TANZANIA WATER WELL PROJECT

WATER CHANGES EVERYTHING

by Winfred Kabuya

We believe in a world where every single person has access to clean and safe water

When a community gets access to clean water, it can change just about everything

Overview

The project develops a classifier to predict the condition of water wells in Tanzania. It targets NGOs and the Tanzanian Government for identifying wells in need of repair and informing future construction decisions.

BUSINESS PROBLEM

This project addresses the lack of access to clean water in Tanzania by developing a predictive model to classify the condition of water wells. It aims to assist NGOs and the Tanzanian government in allocating resources effectively for well repair and construction. The goal is to improve access to safe water for the population of Tanzania.

DATA

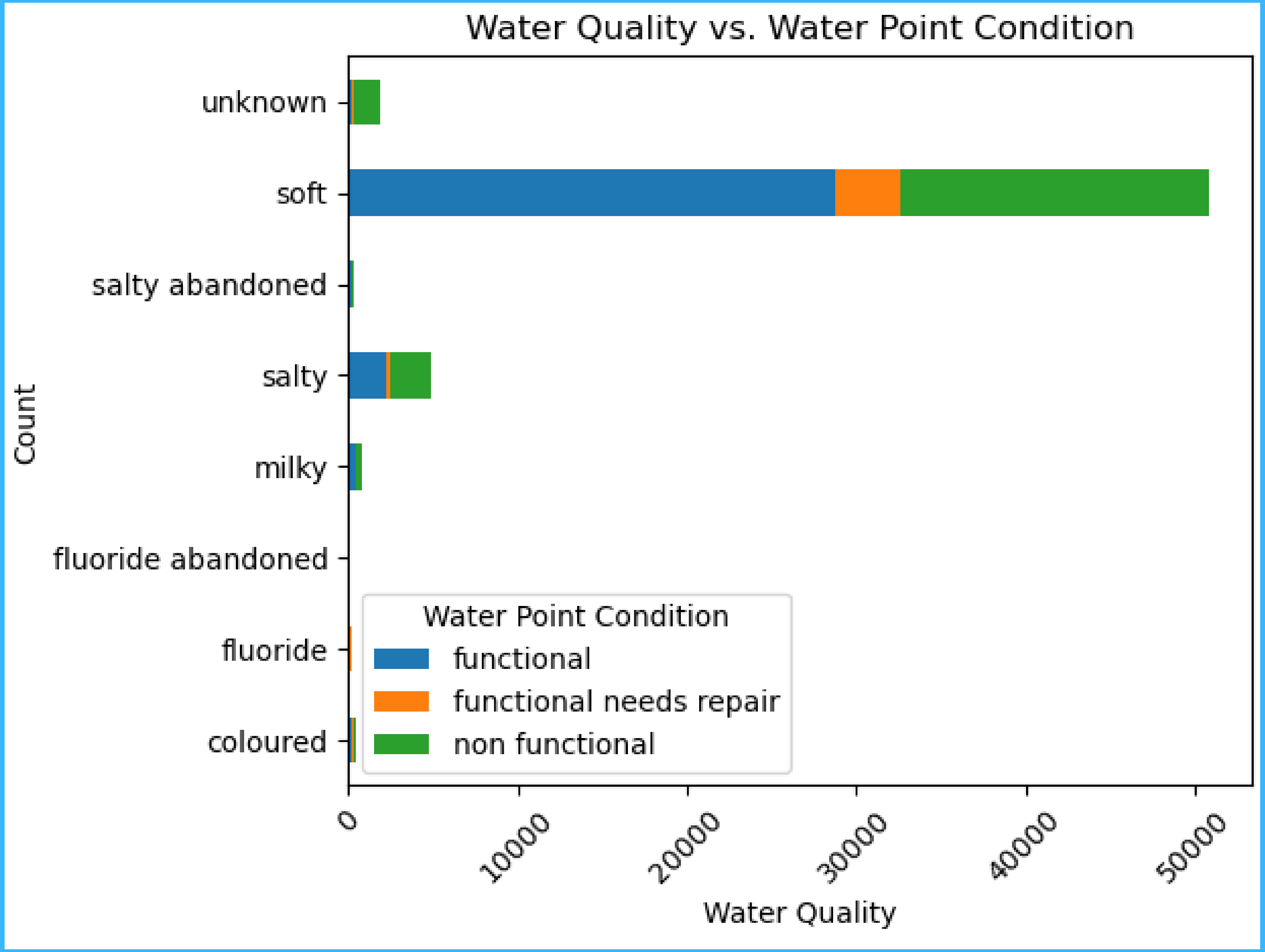
We shall use available datasets from Taarifa and the Tanzanian Ministry of Water to determine the best approach to resolve our problem whereby:

- test_set_values represent the independent variables (features) for the test set, where predictions need to be made.
- training_set_labels represent the dependent variable (status_group) for each row in the training set values. It represents the current operating condition of the waterpoints.
- training_set_values: Represent the independent variables (features) for the training set, used to train the predictive model.

Modelling

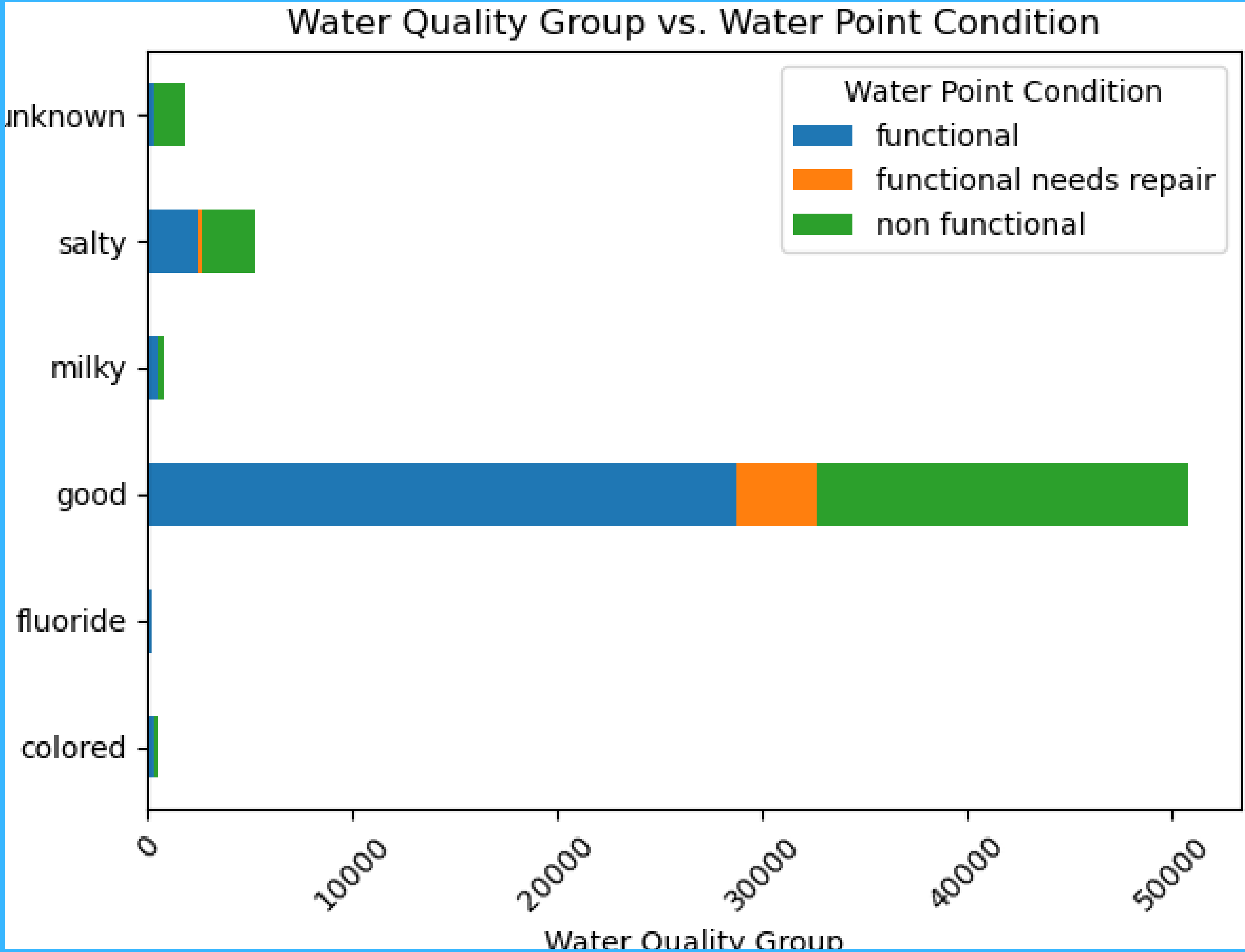
After analyzing the dataset we got the following results:

- Areas with soft water have more functional pumps compared with other areas. Also the areas with soft water have more non functional pumps



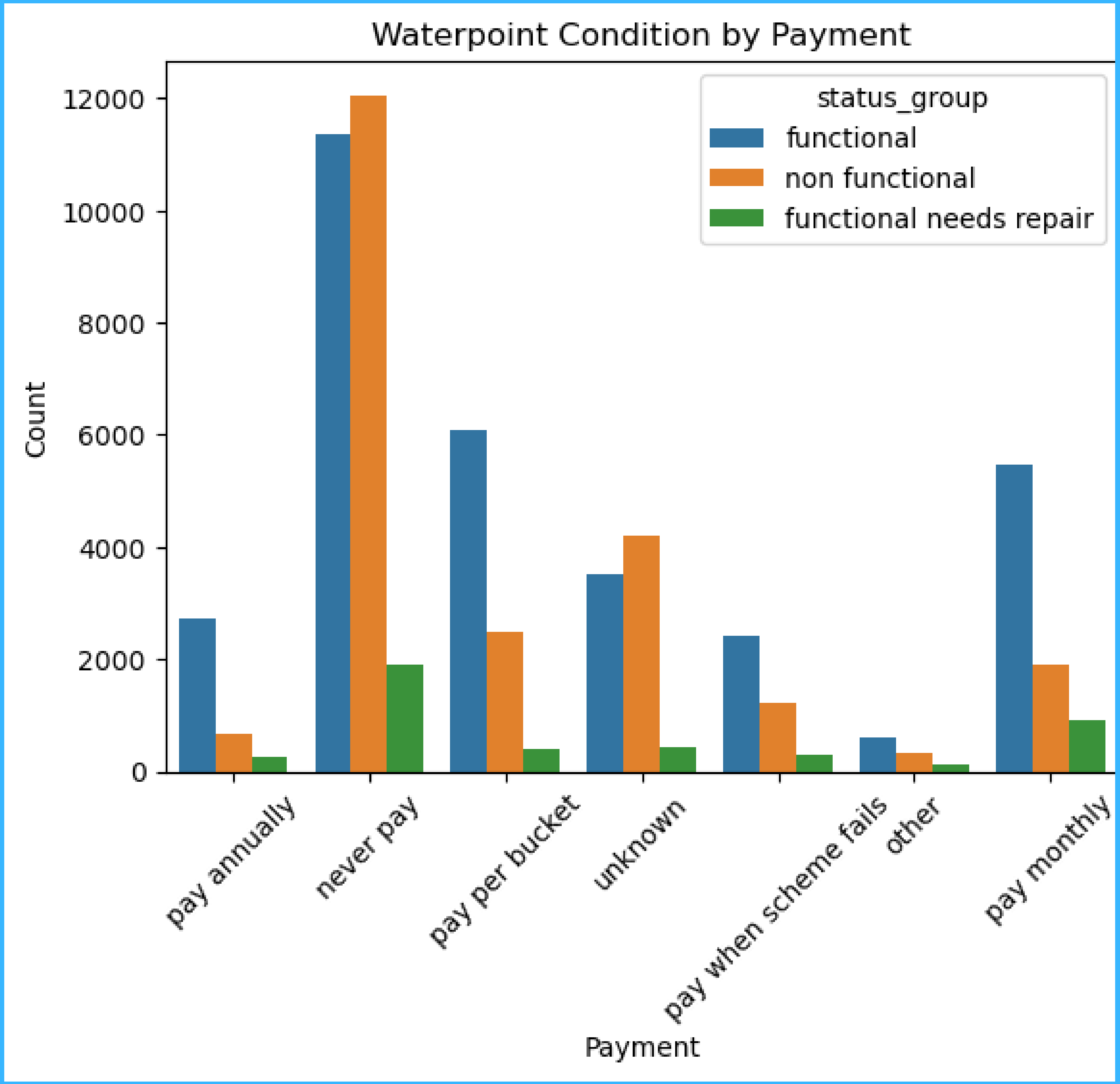
Modelling cont.

- Areas with good water have more functional pumps compared with other areas with different qualities of water. Also the areas with good water have a lot of non functional pumps

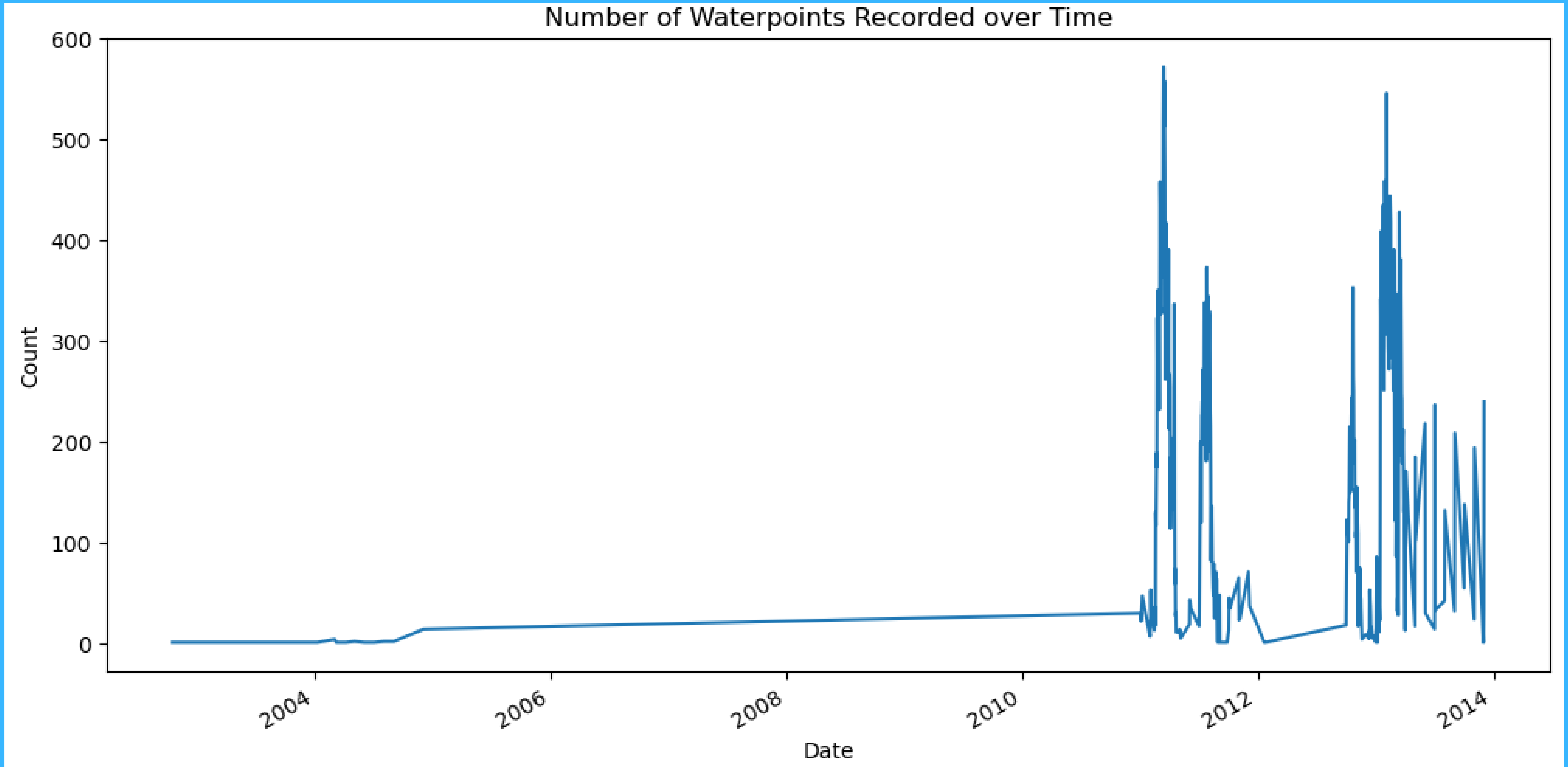


Modelling cont.

- Waterpoints that charge per annum and monthly have more functional pumps and less non fucntional ones compared to those do not charge which have more non functional than functional pumps

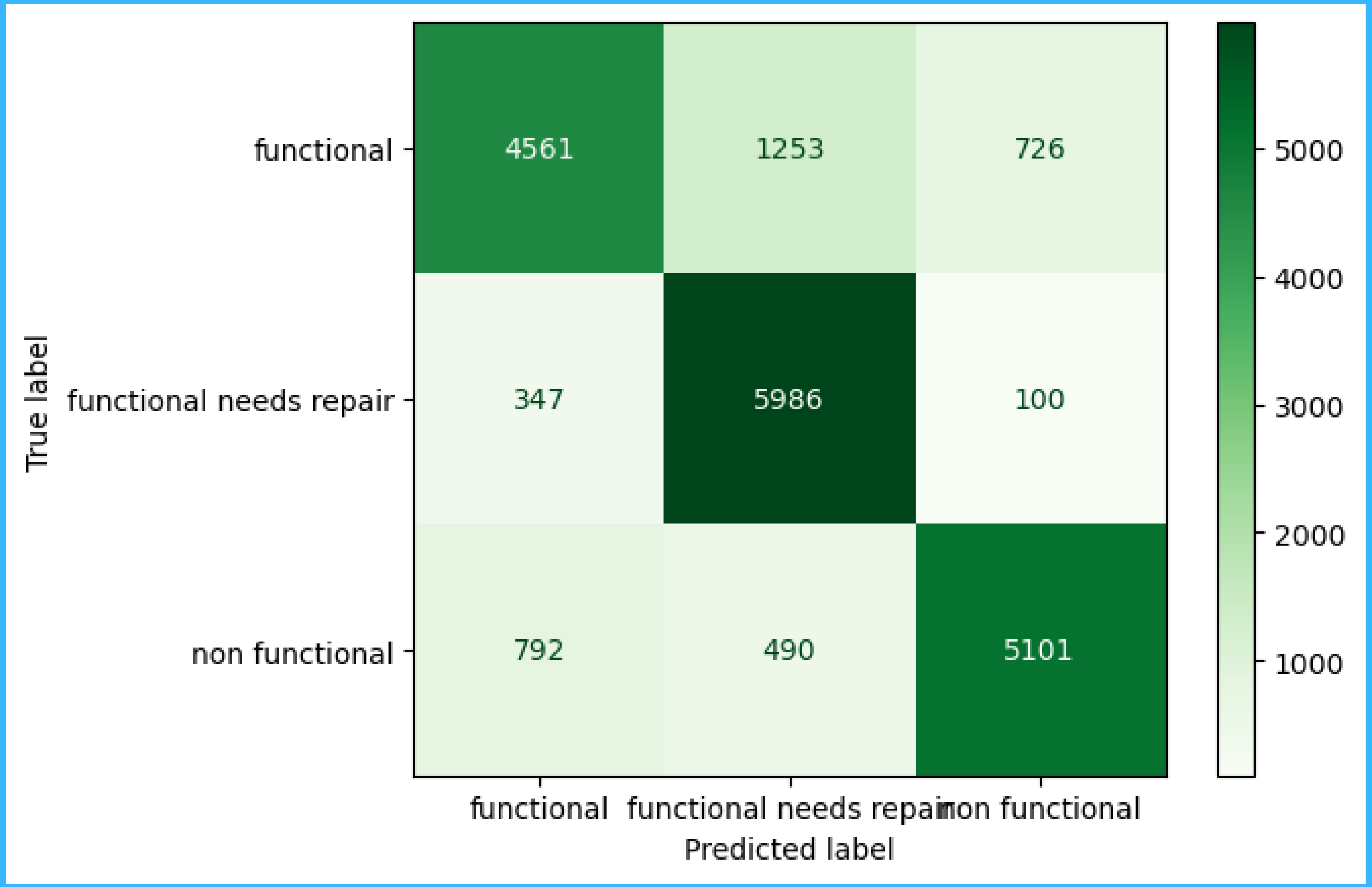


Evaluation



There was an increase in the number of water points between the year 2011 and 2014

Evaluation



True label \ Predicted label	functional	functional needs repair	non functional
functional	4561	1253	726
functional needs repair	347	5986	100
non functional	792	490	5101

The One-hot Coding classifier which was our best model achieved an accuracy score of 80.74% on the validation set, correctly predicting the class for approximately 80.74% of instances . The model's macro-averaged and weighted average F1-scores are also 80.74%, indicating consistent performance across all classes and considering the class imbalance in the dataset.

Recommendations

- Machine learning techniques are valuable for analyzing complex datasets like the Tanzania well dataset.
- Data cleaning, feature engineering, and hyperparameter tuning are important steps in preparing the data and optimizing the model's performance.
- Addressing class imbalance and utilizing cross-validation can improve the model's accuracy and generalization.
- Collecting more data and deepening domain knowledge can further enhance the analysis and understanding of the target variable.

Conclusion

- Collect More Data: Especially for the minority classes, to provide a more representative and balanced dataset for training the model.

- Improve the performance of the model by combining data preprocessing techniques, feature engineering, model selection, and careful evaluation.

- Regular iteration and experimentation with different approaches are essential to achieve better results in predicting the status of water points in Tanzania.