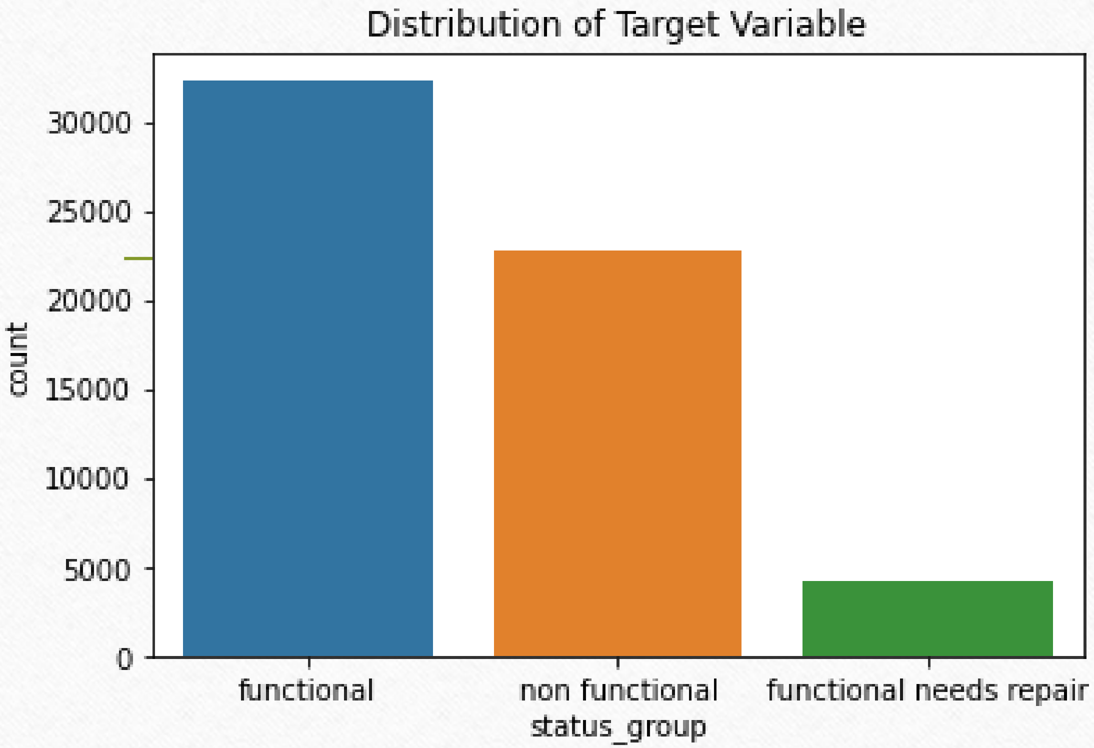
Moringa Phase 3 Project

By Cynthia Njambi Ngethe

Overview

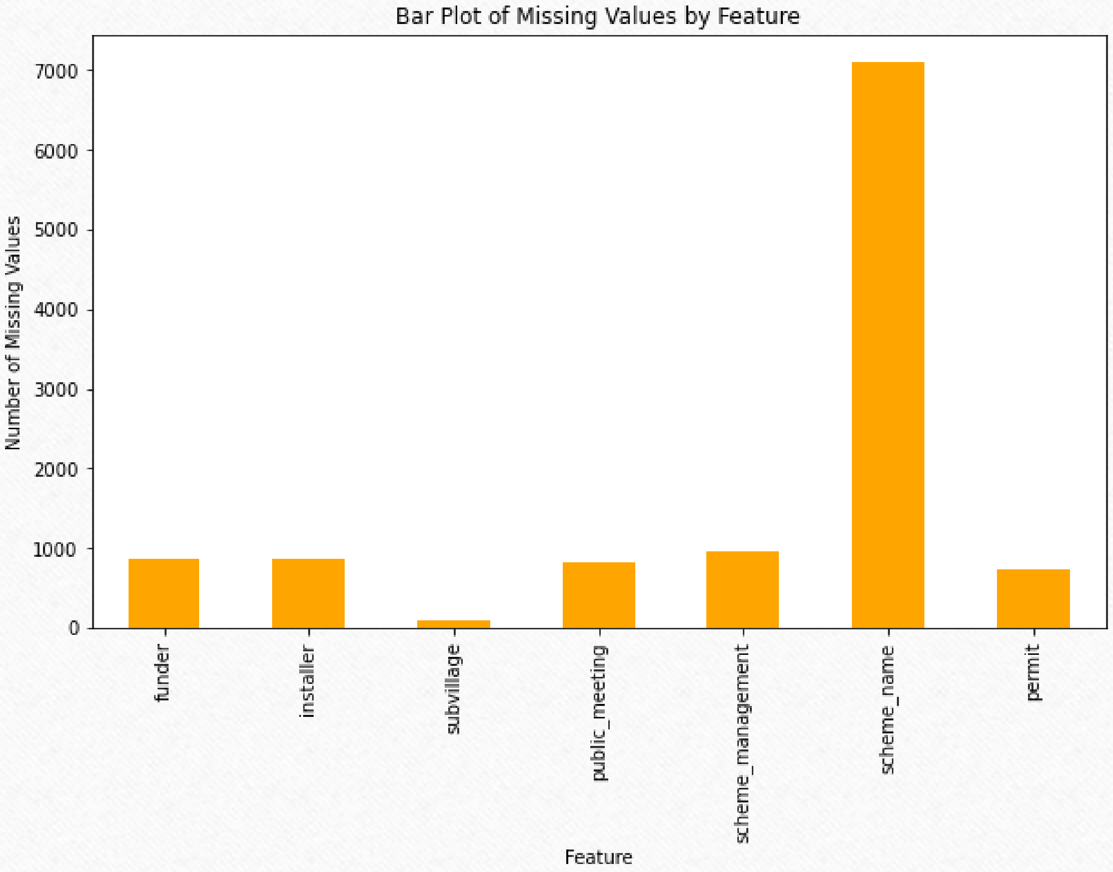
- This project focuses on predicting the functionality of water pumps in Tanzania using a large dataset provided by Taarifa and the Tanzanian Ministry of Water. The goal is to assist NGOs and the Tanzanian government in identifying critical factors that influence pump functionality, enabling them to make informed decisions on maintenance and new installations.

Business & Data Understanding



- I identified two stakeholders, namely: **Non-Governmental Organizations (NGOs):** Involved in providing support for well repairs across Tanzania and the **Government of Tanzania (Ministry of Water)** which seeks to understand patterns in non-functional wells to improve the planning and construction of new water points.
- The dataset comprises 59,400 water points, making it robust for training predictive models.
- I discovered that the pumps that are functional that need repair are: 7.3% (4,317) Since there are more pumps that are not working at all, it will be crucial to predict the need for repairs before complete failure so that maintenance can be done early.

Business & Data Understanding

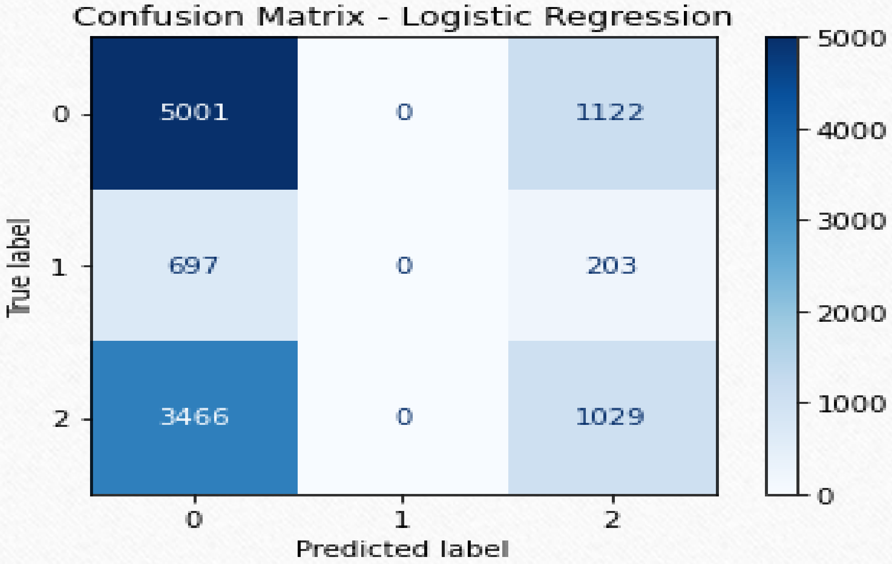


- Based on some funding insights of `amount_tsh` , I discovered that functional water pumps tend to have a higher `amount_tsh`. This means that better-funded pumps are more likely to be functional while lower funded pumps are more likely to be non-functional or needs repair.
- I realized `scheme_name` has the most significant proportion of missing values almost 47.4% of the data which needed to be addressed.
- The features `funder` and `installer` showed insights of could who is responsible for the functionality of the water points.
- The categorical features `public_meeting` and `permit` are values which might reflect the administration or governance aspects of water points.

Modeling & Evaluation

I chose three models: a Logistic Regression model, then a decision tree model and then finally random forest tree.

Logistic Regression Model

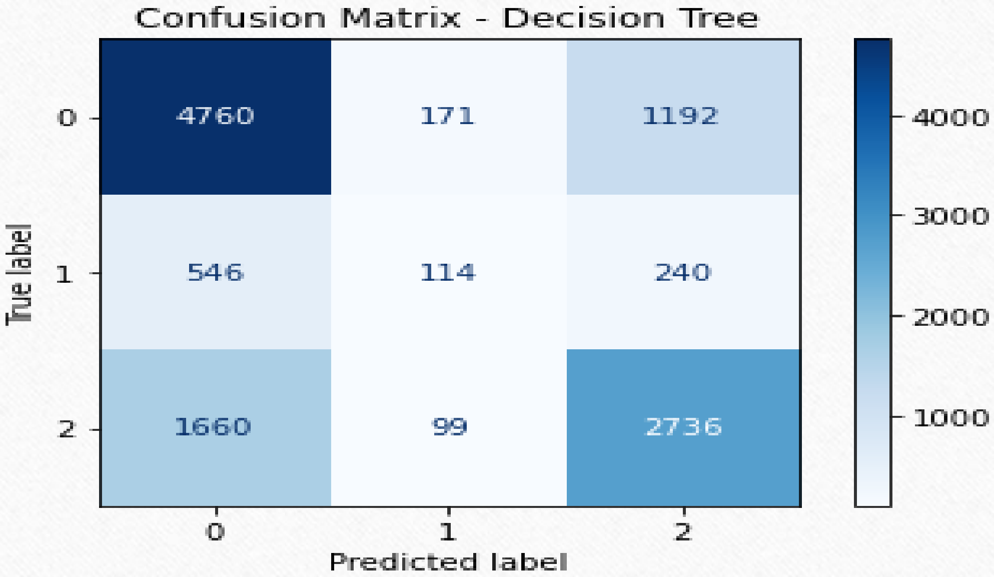


- In the logistic regression model I realized that it is evident that class 1(functional that needs repair) pumps is performing poorly on my model based on the precision, recall and F1-score being 0.
- This may be because that the class is being underrepresented thus the model is not learning enough about it.

Modeling & Evaluation

I chose three models: a Logistic Regression model, then a decision tree model and then finally random forest tree.

Decision Tree Model



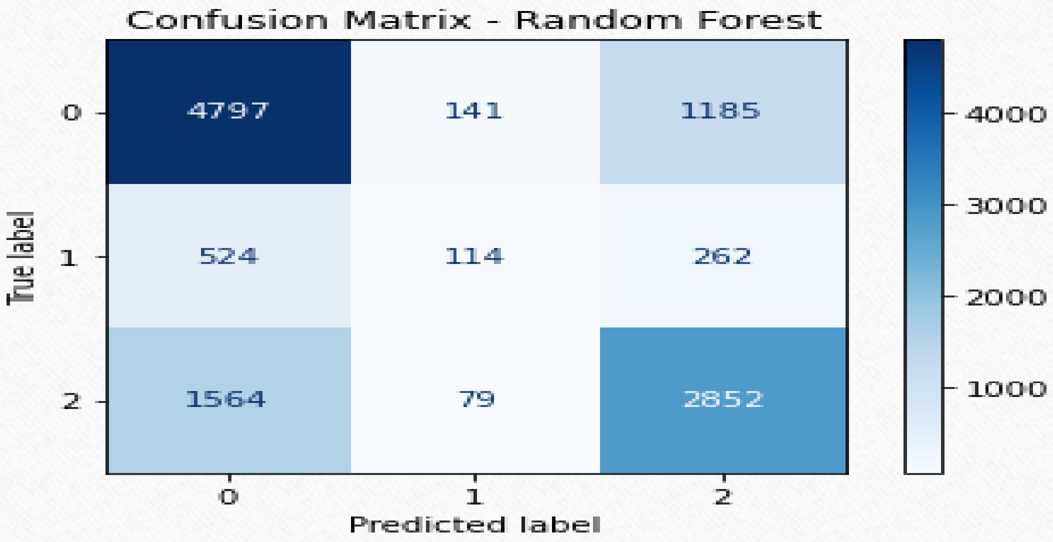
	0	1	2
0	4760	171	1192
1	546	114	240
2	1660	99	2736

- Decision Tree outperforms Logistic Regression in terms of accuracy.
- The model also performs slightly better with predicting class 1 (functional that needs repair) with a precision of 30% and a recall of 13% as compared to Logistic Regression both of which scored 0%

Modeling & Evaluation

I chose three models: a Logistic Regression model, then a decision tree model and then finally random forest tree.

Random Forest Model



Confusion Matrix - Random Forest

True label \ Predicted label	0	1	2
0	4797	141	1185
1	524	114	262
2	1564	79	2852

- Random Forest has the highest accuracy of (0.6740), which indicates better overall performance in predicting the target variable.
- Random Forest also has the best performance in precision, recall and F1-score.

Recommendations

- Focus on improving conditions in areas with large populations, ensuring proper installations by reliable contractors and scrutinizing the role of different funders in pump maintenance and quality
- Use the model's predictions to schedule maintenance activities for pumps that are at a higher risk of failure, optimizing resource allocation.
- Implement alert systems based on model predictions to notify maintenance teams about pumps needing attention.

Conclusions

- Predicting pump functionality accurately can reduce maintenance costs by prioritizing repairs and replacements for pumps likely to fail soon.
- Train maintenance staff to understand and use model predictions effectively as part of their workflows.
- The volume of water extracted can influence pump functionality. Higher volumes may lead to quicker wear and tear.