Tanzania Water Well Project

Cameron Ladd and Wonuola Abimbola

Problem

- Tanzania, as a developing country, struggles with providing clean water to its population of over 57,000,000.

- There are many waterpoints already established in the country, but some are in need of repair while others have failed altogether.

- The Government of Tanzania has hired us to make predictions about the conditions of the water well status using information contained in the provided dataset

Goal

- Build a model that accurately predicts the conditions of a water well status using information provided in the data

- Maximize Recall score for class 0 and 2 while maximizing precision score for class 1.

Data Exploration

- Early in the data exploration process, we discovered that is a class imbalance in the data set

- There are more instances of the 'functional' (1)and 'non-functional' (0) class than there are of the 'functional but needs repair' class (2) in the data.


class distribution
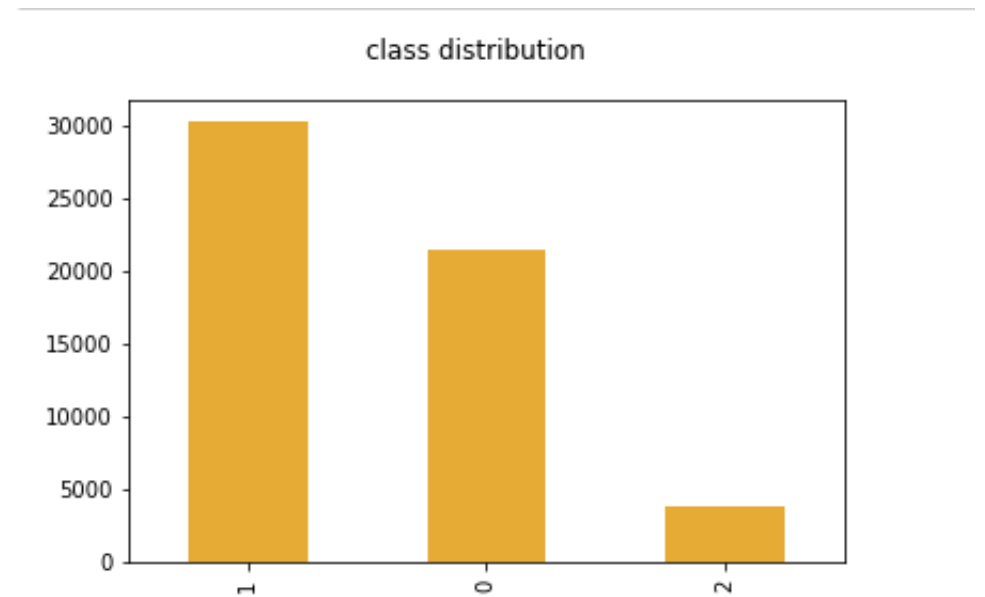
Basin

- There are some basins that are more likely to have higher amount of functional wells.

- Two regions that stick out are 'Ruvuma/Southern Coast' and 'Lake Rukwa', because seem to have a higher number of non-functional wells compared to functional ones!
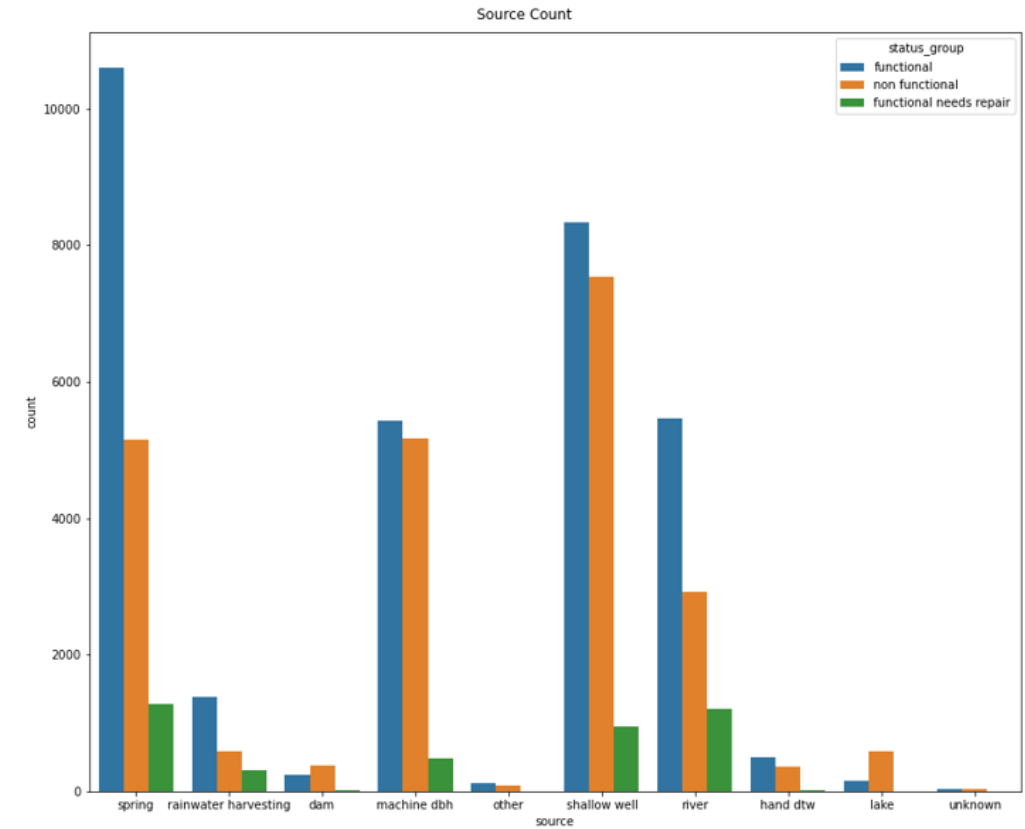

Basin Count
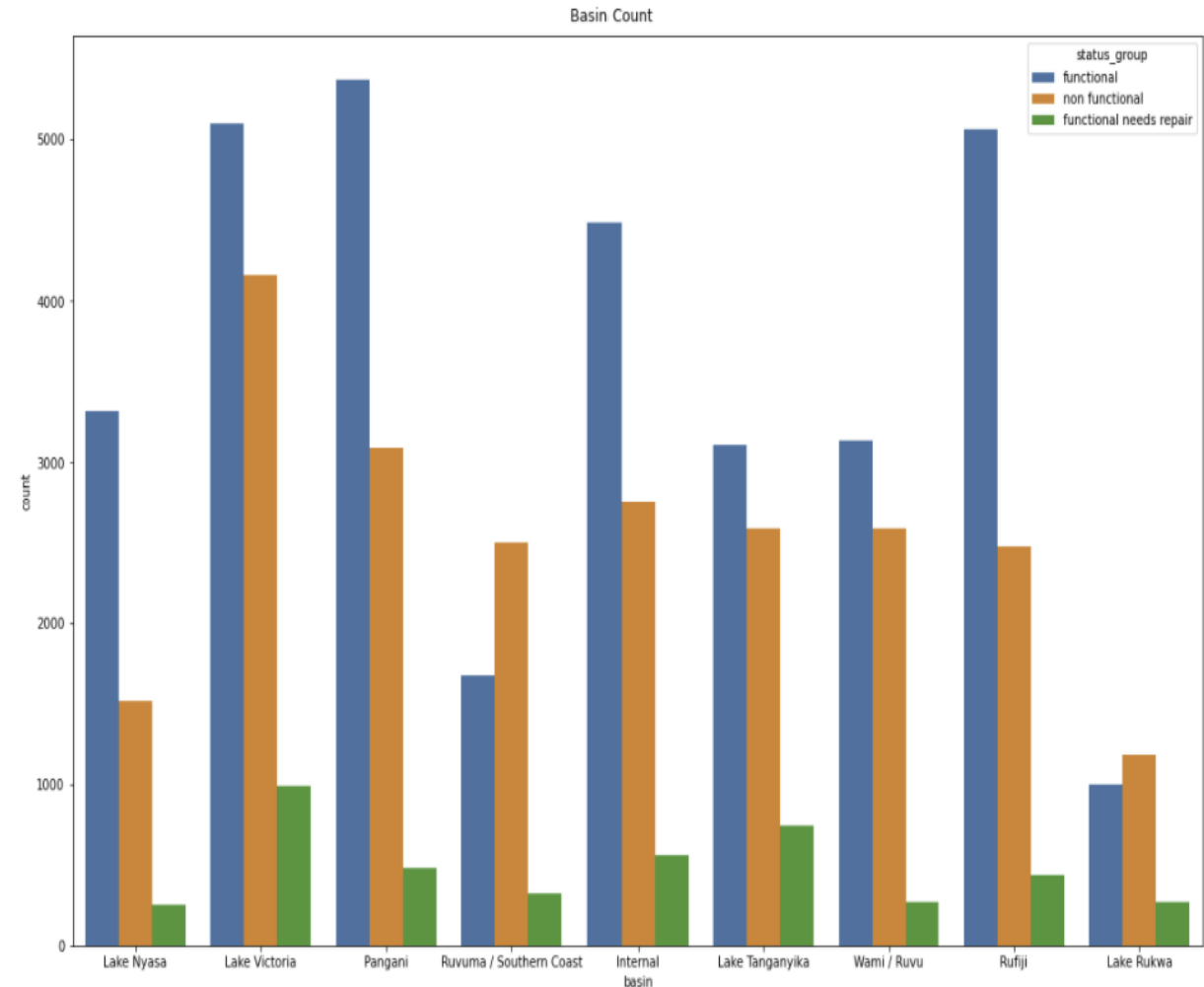
Quality Group

- Clearly it's more likely for a water well to be functional when the water quality is good
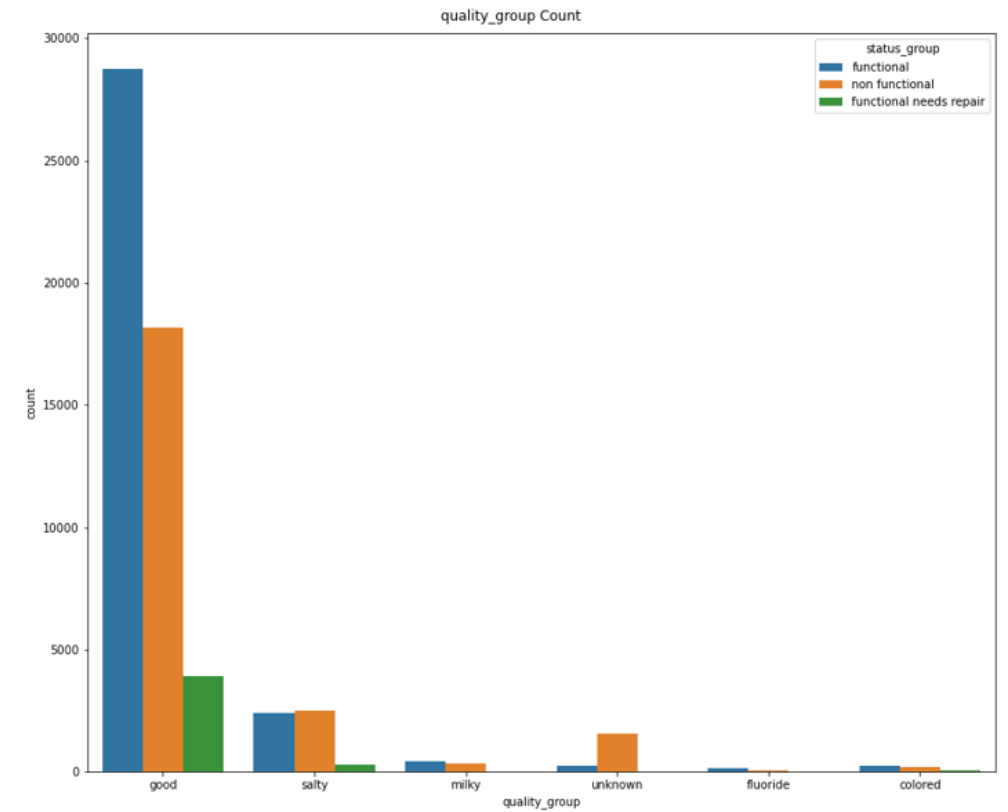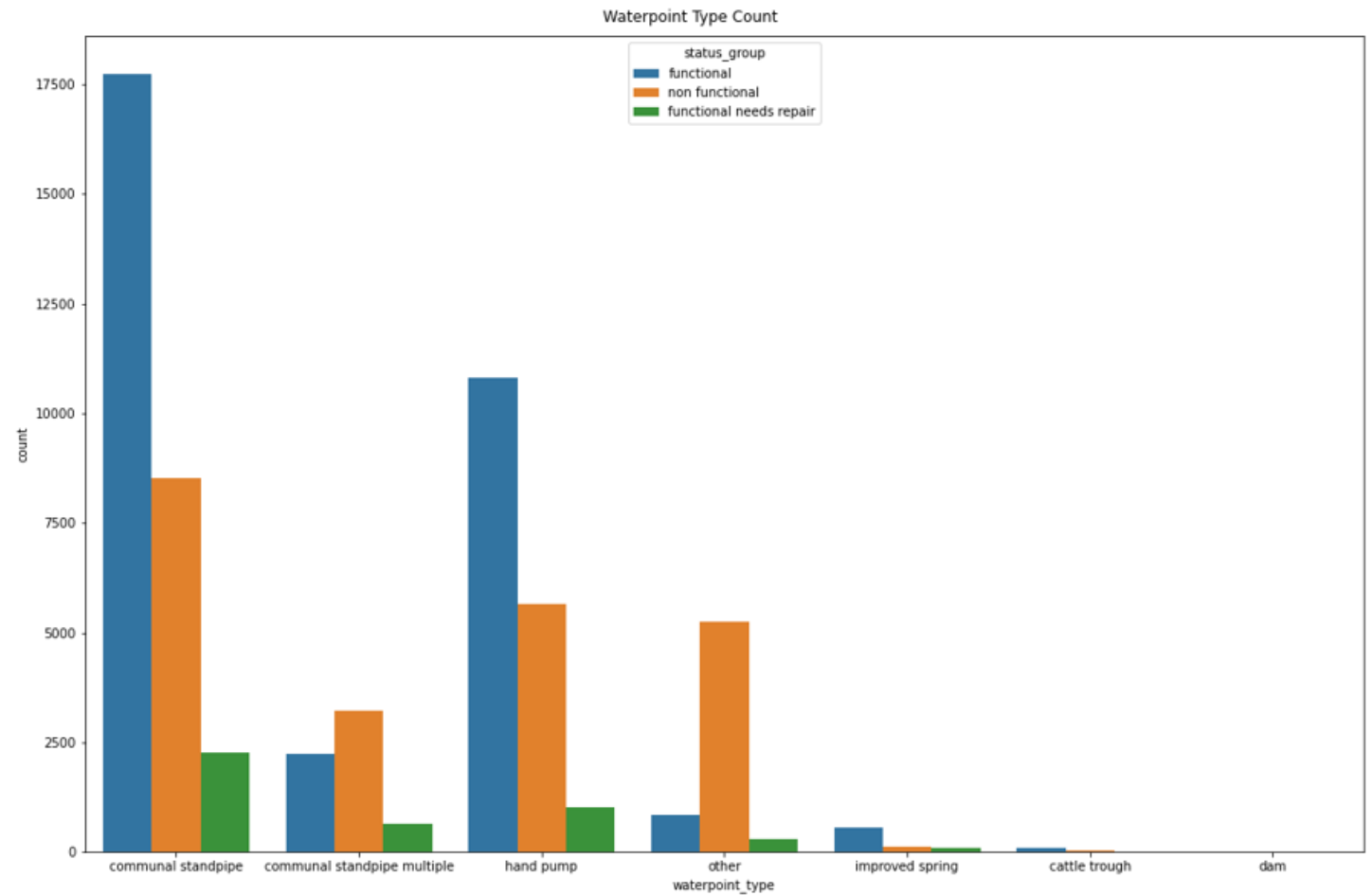
Source

- Spring and shallow well sources have the highest number of functional wells overall.

- Sources like dams, lakes and rainwater aren't used for water wells often

Waterpoint Type

- Communal standpipe has the highest number of functional wells followed by handpump.
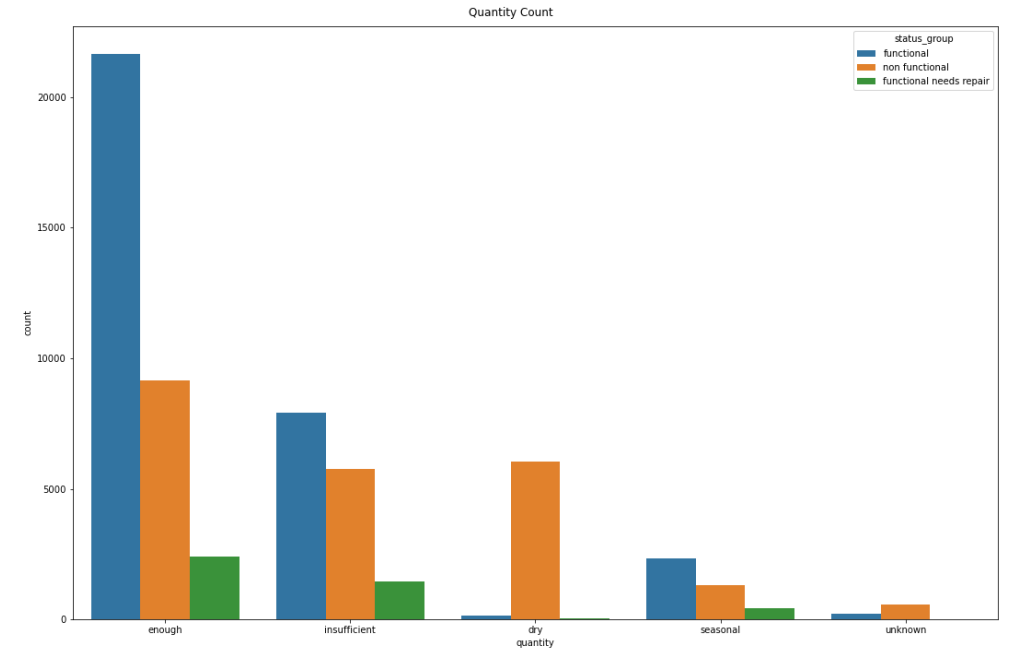

Waterpoint Type Count

Region

- Most of the regions have a higher rate of functional wells with the Irina region being the highest.

- Regions like Rukwa and Lindi have a high incidence of non-functional to functional waterwell ratio.

Region Count
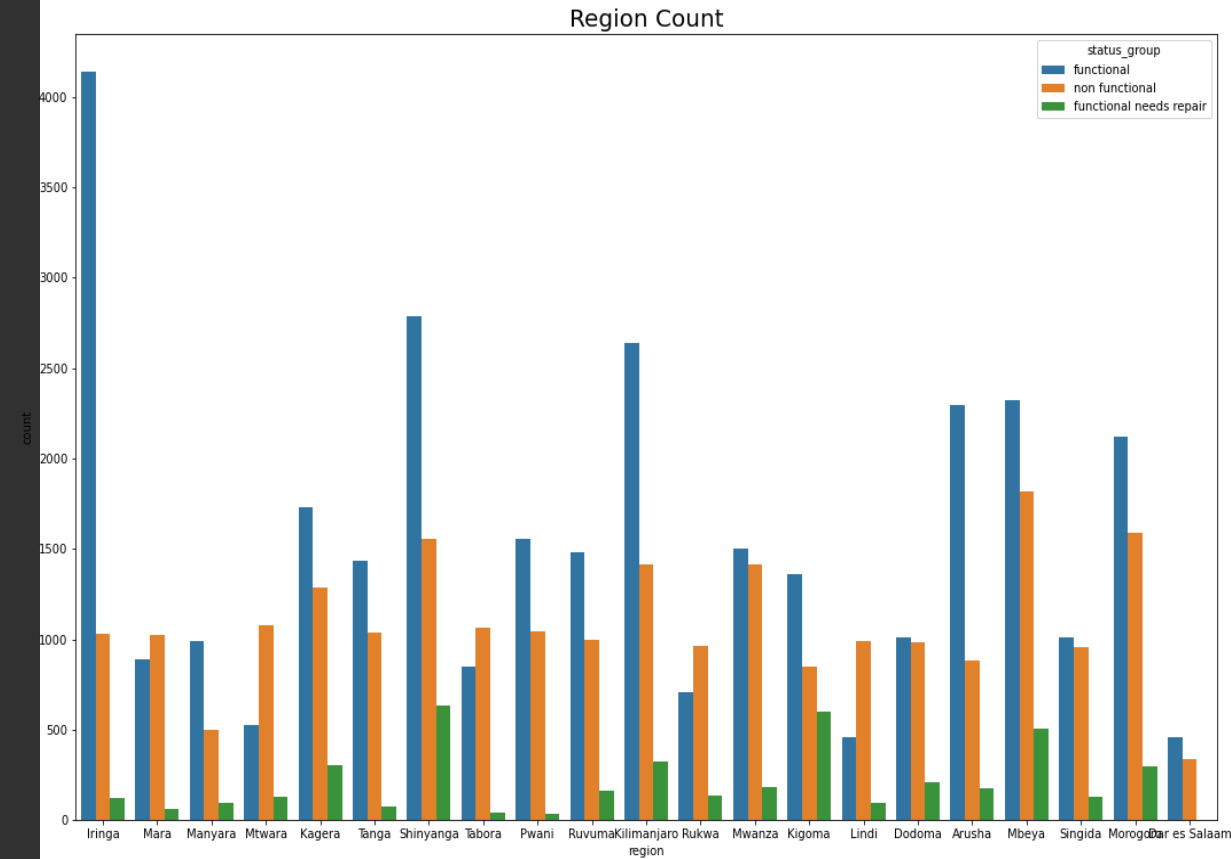
Quatity

- Waterwells that have "enough" water quantity are more likely to be functioning.

Quantity Count

Extraction Type Class

- The method of extraction of the water seems to have an effect on the functionality of the water well.

Extraction Type Class Count
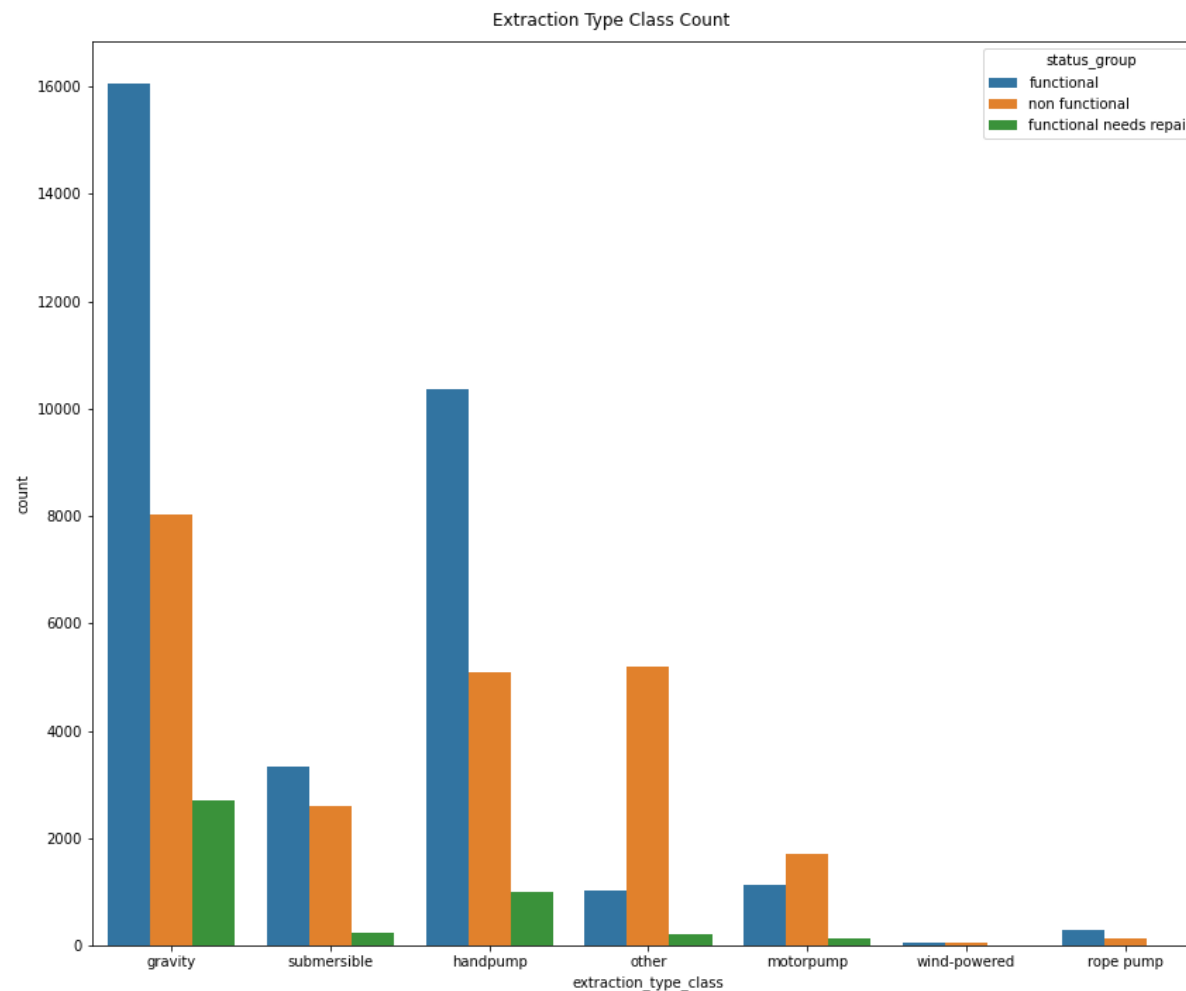
Analysis

- Overall, the model we decided to make use of, had an accuracy of 79%, the highest accuracy out of any of our models.

- Another reason why this model was chosen over the others was because it had a high precision score for the functional class while also having a high recall score for the non-functional and the needs repair class

Conclusion

- In conclusion, the best model based on our research is XGBoost.

- We chose this based on accuracy score, recall score for classes 0 and 2, and precision score for class 1.

- Baseline accuracy of 54.3%!

- Our overall accuracy was 79%, significantly higher!