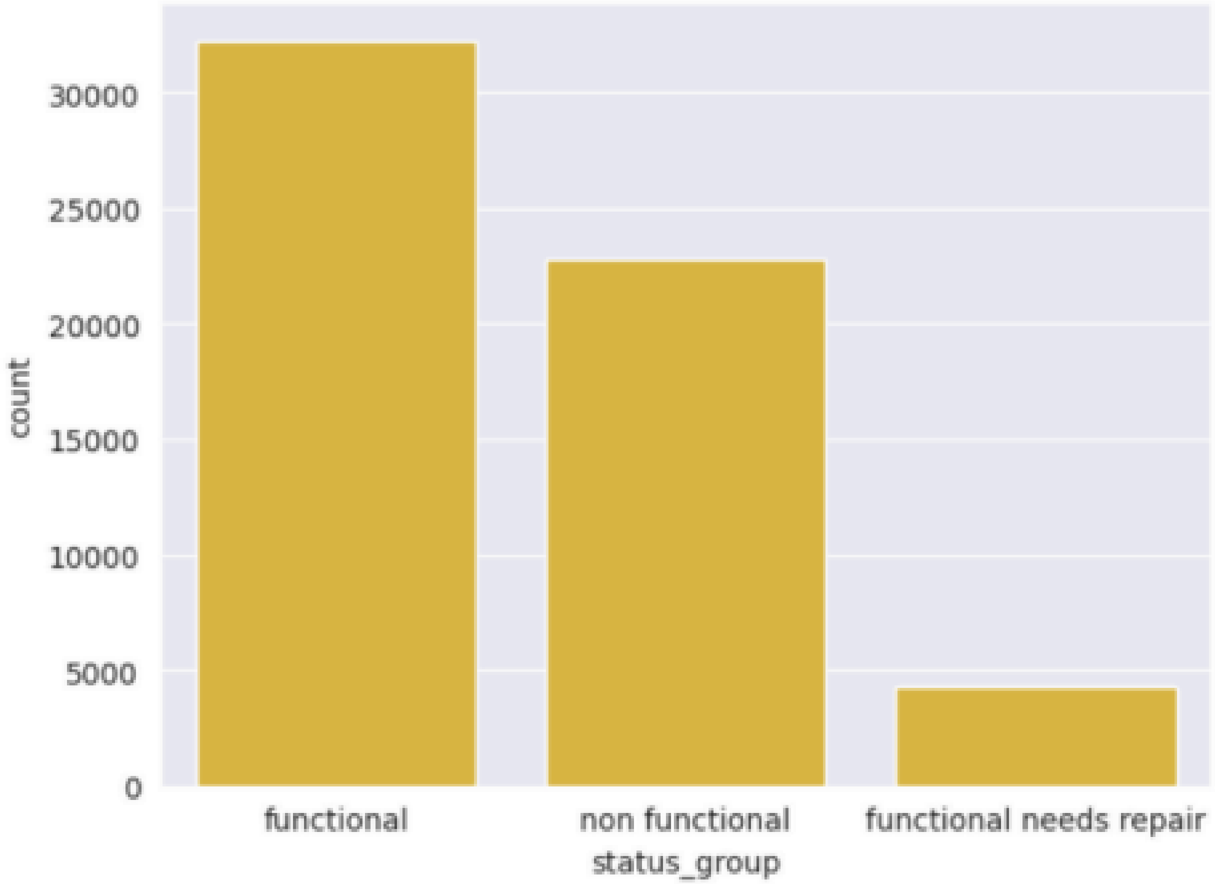
Optimal Performance

Tanzanian Waterwell Project

EDA Process

Data Exploration


Well Functionality Count
4000
3500
3000
2500
2000
1500
1000
500
0
iringa
shinyanga
kilimanjaro
mbeya
arusha
morogoro
kagera
pwani
mwanza
ruvuma
tanga
kigoma
singida
dodoma
manyara
mara
tabora
rukwa
mtwara
daressalaam
lindi
Region


count
30000
25000
20000
15000
10000
5000
0
functional
non functional
functional needs repair
status_group

Multiclass data:

Functional

Not Functional

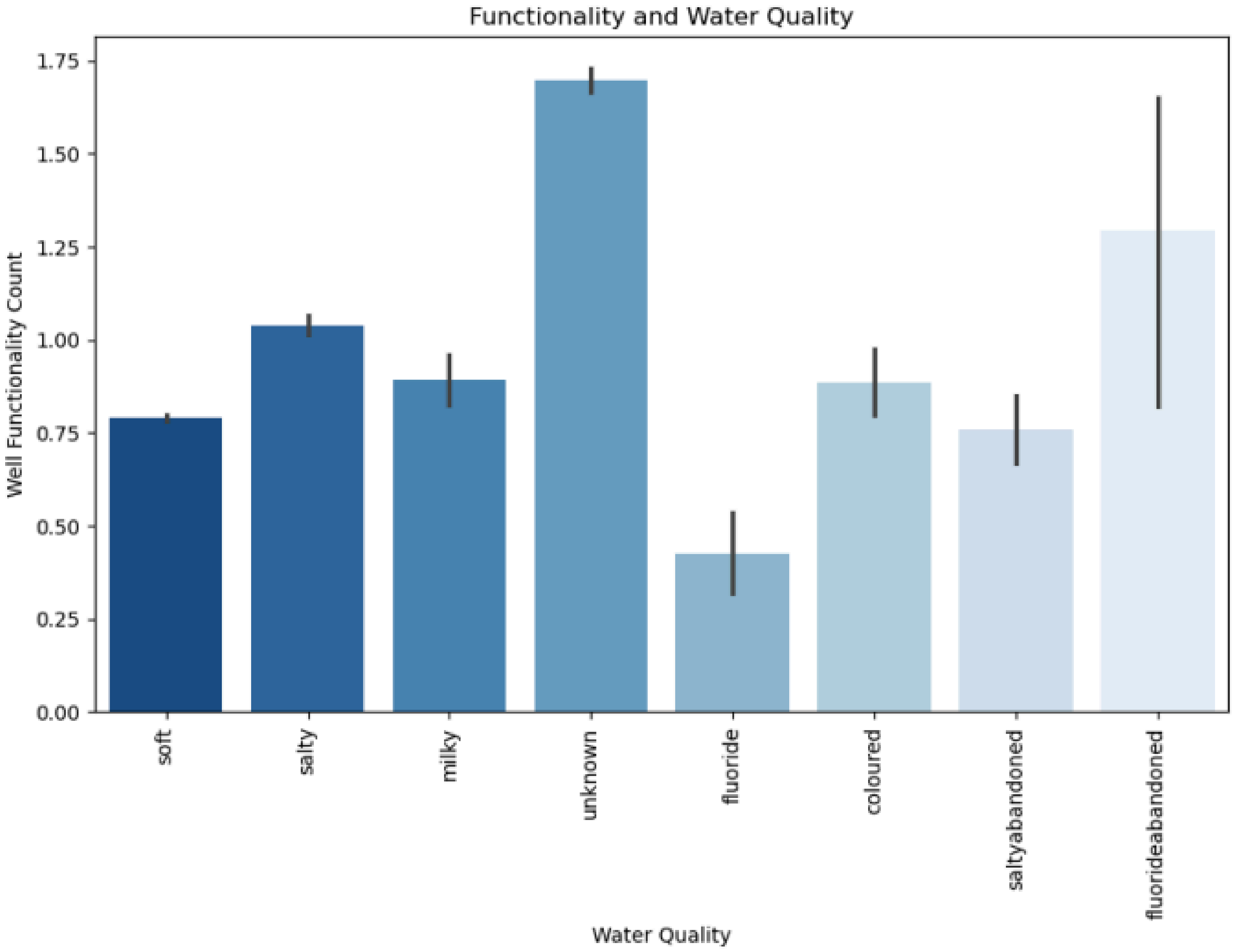
Functional Needs Repair

Looked at features like Functionality by Region

Tanzania Waterwells Project

Water Qualtiy Analysis

Functioning waterwells + Water Quality



Functioning did not always mean clean water. Unknown, Flourdide Abandoned, and Salty top the Water Quality list.

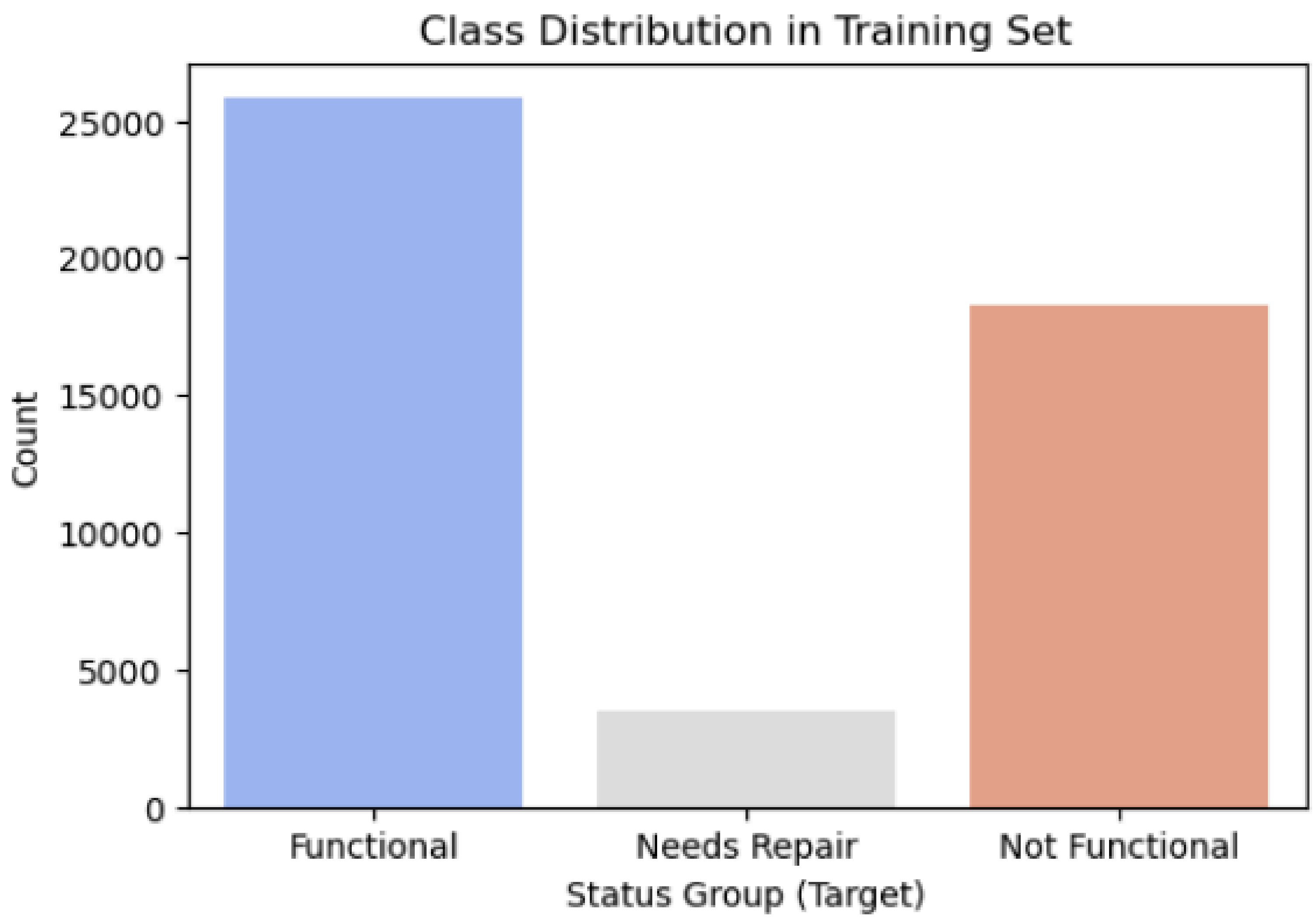
Soft(good) was near the bottom of the quality list.

Tanzanian Waterwells Project

Model Selection

Overweighted class distribution
Reduced cardinality
Handled missing values

Decision Tree was baseline
model Initial Accuracy 75%
Identified Class 0 and 2 well

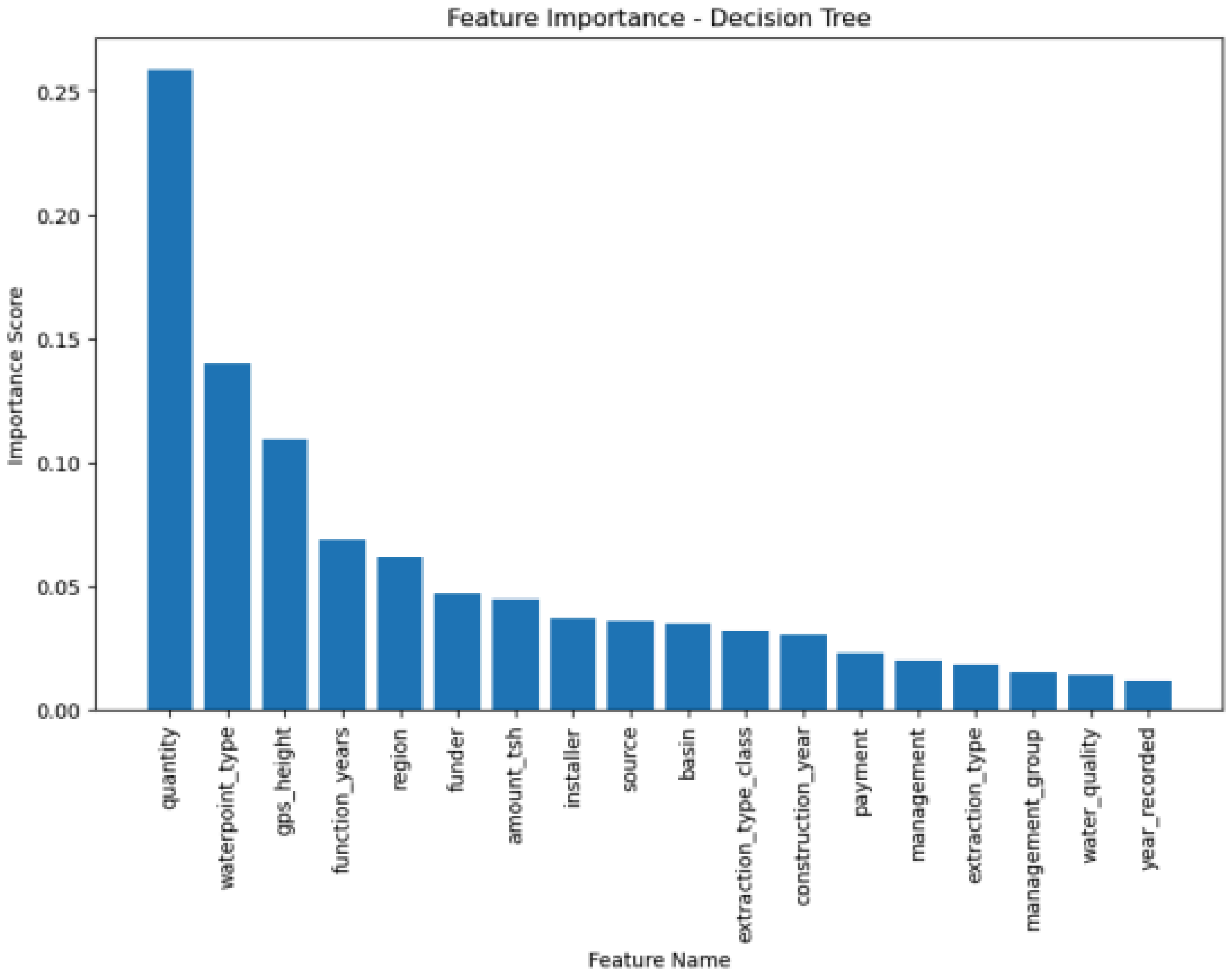


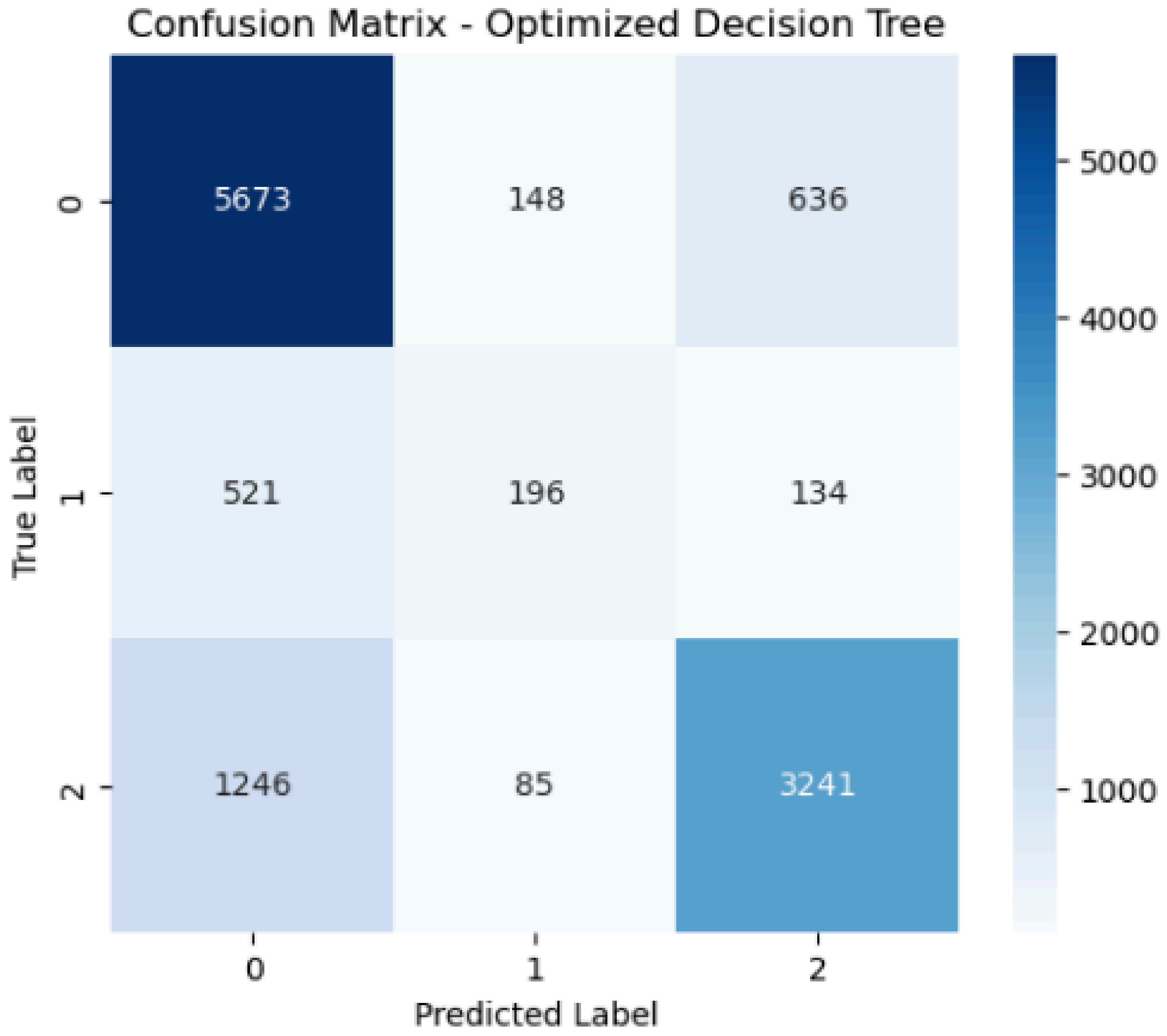
```
Decision Tree Accuracy: 0.7505
              precision    recall  f1-score   support

           0       0.78      0.82      0.80      6457
```

Struggled to identify Class 1

Decision Tree Model



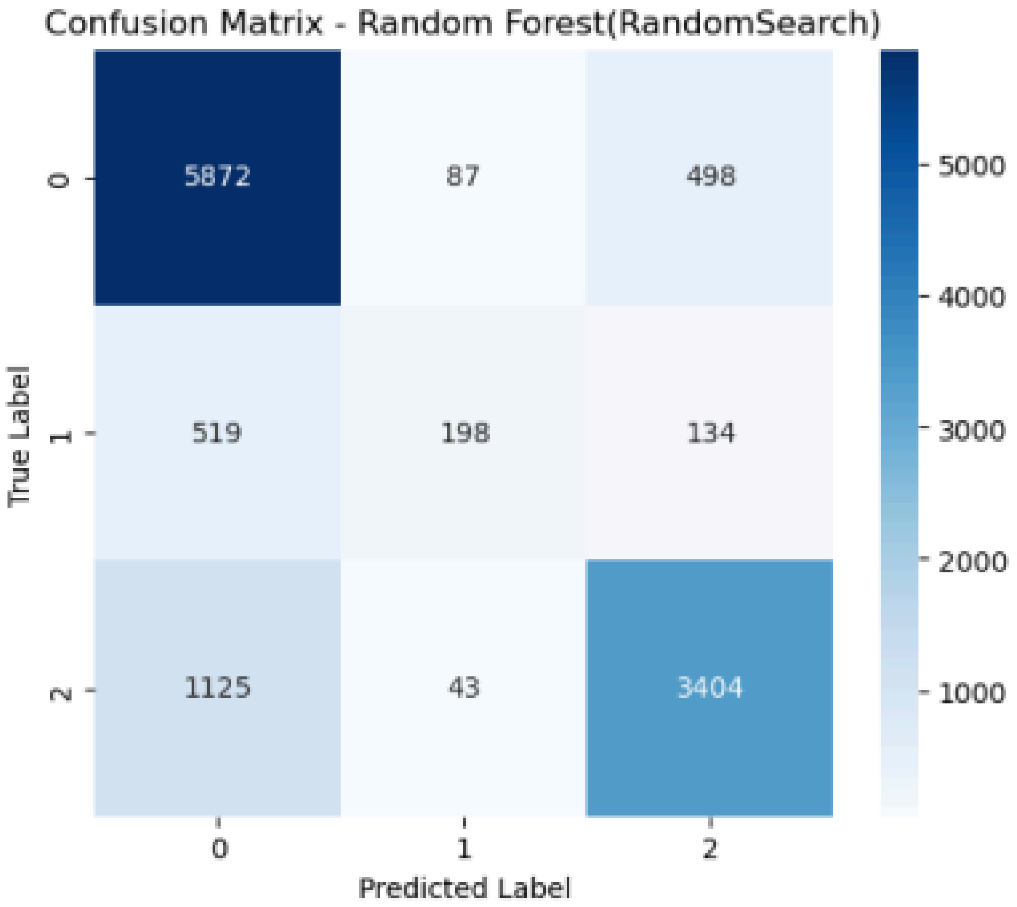


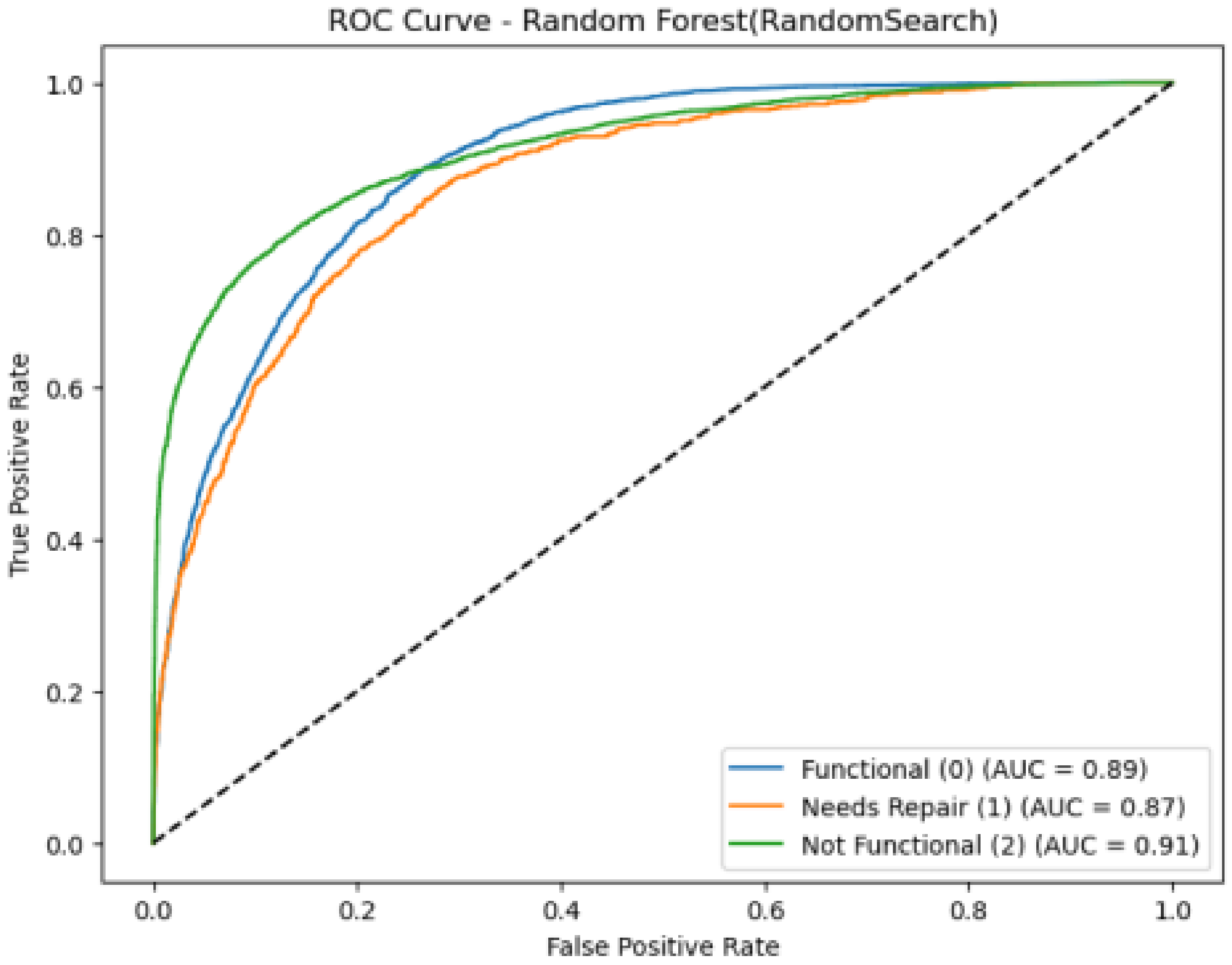
Max accuracy score of 76% 23% recall and 46% precision False positives increased

Further EDA showed model's feature importance Tuned hyperparameters

Weighted the model, and used SMOTE

Tanzania Waterwells Project





Random Forest was the optimal model

Initial accuracy score was 79%

Increase in precision score for Class 1 (60%) and Class 2 (84%)

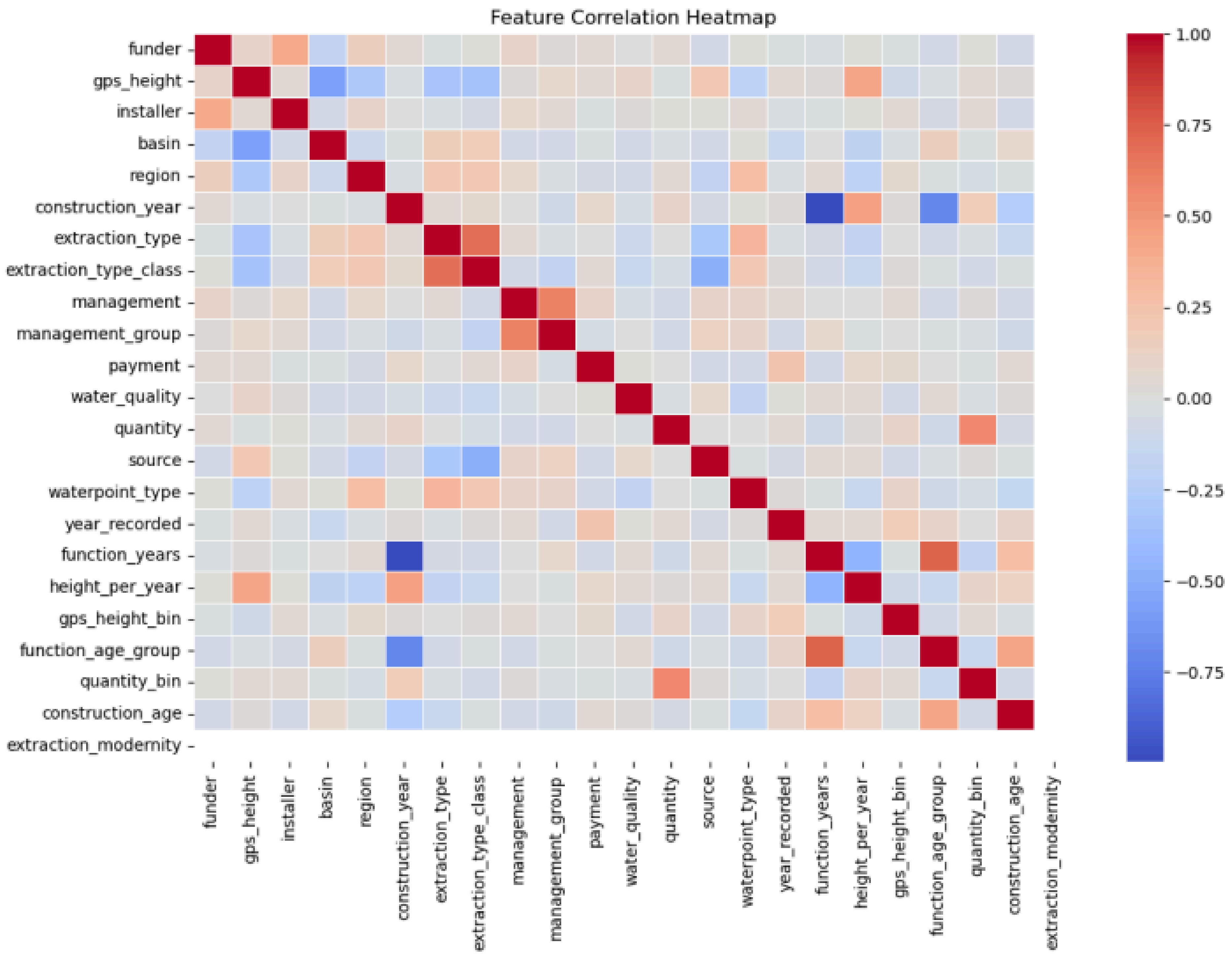
Recall score was 23% for Class 1

Tanzania Waterwells Project

Model identified Class 0 and Class 2 very well

Model struggled to distinguish Class 1 from other classes.

Correlating Features



Identify Correlating Features

Engineer features based on correlation

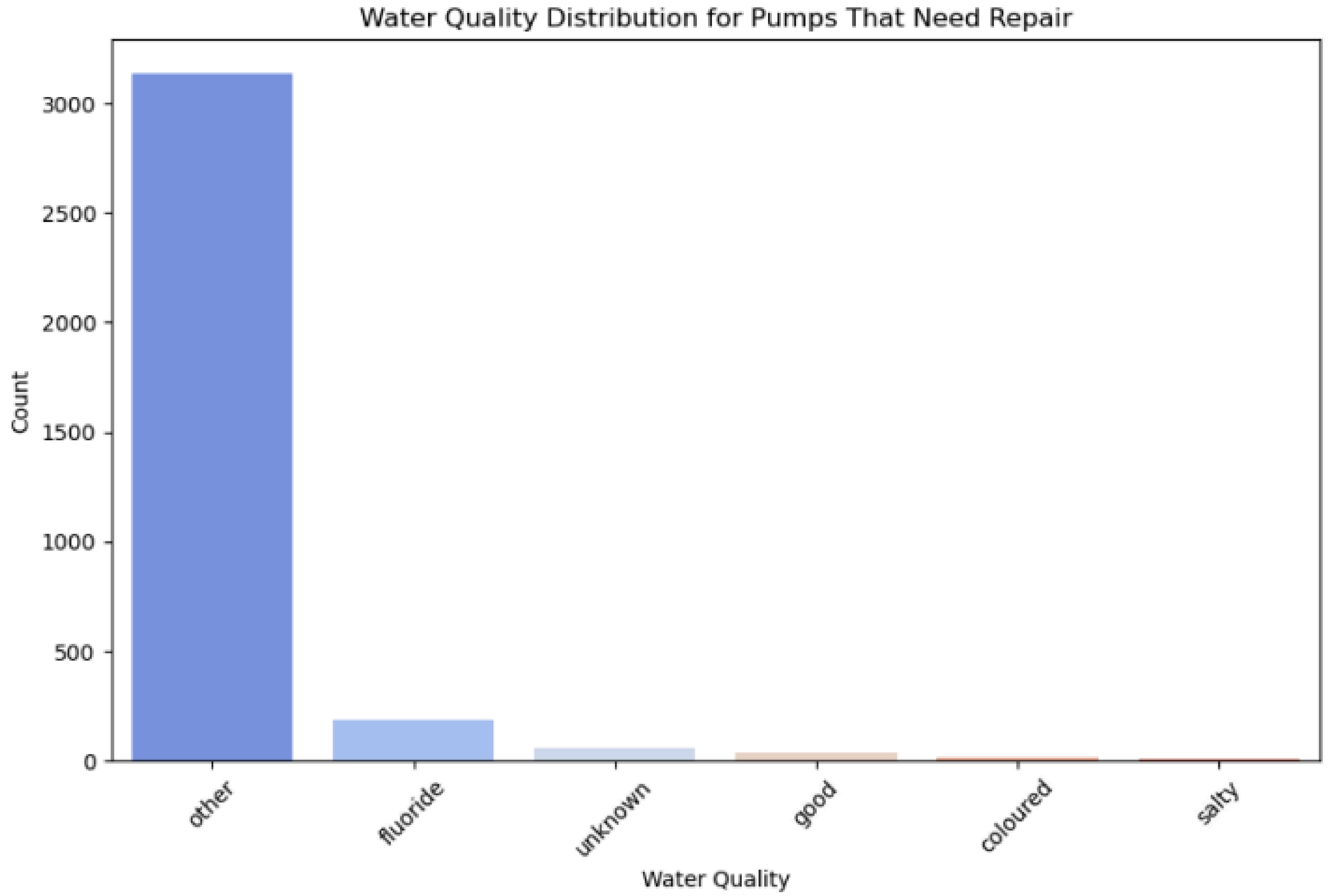
Binned features like gps_height and function_years.

Raised accuracy score to 80%

But we could do better...

Tanzanian Waterwells Project

Random Forest Error Analysis



Dived into false positives

Broken pumps = poor water quality

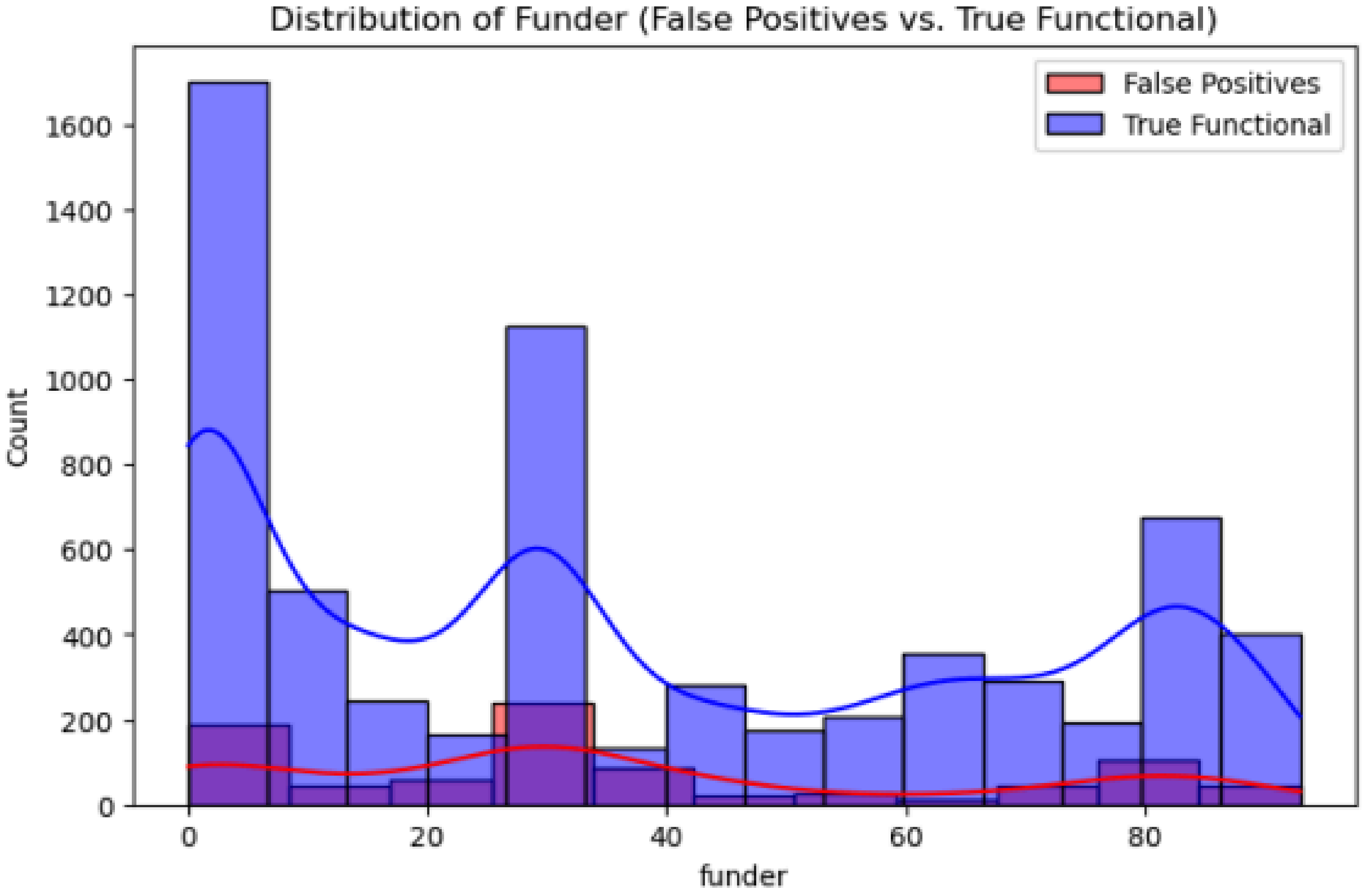
Grouped Class 1 with Class 2

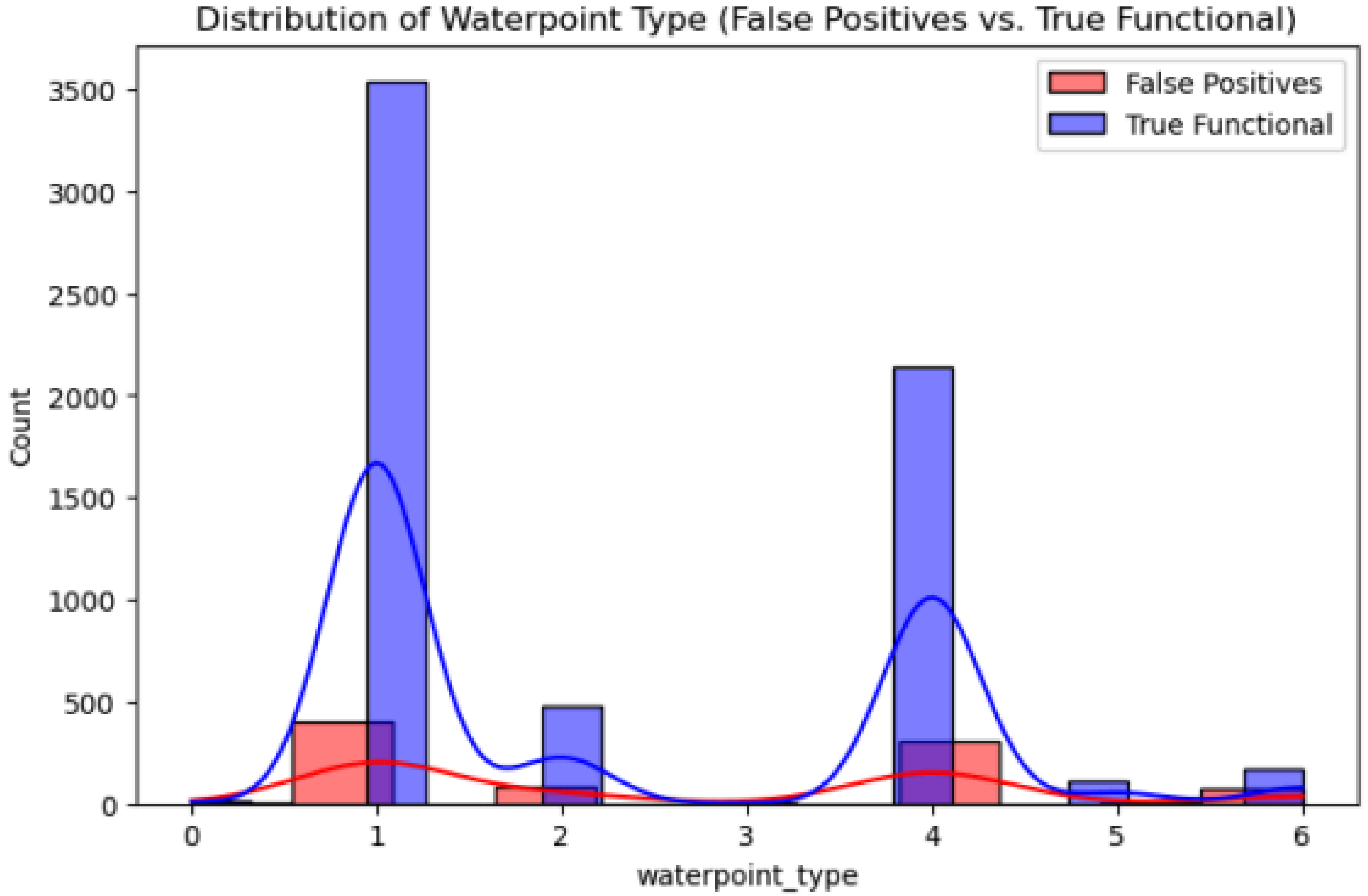
Model’s accuracy score was 81%

200 less false positives

.0025 STD shows model is stable

Error Analysis





Identifying

top false positive features

Feature engineering on top offenders

Accuracy score ultimately ends up around 81%

Help bring clean water to Tanzania

Thank you!