Predicting Waterpoint Functionality in Tanzania

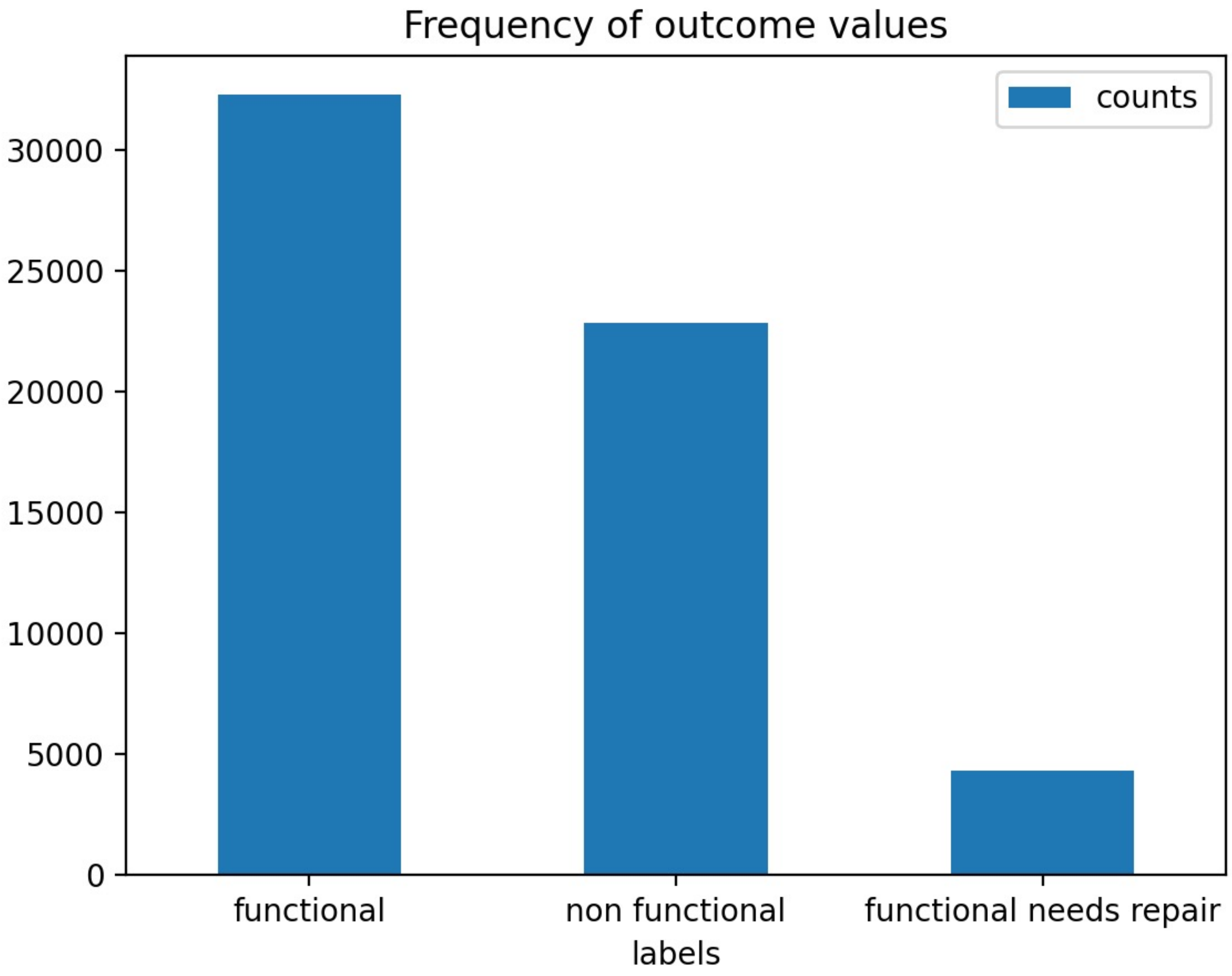
Caitlin Snyder

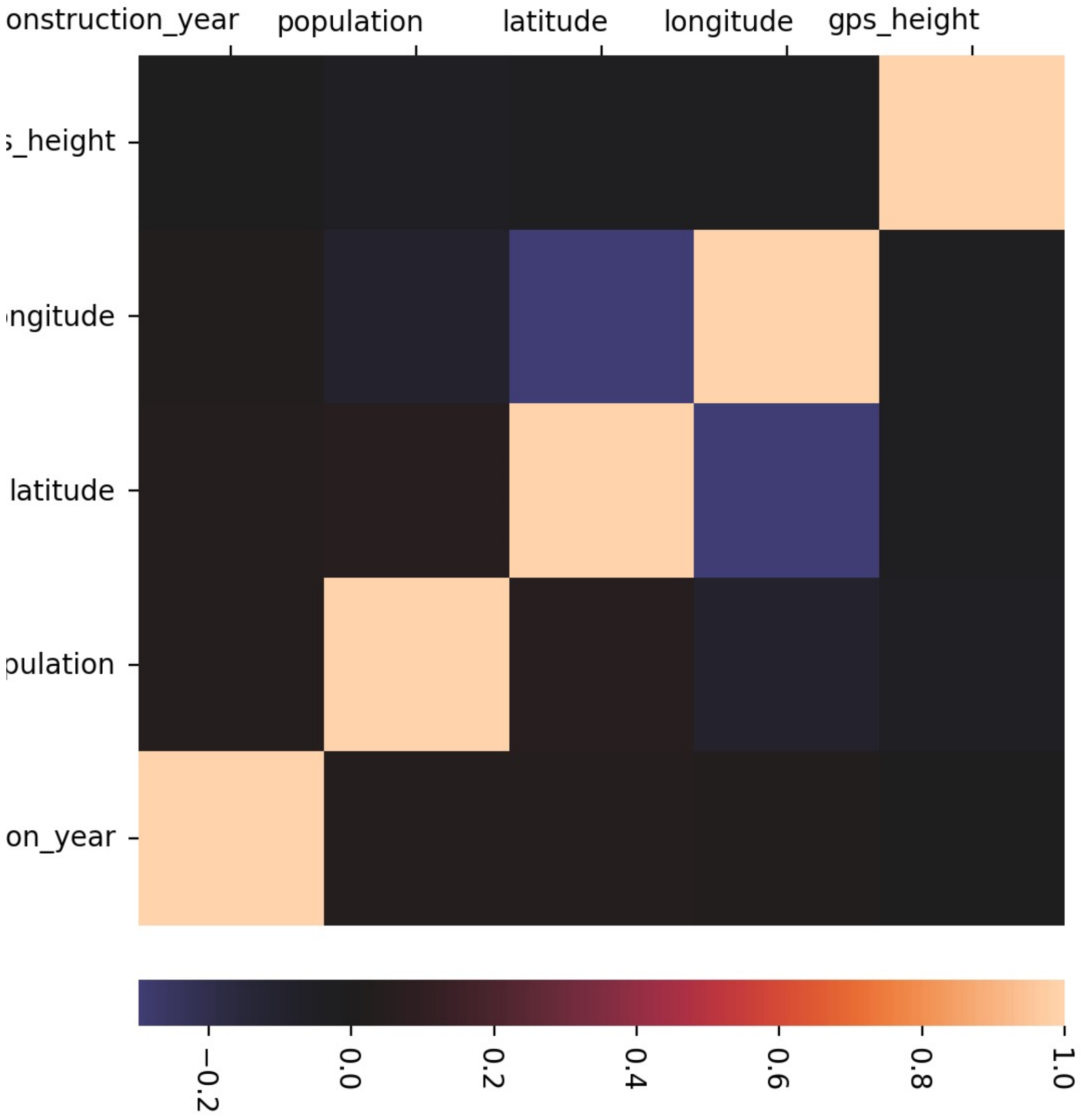
Flat Iron - Data Science

Module 3

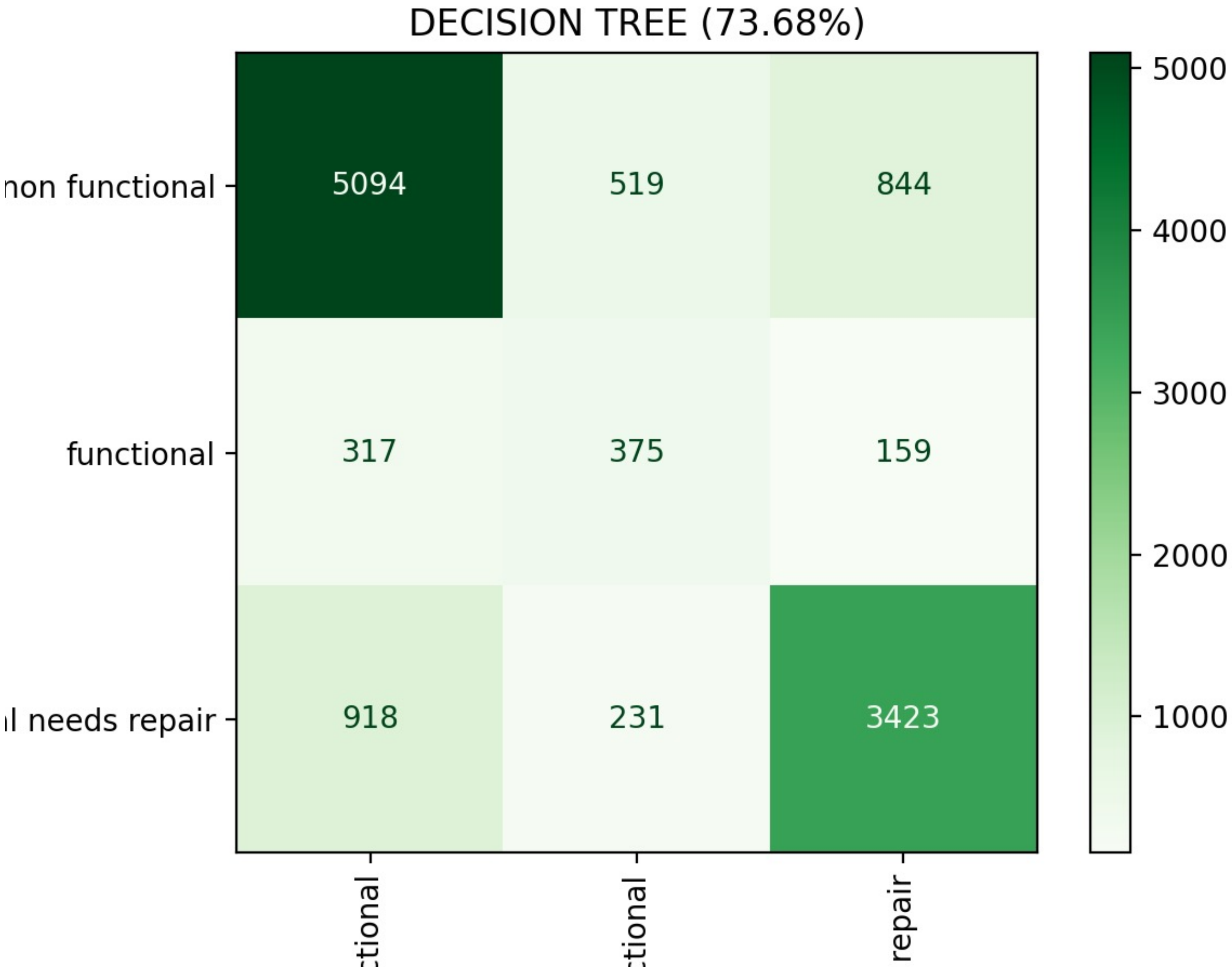
Research Question

Given a set of waterpoint characteristics, can we accurately predict whether that waterpoint is functional, non-functional, or functional but in need of repair?

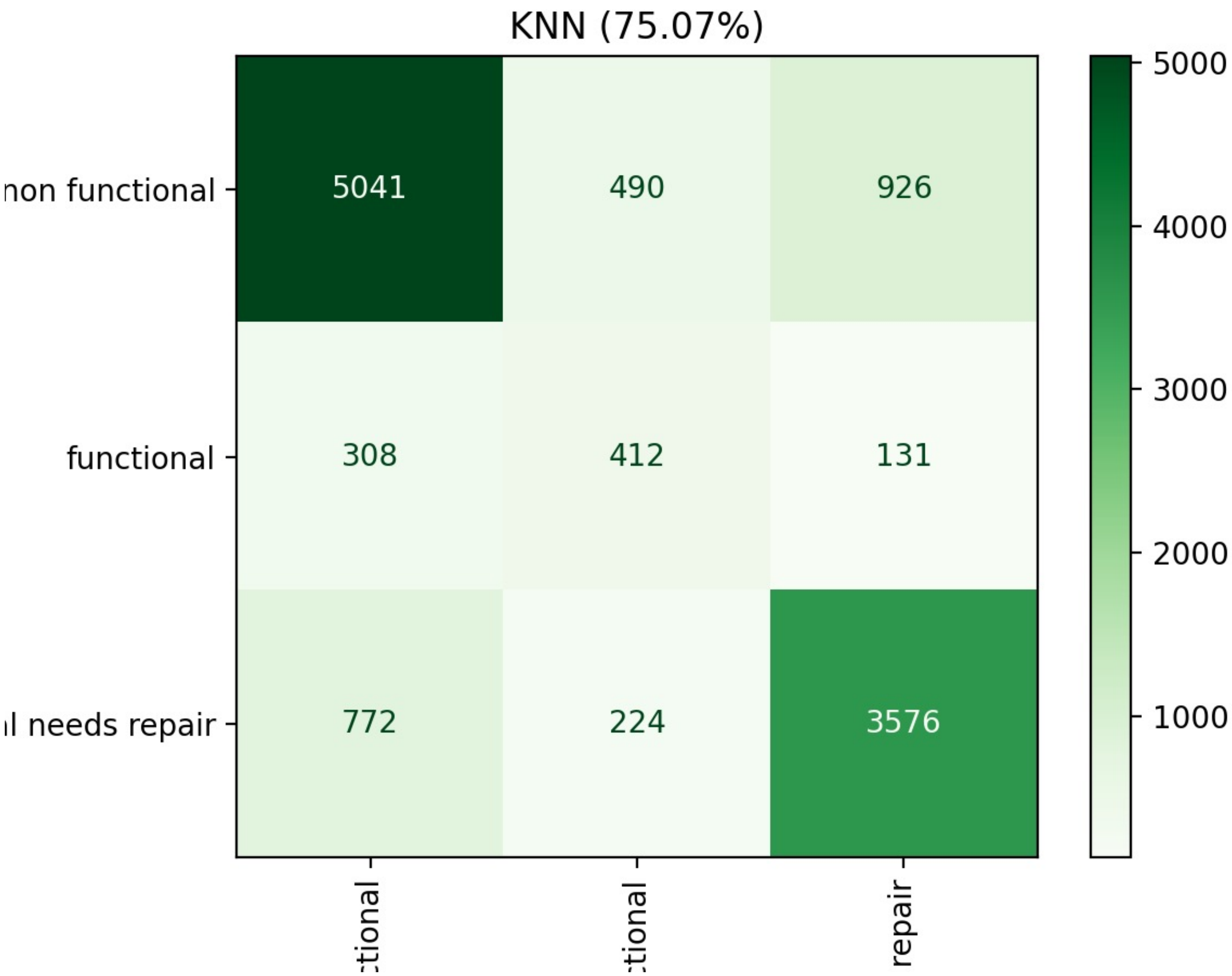

Frequency of outcome values
counts
30000
25000
20000
15000
10000
5000
0
functional
non functional
functional needs repair
labels


onstruction_year
population
latitude
longitude
gps_height
_height
ngitude
latitude
pulation
on_year
−0.2
0.0
0.2
0.4
0.6
0.8
1.0

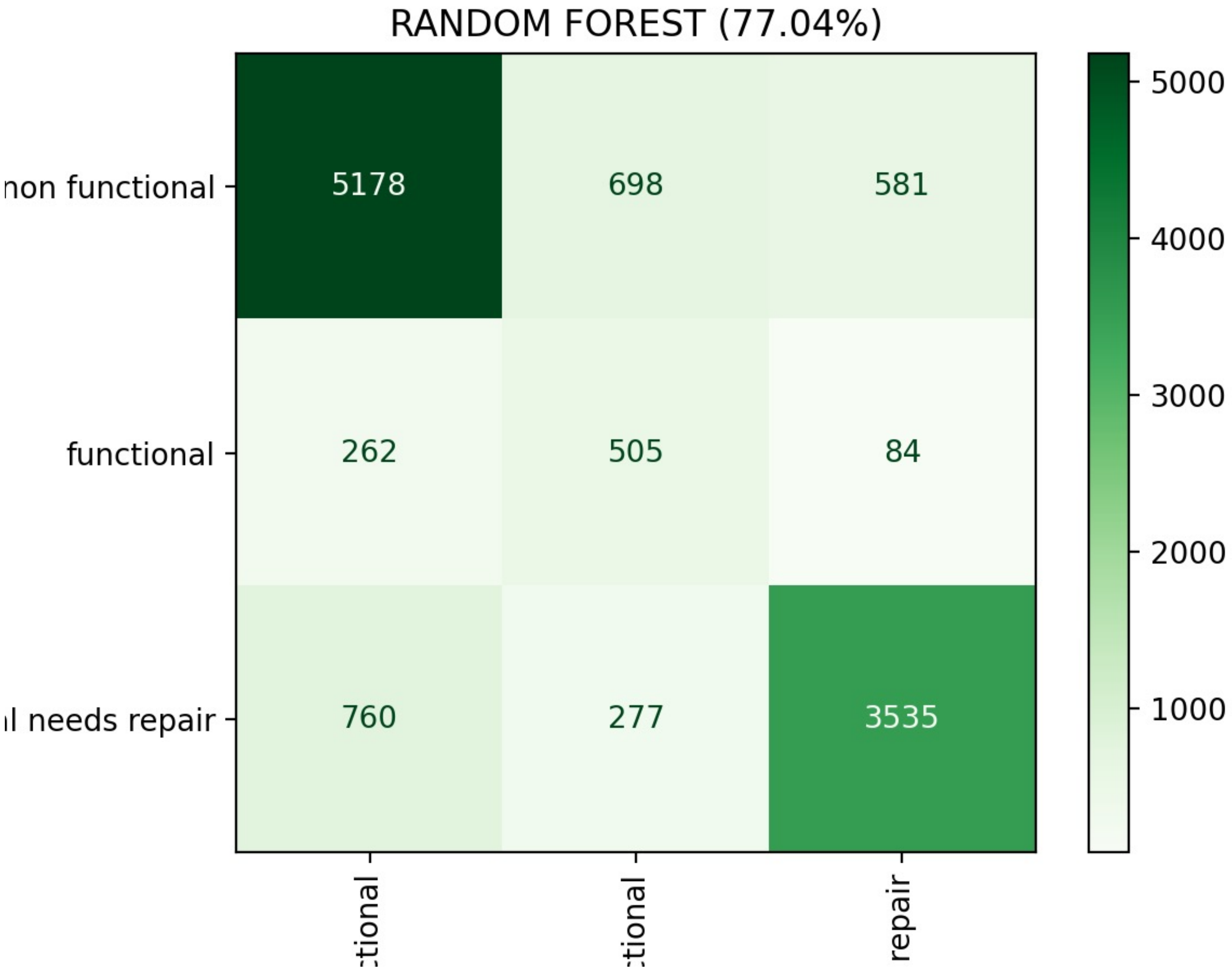
Evaluating the model: *Decision Tree*



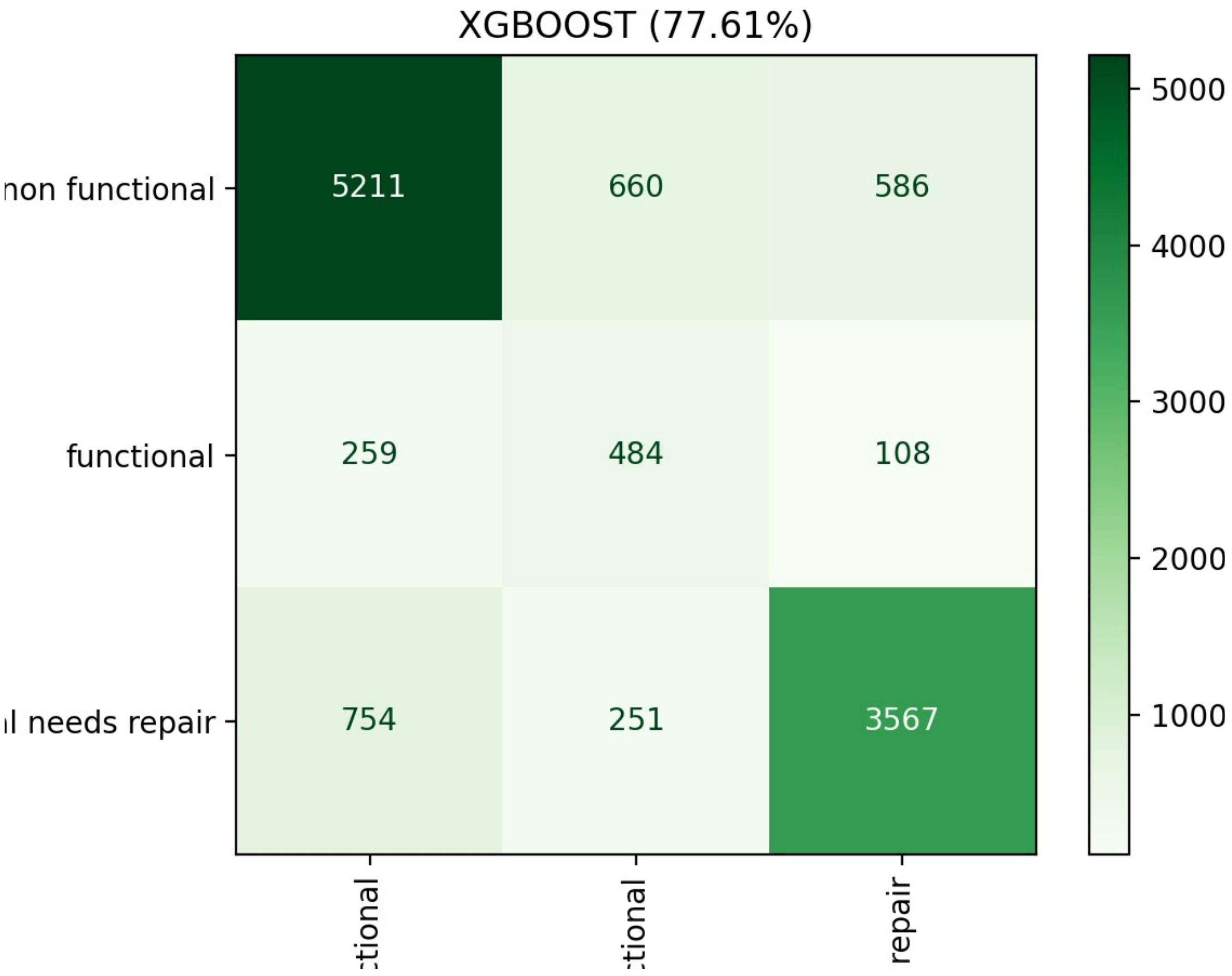
Evaluating the model: KNN



Evaluating the model: *Random Forest*



Evaluating the model: *XGBoost*



XGBOOST (77.61%)

	tional	tional	repair
non functional	5211	660	586
functional	259	484	108
l needs repair	754	251	3567

Take-aways

- Although XGBoost and Random Forest produced very very similar accuracy scores, XGBoost *slightly* (0.2%) outperformed Random Forest.
- The intensive memory requirements of these classifiers would be best served using cloud computing resources.

Future avenues for exploration

- Does inclusion of elevation enhance the model's predictive accuracy?
- What is the effect of including frequency of conflict incidences related to water-resource usage?

Sources

- https://towardsdatascience.com/fuzzywuzzy-find-similar-strings-within-one-column-in-a-pandas-data-frame-99f6c2a0c212
- https://towardsdatascience.com/fuzzywuzzy-fuzzy-string-matching-in-python-beginners-guide-9adc0edf4b35
- https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/
- https://stackabuse.com/the-naive-bayes-algorithm-in-python-with-scikit-learn/
- https://medium.com/@erikgreenj/k-neighbors-classifier-with-gridsearchcv-basics-3c445ddeb657
- https://medium.com/vickdata/a-simple-guide-to-scikit-learn-pipelines-4ac0d974bdcf
- https://datascience.stackexchange.com/questions/60862/if-i-have-negative-and-positive-numbers-for-a-feature-should-minmaxscaler-be-1

Sources *cont'd*

- https://stackoverflow.com/questions/41899132/invalid-parameter-for-sklearn-estimator-pipeline
- https://lifewithdata.com/2021/04/02/how-to-build-machine-learning-pipeline-with-scikit-learn-and-why-is-it-essential/
- https://stackoverflow.com/questions/63467815/how-to-access-columntransformer-elements-in-gridsearchcv
- https://openscoring.io/blog/2020/10/24/converting_sklearn_imblearn_pipeline_pmml/
- https://imbalanced-learn.org/stable/over_sampling.html
- https://towardsdatascience.com/imbalanced-class-sizes-and-classification-models-a-cautionary-tale-part-2-cf371500d1b3

Thank you!