CLASSIFICATION OF WATER WELLS IN TANZANIA

Rebecca sharon Kulundu

Outline

1. Introduction
2. Business Understanding
3. Methodology
4. Results
5. Conclusions
6. Recommendations

Introduction

Providing clean water to the population of over 57 million in Tanzania is a critical task due to the dire condition of some established water points. To address this challenge, **a classifier will be built to predict the condition of water wells** using factors such as the type of pump and installation date. The tool will aid NGOs in locating wells that require repair and the Tanzanian government in identifying patterns in non-functional wells, which can inform the construction of new wells.

Business Understanding

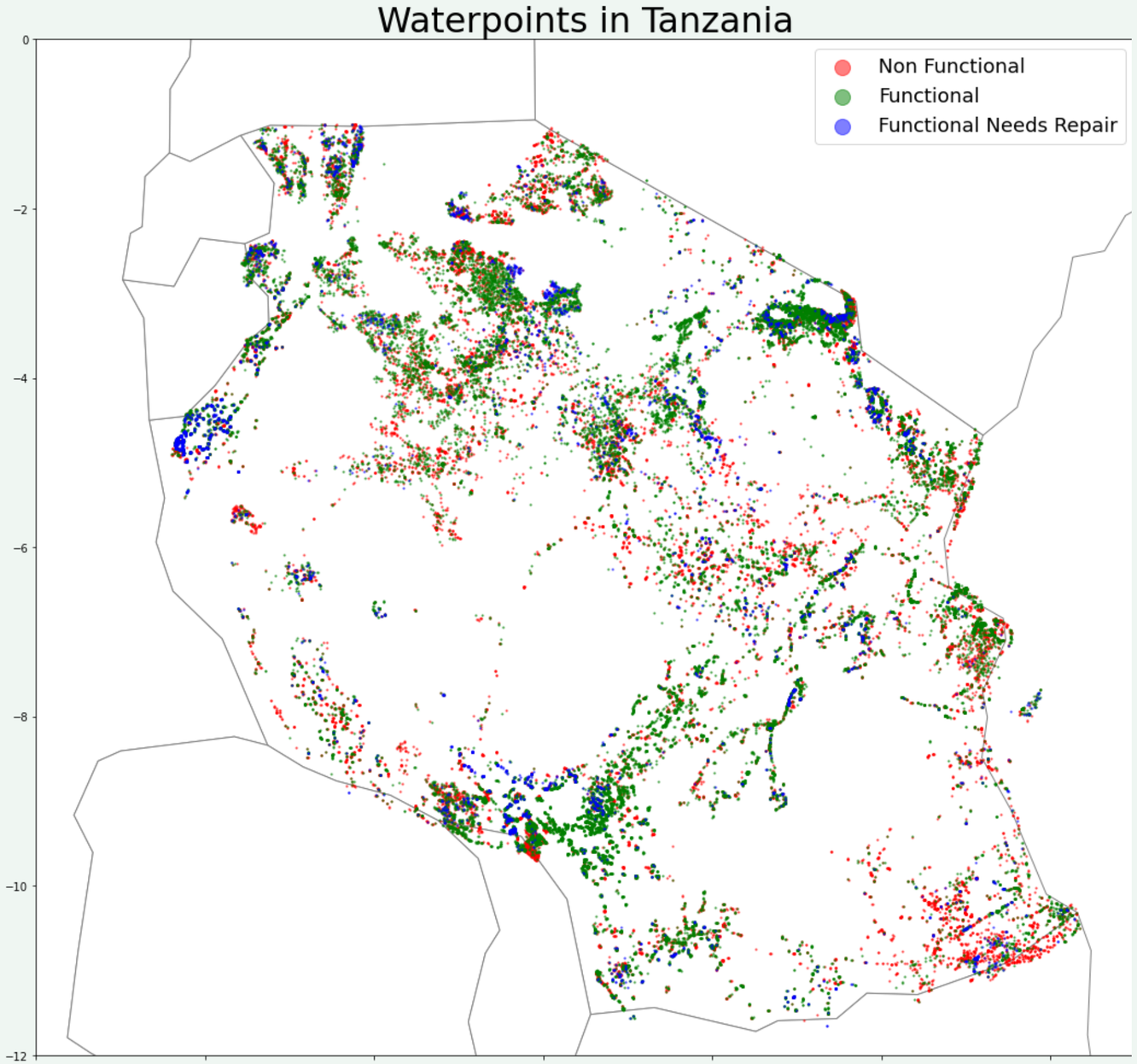
Given that the Tanzanian Ministry of Water has limited resources, how can available resources be used to efficiently maintain and expand the water system to provide clean water to the current population of 60 million people?
Specifically
If we can predict the status of a waterpoint with a high enough accuracy:
1: The Ministry will be able to know the status of any waterpoint without having to make a site visit.

2: By saving resources eliminating useless site visits, those resources can be allocated elsewhere.

Methodology

1. Conduct an exploratory analysis of the dataset to understand the relationship between the data features and labels. After identifying a significant amount of missing data in the dataset, clean the data and impute the missing values. The missing data appeared to be a result of unavailable data during collection or insufficient collection methods.
2. The dataset is imbalanced, requiring oversampling to generate synthetic data and balance the classes.
3. Supervised learning classifiers were implemented, including Logistic Regression, Logistic Regression with SMOTE oversampling, Decision Tree with SMOTE oversampling, XGBoost with SMOTE oversampling, and Random Forest. The best performing classifier, Random Forest, underwent hyperparameter tuning.

Data Used


Waterpoints in Tanzania
Non Functional
Functional
Functional Needs Repair
0
−2
−4
−6
−8
−10
−12

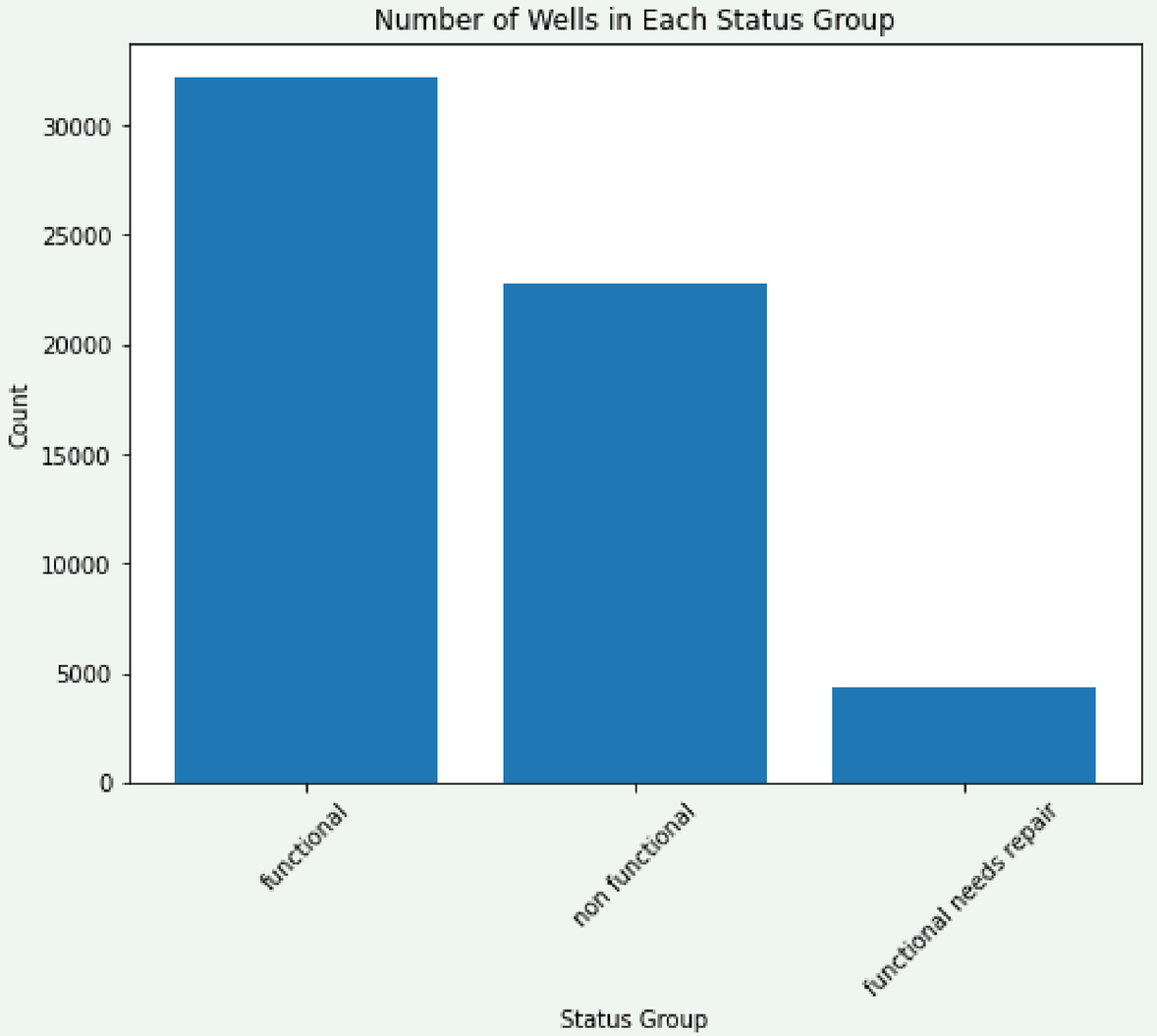
Description of data

- Data required to analyze the condition of water points in Tanzania was given by the Ministry of Water and Taafifa, which contained information such as the location, water quality and quantity, users, type of water point, and construction year, among others.
- Nearly 60,000 water points are distributed throughout the country and are displayed in the map above with their present status, color-coded as Non-Functional, Functional, or Functional Needs Repairs.

Data exploration

- From this visualization, one can tell that the data given is imbalanced with a majority of functional water points, which is beneficial for individuals but can make it difficult for a model to generate forecasts.

- To establish an even distribution of water point classes, an oversampling method was utilized, creating fresh data similar to existing data.

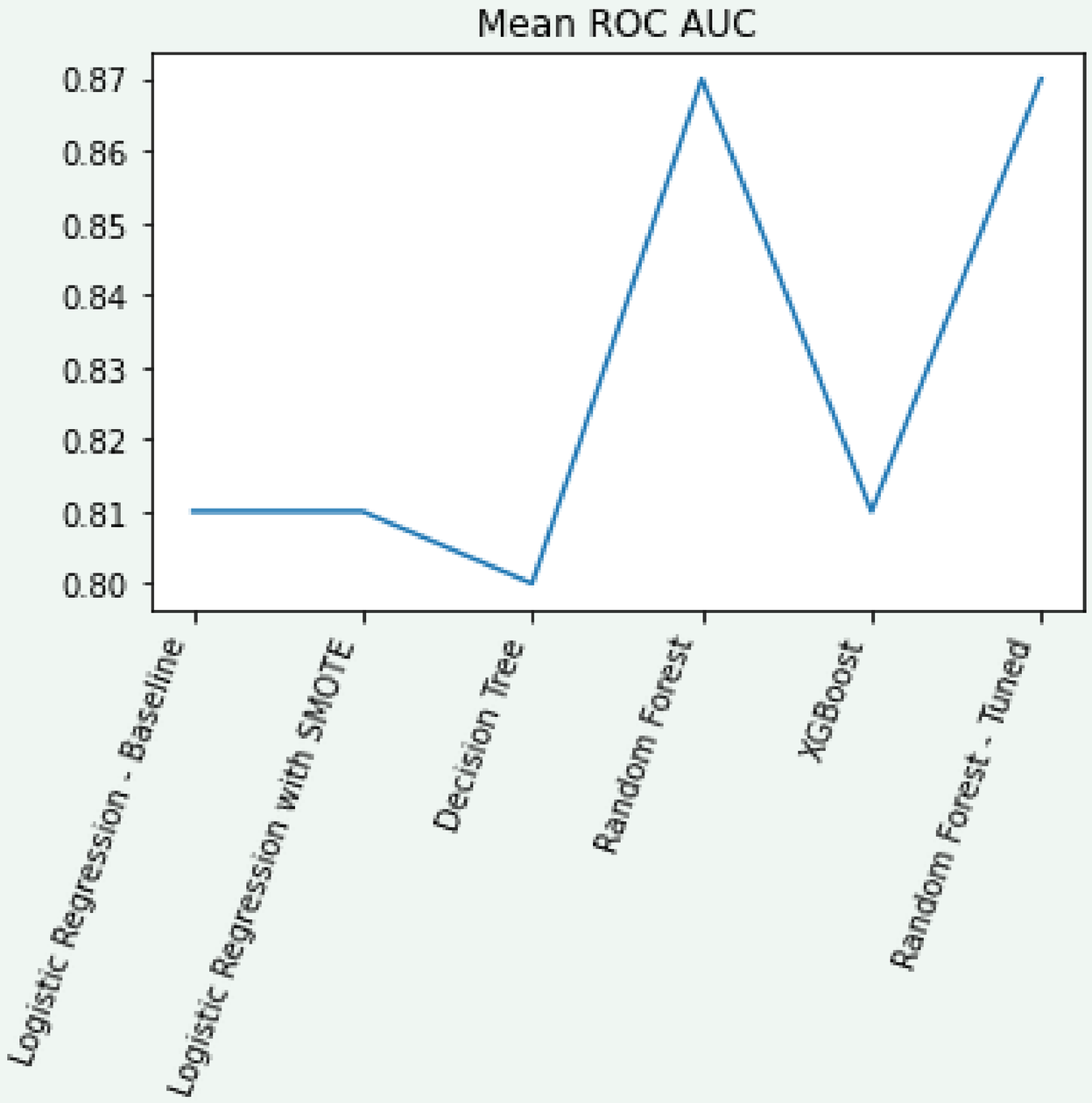


Supervised learning models used and the results.

Recall				
Model	**Functional**	**Functional Needs Repairs**	**Non Functional**	**Overall Accuracy**
Logistic Regression - Baseline	0.91	0.01	0.62	0.73
Logistic Regression w/ SMOTE	0.59	0.65	0.61	0.6
Decision Tree	0.74	0.46	0.7	0.71
Random Forest	0.75	0.60	0.72	0.77
XGBoost	0.69	0.57	0.61	0.65

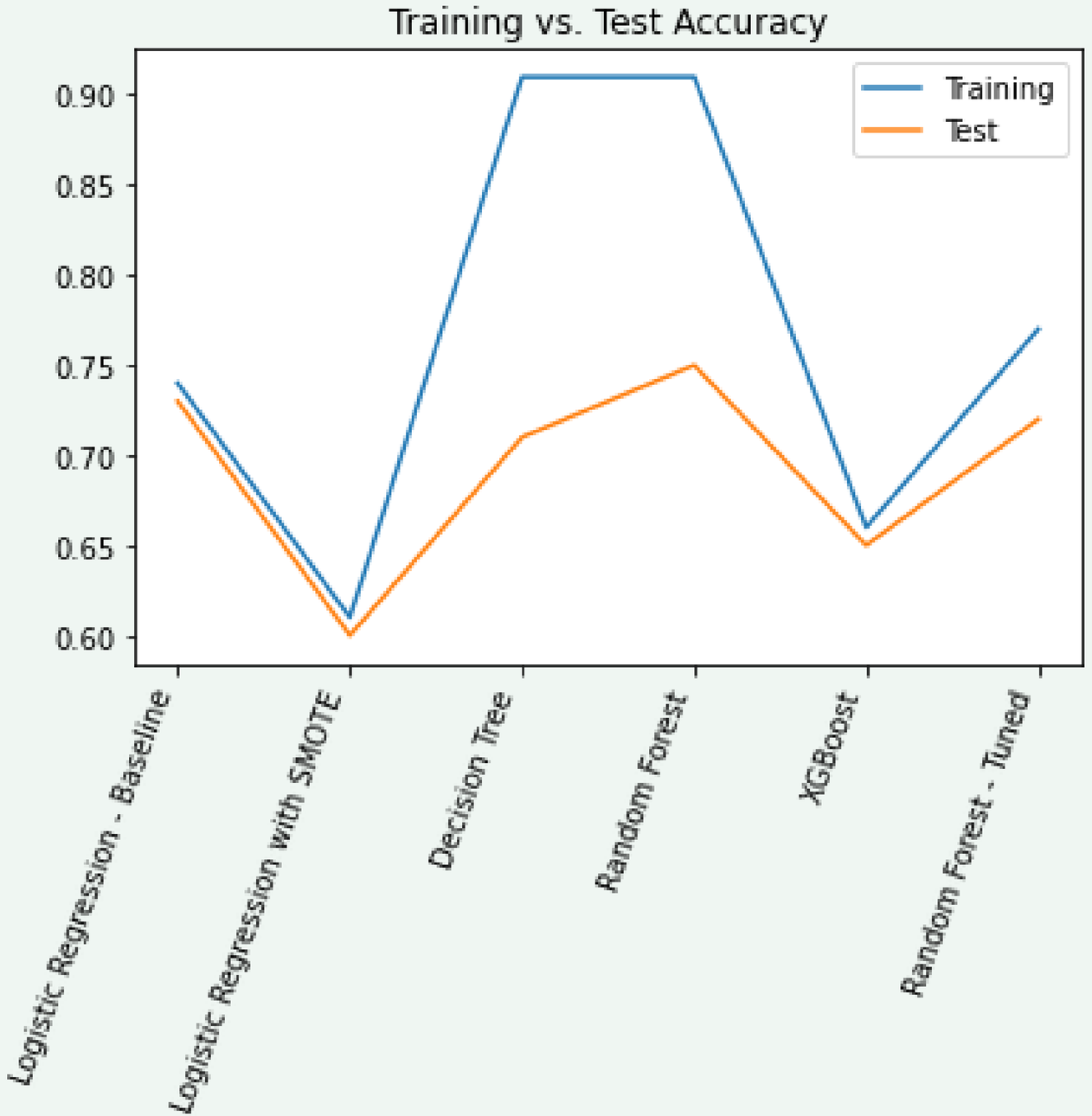
Results

- **Supervised learning classification** was employed to develop a model that could accurately predict classes/status of water points. The objective was to ensure that the model could correctly forecast the class of new water points beyond the ones on which the model was trained.
- **Five supervised learning classification models** were employed based on their overall accuracy and a balanced recall for each class. It was important that the models could correctly predict each class rather than just one or two of them.
- **The Random Forest Classifier model**, which is a collection of decision trees that vote on the class of each water point, was the most successful type of model. The model achieved 72% overall accuracy and recall rates of 75% and 72% for the two most extensive classes (Functional and Non-Functional, respectively) and 60% for the smallest class (Functional Needs Repairs).In this context, recall refers to the accuracy within a specific class.

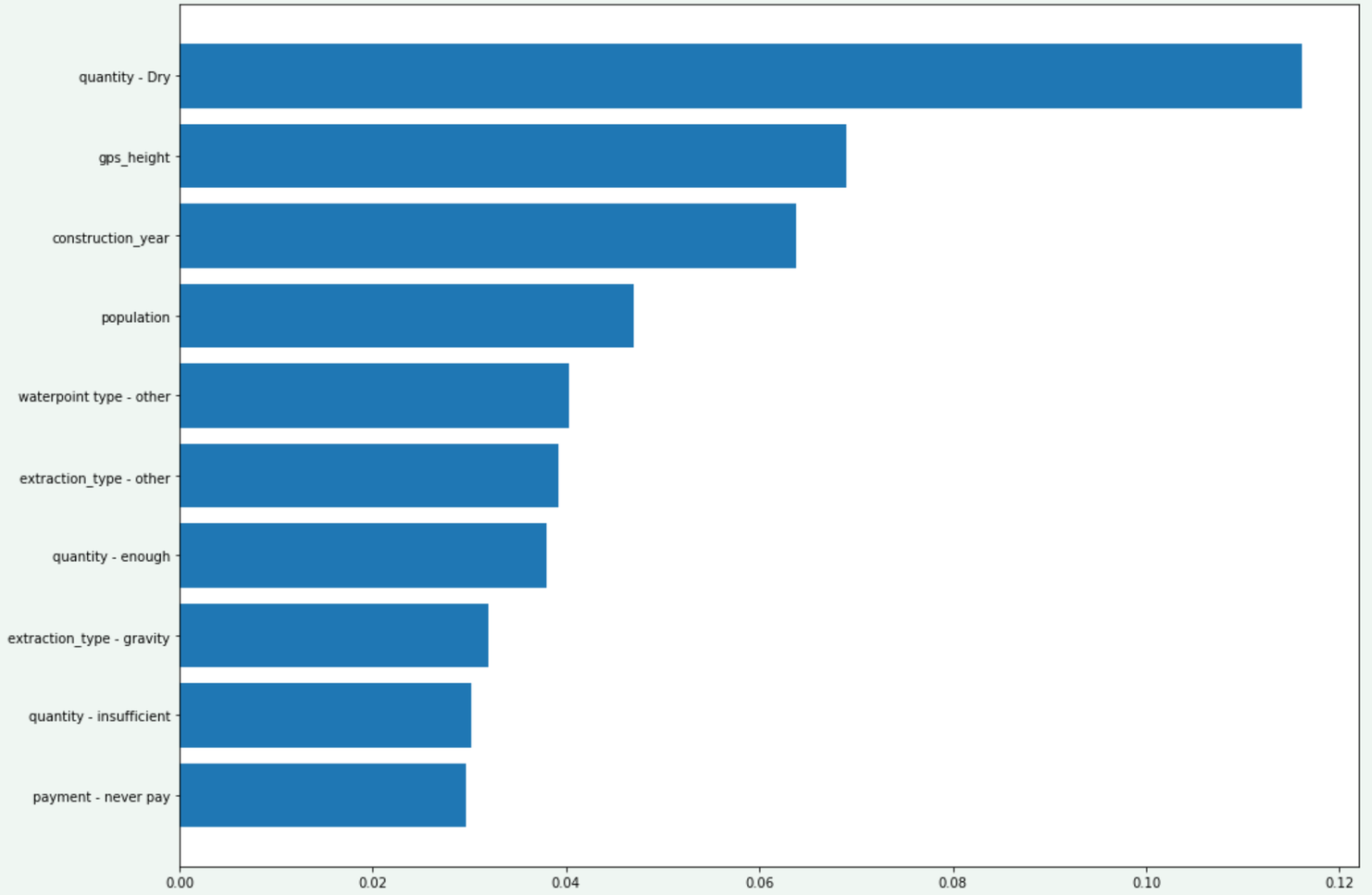


Mean AUC, a secondary performance metric in this case, but one that reflects,
the primary metrics of accuracy and balanced recall.

Training Vs. Test Accuracy for each
class for each classification model



- Feature Importances as determined by the best performing Random Forest Classifier
- The Random forest classifier found that the most important feature for classifying a waterpoint, was whether it was dry.
- Meaning regardless of the functionality of the waterpoint,there was no water.



Conclusion

- **By utilizing a Random Forest Classifier**, I obtained a 72% accuracy score and recall values of 0.75, 0.60, and 0.72 for the three categories, 'functional', 'functional needs repairs', and 'non-functional'. The minority class was the category with a recall of 0.60.
- **The Ministry of Water could benefit significantly from this capability to predict water point statuses with a 72% accuracy** since it would enable them to prioritize site visits.

Examine troublesome categorizations.

		Functional	Functional Needs Repairs	Non Functional
	Functional			
True	Functional Needs Repairs			
	Non Functional			
			Predicted	
		BAD	OKAY	GOOD

Examine troublesome categorizations.

- Our current accuracy stands at 72%, but there is room for improvement. Our objective is to accurately classify waterpoints into three categories - functional, functional needs repairs, and non-functional.
- The table uses green to indicate a correct classification, yellow for an incorrect but acceptable classification, and red for an incorrect and unacceptable classification.
- To improve our accuracy, we must focus on correcting misclassifications in the red areas. For instance, if a non-functional waterpoint is wrongly classified as functional needs repairs, it's acceptable because maintenance staff can still identify and fix the problem.
- However, if a non-functional waterpoint is wrongly classified as functional, it's unacceptable as it will remain non-functional and neglected.

Recommendations.

- **I recommend the Ministry of Water utilize the predictive model to develop a strategy for prioritizing waterpoint site visits.** The model recommends that waterpoints predicted to be 'functional needs repairs' and 'non functional' be given priority status, while waterpoints predicted to be 'functional' be inspected on a routine basis and to check for any misclassified waterpoints.

- **By creating a more effective maintenance program, the Ministry can reduce costs and allocate those savings** towards expanding the water infrastructure.

- Additionally, the accuracy of the model and the Ministry's improved maintenance program can be used to **showcase the government's progress and attract international aid.**

Limitations and next step to take.

- **To improve the model, maintenance records should be integrated to reflect repairs made to the waterpoints,** to prevent the same waterpoints from being classified as needing repairs year after year. Historical maintenance records can also be integrated into the model.
- **To investigate misclassifications in multiclass classification,** it's important to recognize that some misclassifications are more problematic than others.
- **Additional classifiers can be used to determine which non-functional waterpoints to prioritize for maintenance**, rather than simply assuming that all non-functional waterpoints should be visited,they should be split into visit or don't visit.

THANK YOU