PREDICTING THE FUNCTIONALITY OF TANZANIA WATER WELLS

Overview

- The project aims to develop 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.

Problem Statement

- Access to safe and consistent drinking water is a major challenge in Tanzania.
- To solve this issue, the Tanzanian government has made investments in the building of water wells in collaboration with a number of NGO's.
- However, the sustainability and functionality of these wells remain uncertain, with many of them falling into disrepair or becoming non-functional over time.

Overview

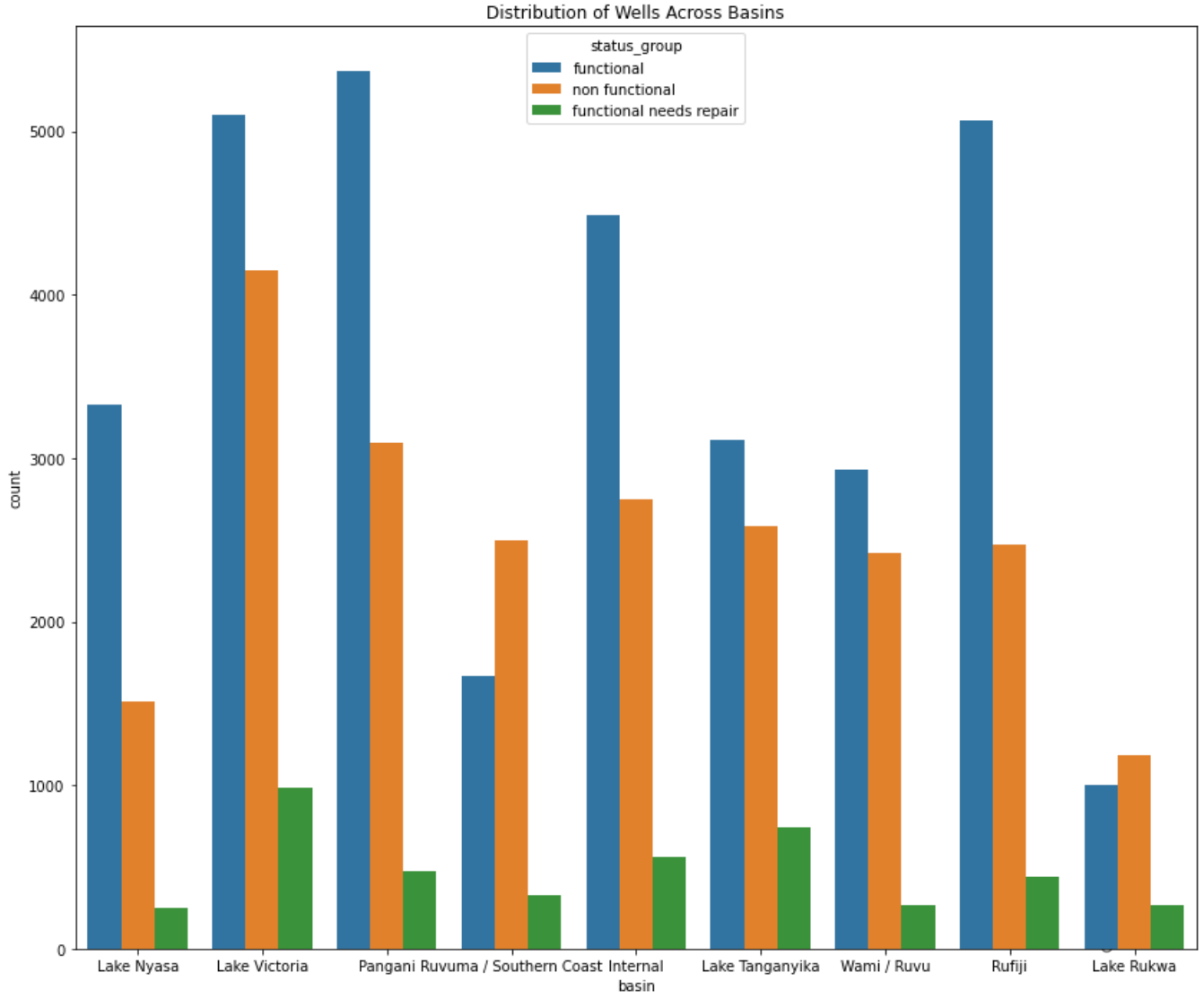
- The project aims to develop a classifier to predict the condition of water wells in Tanzania.
- It targets NGOs and the Tanzanian Government for identifying wells in need of repair and informing future construction decisions.

Data

- The data is sourced from Taarifa and the Tanzanian Ministry of Water.
- For the purposes of our evaluation, I am utilizing the Training Set Labels and Training Set Values.
- This will be followed by testing it on unseen data(Testing set values).

Data Analysis

- I analysed different columns and used my target variable as the hue.
- The columns include basin, funder, installer and population.
- Various eda visuals were used: bar charts, count charts, scatter plots, heat maps.

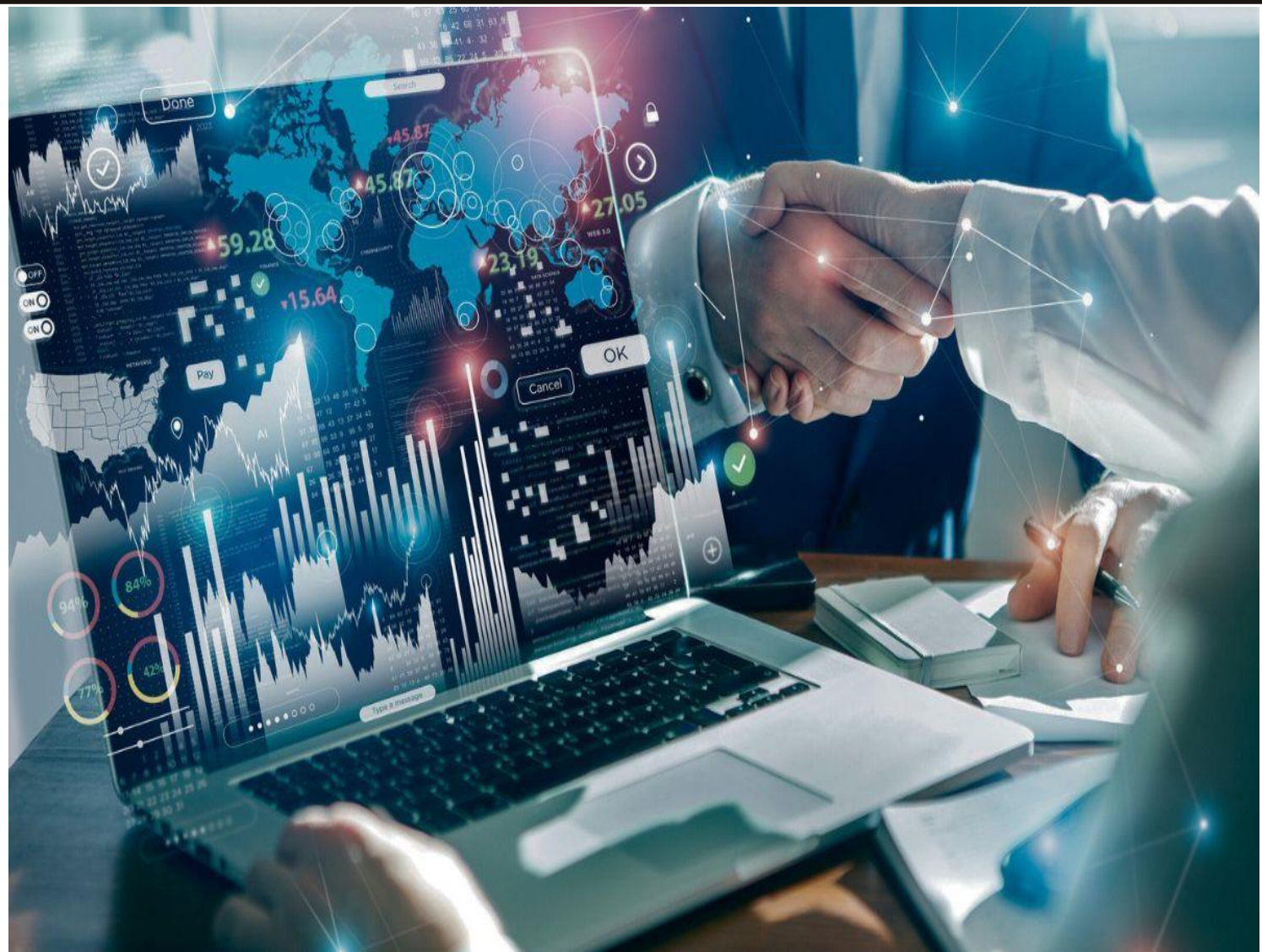


Modeling

- I split the data into:

1. Training – 80%
2. Testing – 20%

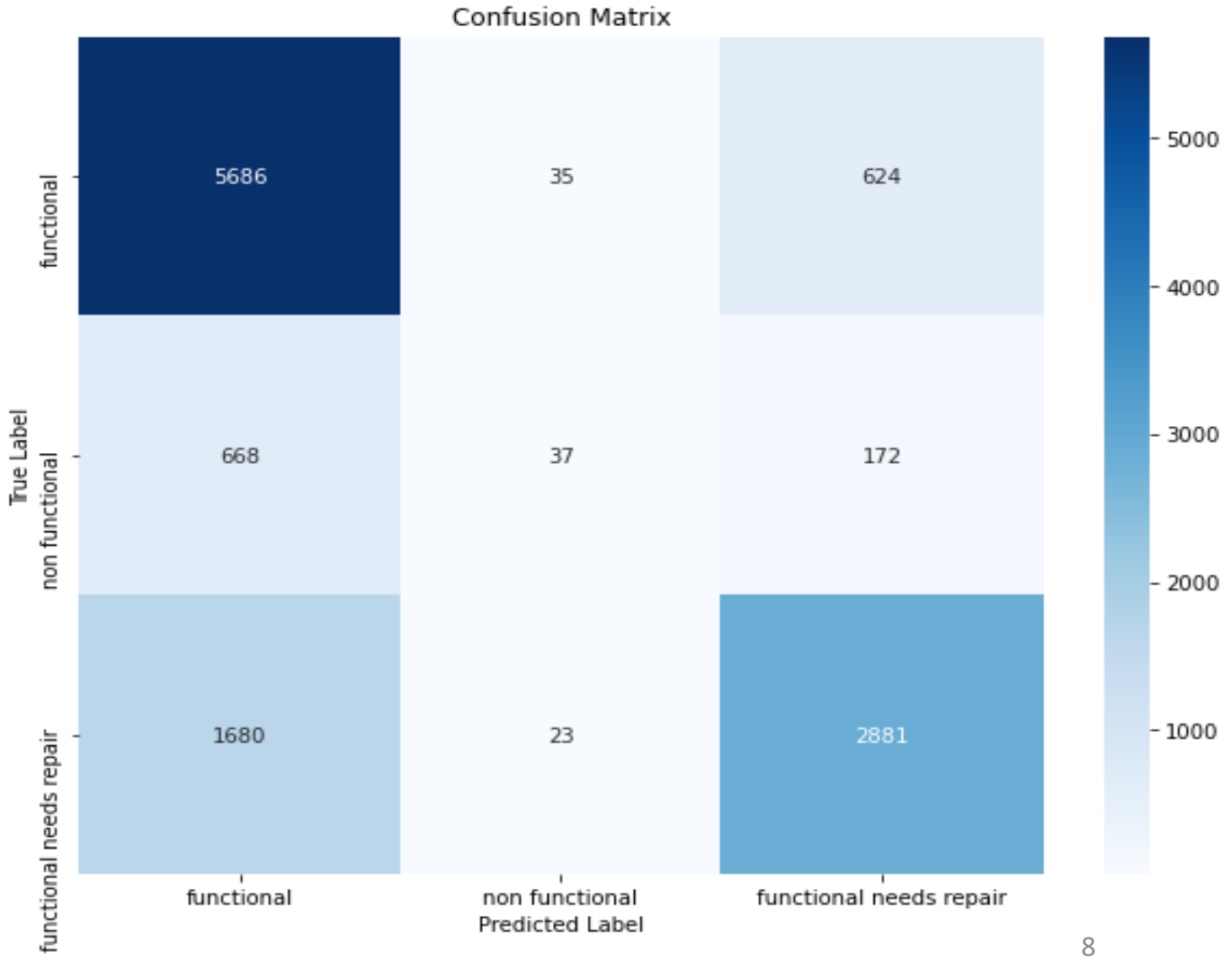
8 0% 2 0%



Results

Baseline model: **Logistic Regression**

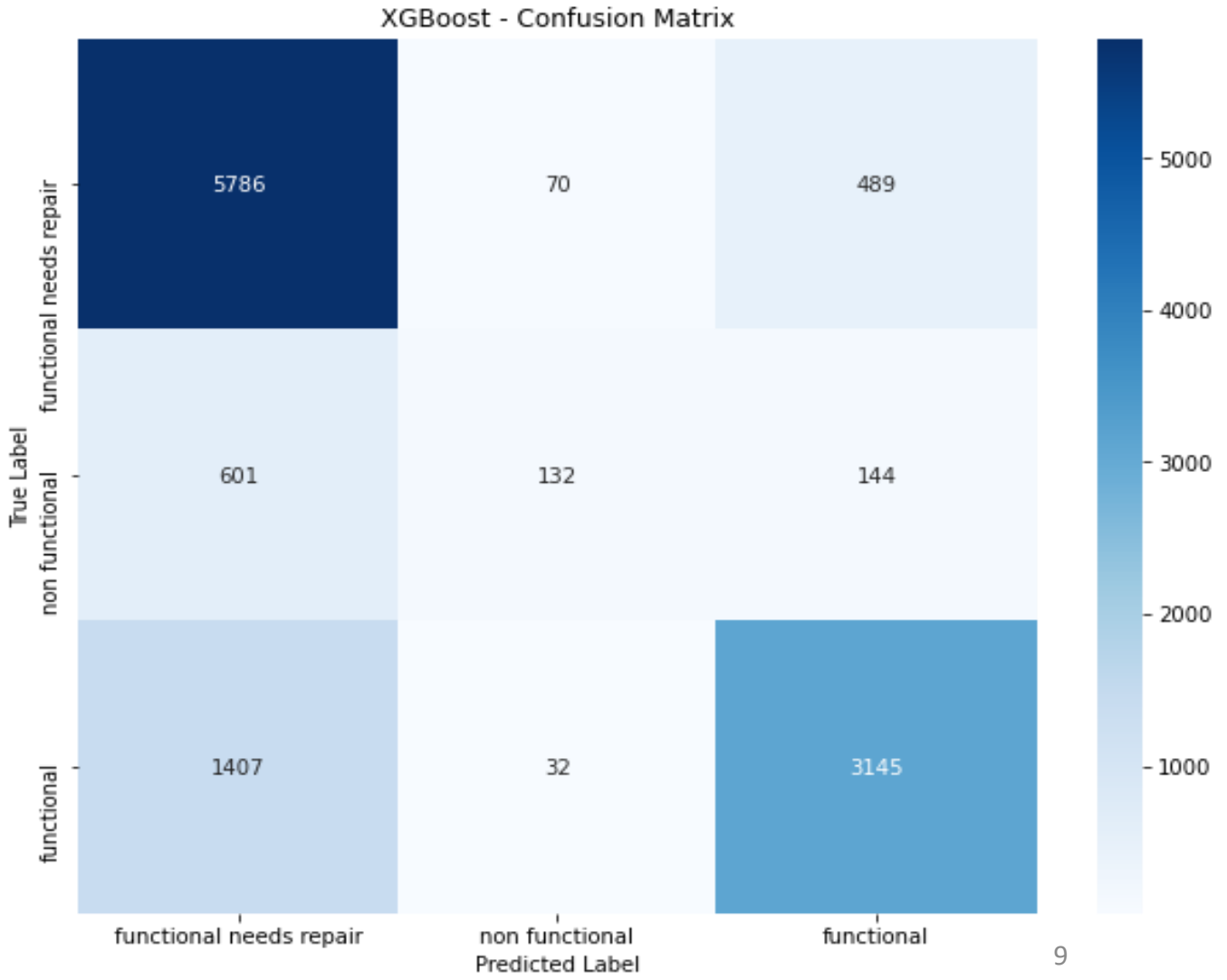
- Classifier was about 72.87%accurate on our testing data with an F1=score of about 70%.



Results(cont'd)

XGBoost Model

- The XGBoost Classifier was about 76.77% accurate on our testing data with an F1=score of about 74.96%.



Recommendations

- Use the XGBoost classifier to predict wells functionality.
- Gain deeper understanding of the factors that affect well functionality.

Next Steps

1. Improve class imbalance by using techniques such as oversampling or undersampling to balance the class distribution in the training data.

2. Use Hyperparameter tuning. Using Grid search or random search can help improve the model's ability to generalize and make accurate predictions on unseen data.

3. Incorporating Cross-validation. This will help identify whether the model's performance is consistent across different subsets of the data and reduce the risk of overfitting.

4. Gain a deeper understanding of the domain and the factors that influence the functionality of waterpoints.

Thank You!

Benson Kamau

Kamauben.kaguru@gmail.com

github.com/Kamaukaguru