ReDI School of Digital Integration

Data Circle 2025
ReDI School

Water pump Functionality project

Group 7: James Donahue, Fatemeh Ebrahimi, Kateryna Ponomarova

Project Overview:

Goal: Build models to classify Tanzanian water pumps as “functional,” “needs repair,” or “non-functional.”

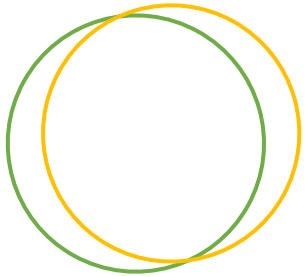
Why it matters: Reliable pumps → better maintenance planning → sustained access to clean water.

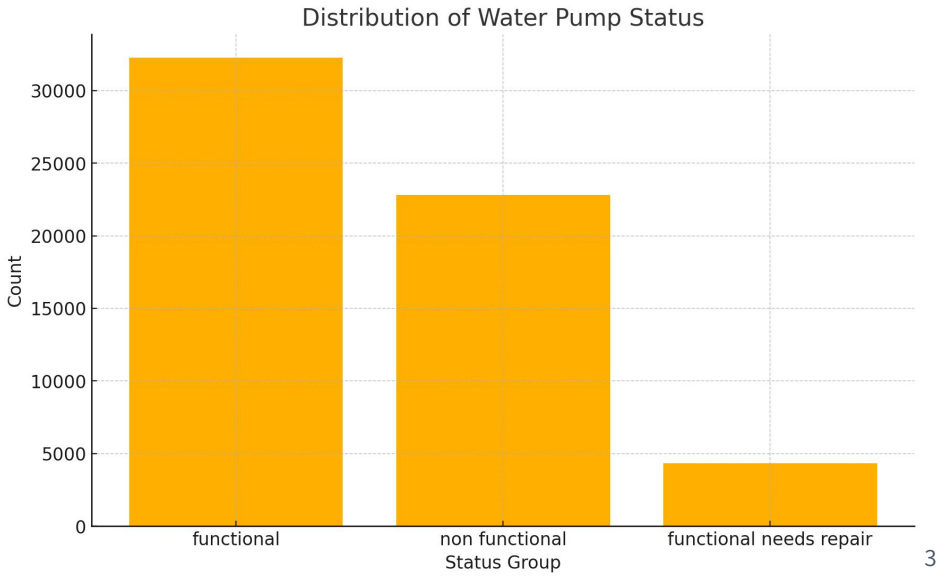
a typical hand pump in Tanzania

Data Introduction

Dataset at a Glance

- Samples: 59,400 pumps
- Features: 40 predictors (numeric, categorical, geospatial, temporal)
- Label: status_group (3 classes)





Overview of data:

1. Handling Missing & Raw Features
2. Numeric Transformations : amount_tsh, yearsq
3. Categorical Encodings
4. Feature selection: Keep only:
   - **Numeric**: amount_tsh, gps_height, longitude, latitude, population, construction_year
   - **Binary/coded**: extraction_type_class, payment, water_quality, quantity, source, waterpoint_type, scheme_management
   - **Dummies**: basin_*
5. Engineered Interaction Features

Feature Name	Computation
tshXpayment	amount_tsh × payment
extractXsource	extraction_type_class × source
popXtsh	population × amount_tsh
popXquant	population × quantity
popXsource	population × source
extractXheight	extraction_type_class × gps_height
typeXsource	waterpoint_type × source
typeXyear	waterpoint_type × construction_year
yearXpop	construction_year × population
quantXsource	quantity × source
yearsq	√(construction_year + 1)

Model selection

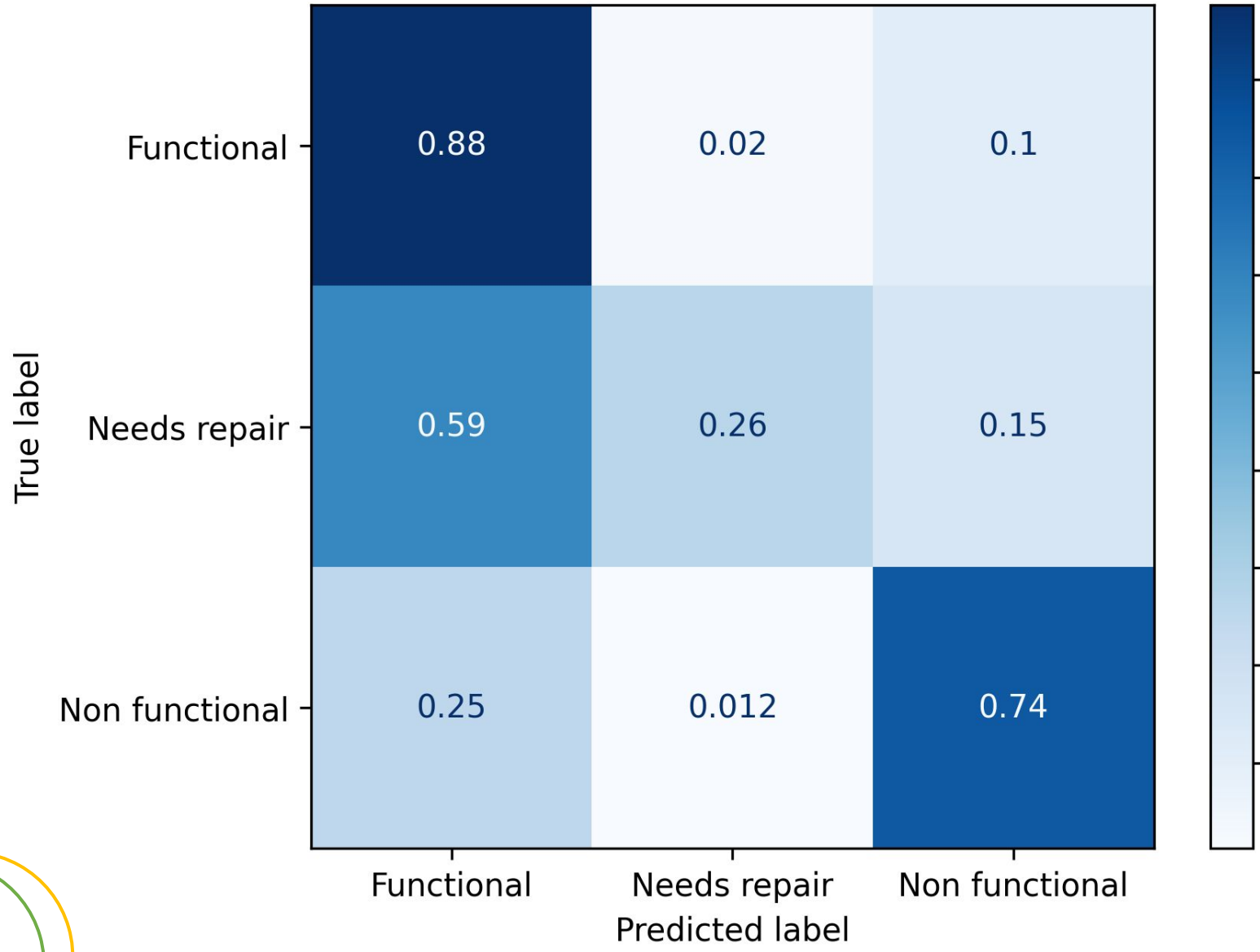
```
Best Estimator: XGBClassifier( ...
    eval_metric=auc,
    gamma=0,
    learning_rate=0.3, max_bin=None,
    max_delta_step=1,
    max_depth=8,
    min_child_weight=None,
    num_class=3, ...)
```

Model selection

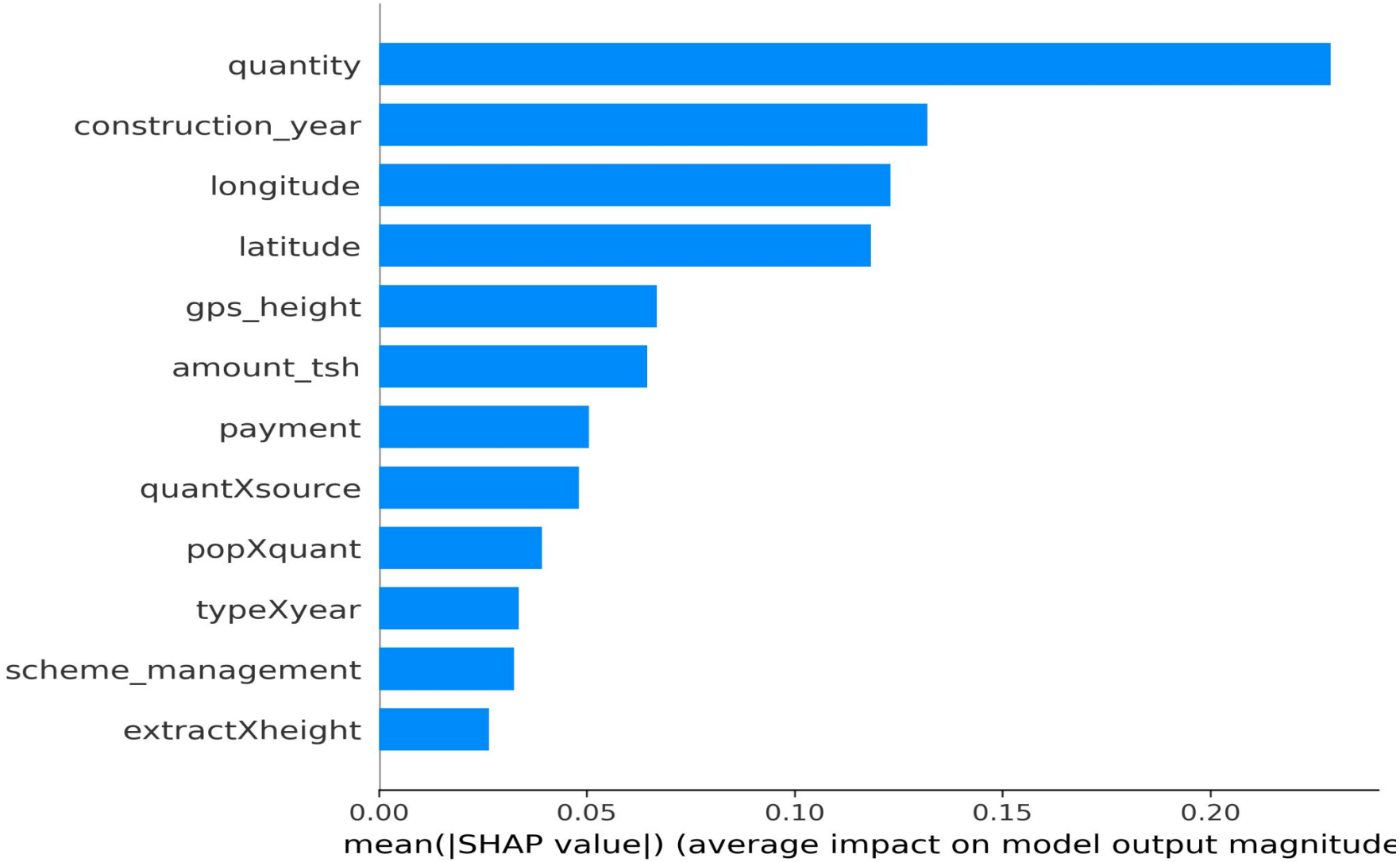
Classification Report on Test Set:

	precision	recall	f1-score	support
functional	0.77	0.88	0.82	9678
functional needs repair	0.55	0.25	0.34	1295
non functional	0.81	0.73	0.77	6847
accuracy			0.78	17820
macro avg	0.71	0.62	0.64	17820
weighted avg	0.77	0.78	0.77	17820

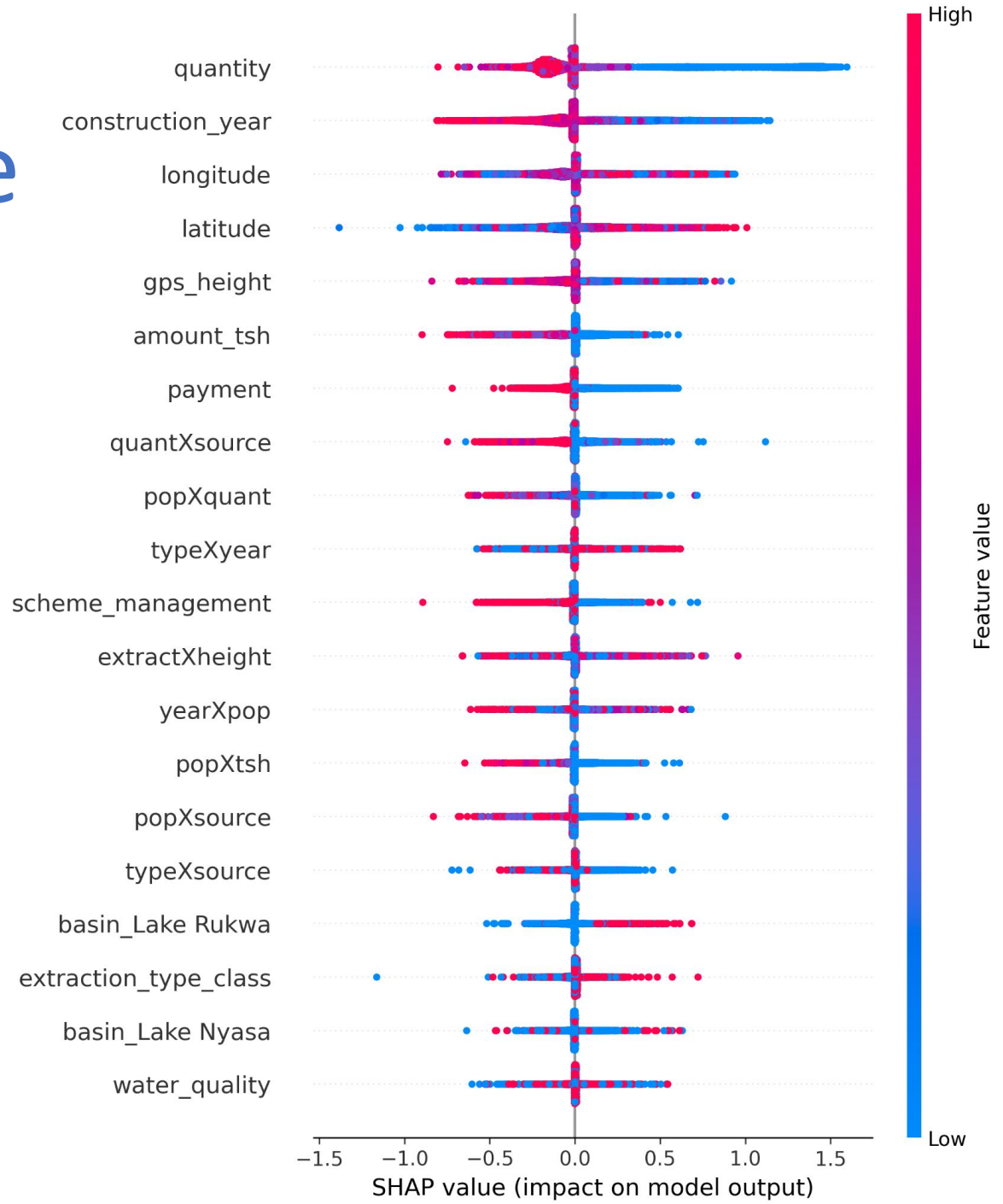
Prediction:


Functional
Needs repair
Non functional
True label
0.88
0.02
0.1
0.59
0.26
0.15
0.25
0.012
0.74
Functional
Needs repair
Non functional
Predicted label

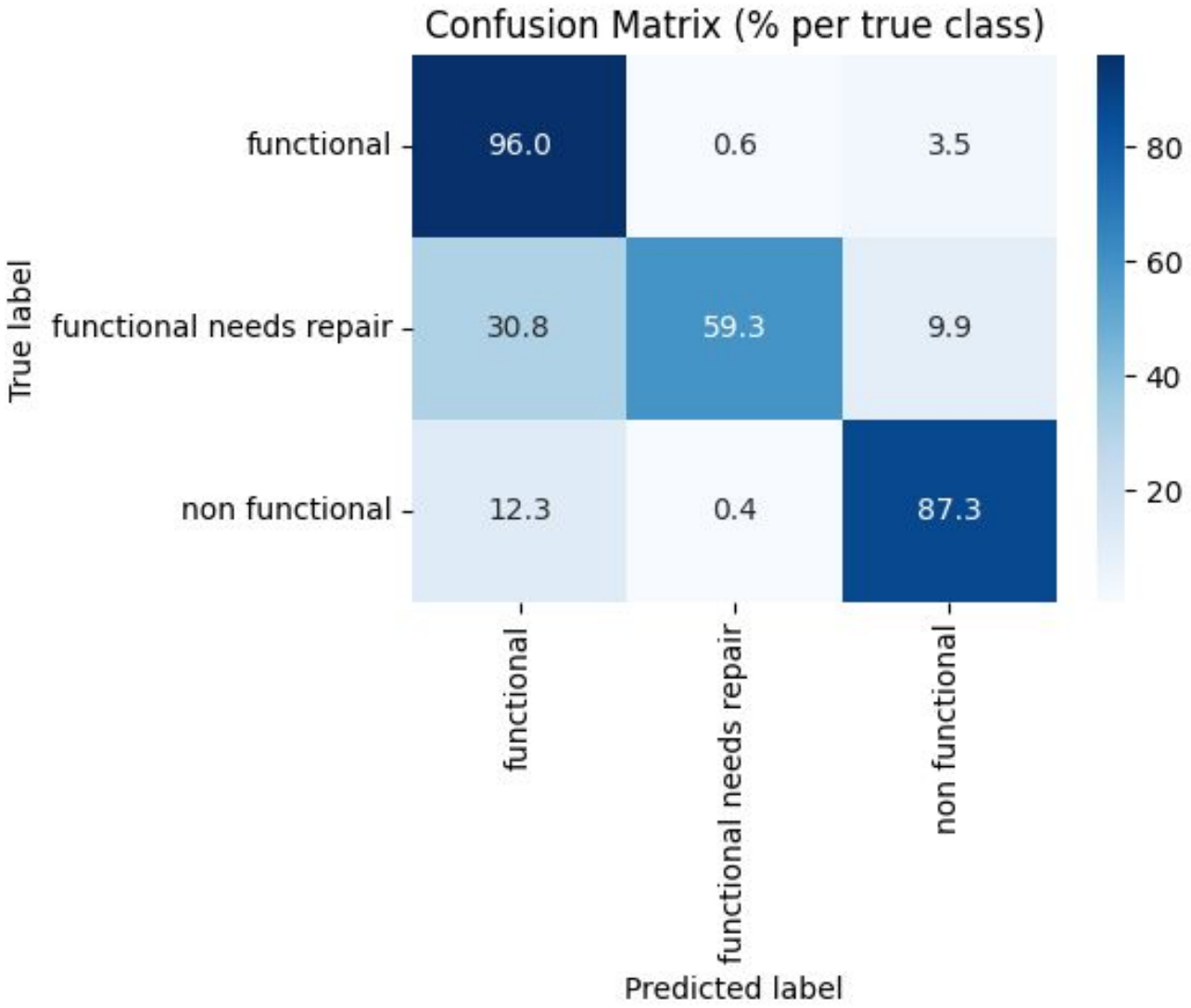
Feature Importance



Feature Importance



Prediction:


Confusion Matrix (% per true class)
True label
functional
functional needs repair
non functional
96.0
0.6
3.5
30.8
59.3
9.9
12.3
0.4
87.3
80
60
40
20
functional
functional needs repair
non functional
Predicted label

Submission score for XG-boost

Best score	Current rank
0.7047	#6939