Predictive Analytics for operational wells in Tanzania

DrivenData. (2015). Pump it Up: Data Mining the Water Table. Retrieved [December 3 2024] from https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table.

Business Overview

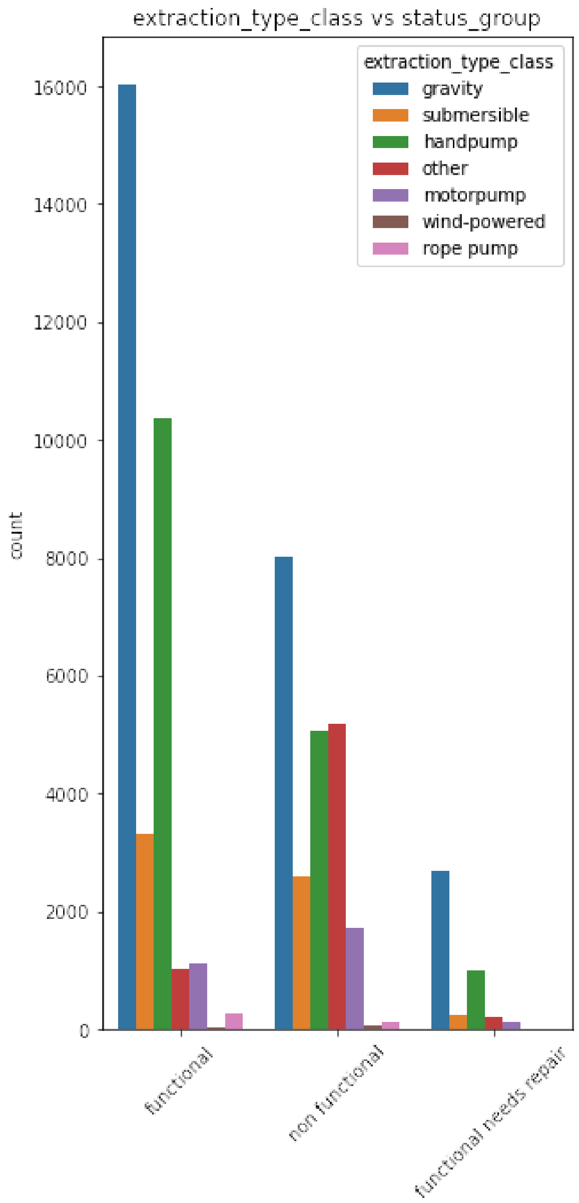
- Shortly after gaining independence, the Government of Tanzania initiated a policy aimed at providing free potable water to all rural inhabitants by 1991. This policy, consolidated in 1971, placed the responsibility of developing, operating, and maintaining water supply systems on the government, with no cost recovery. Many projects were funded by donors, particularly from Sweden, during the 1970s. This project has led to the building of many waterpoints to the day our dataset was collected. Some of these waterpoints are still functional while others are functional but in need of repair while others have lost their functionality

Objectives

- The primary objective of this project is to build a model that can predict if a waterpoint is either functional or nonfunctional given a set of independent variables. This information can be used by the Tanzanian Government or other stakeholders in identifying which waterpoints might need repair based on their characteristic

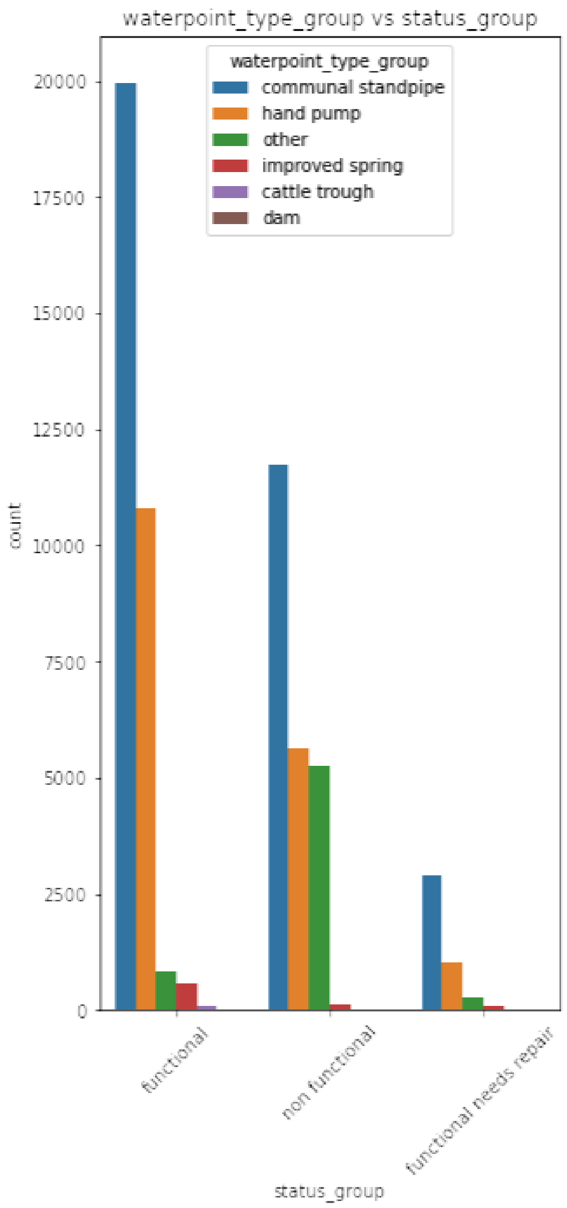
Extraction type vs status group

- The figure shows that water points that used gravity as a way to extract water had the most functioning wells. This indicates that pumps that used gravity



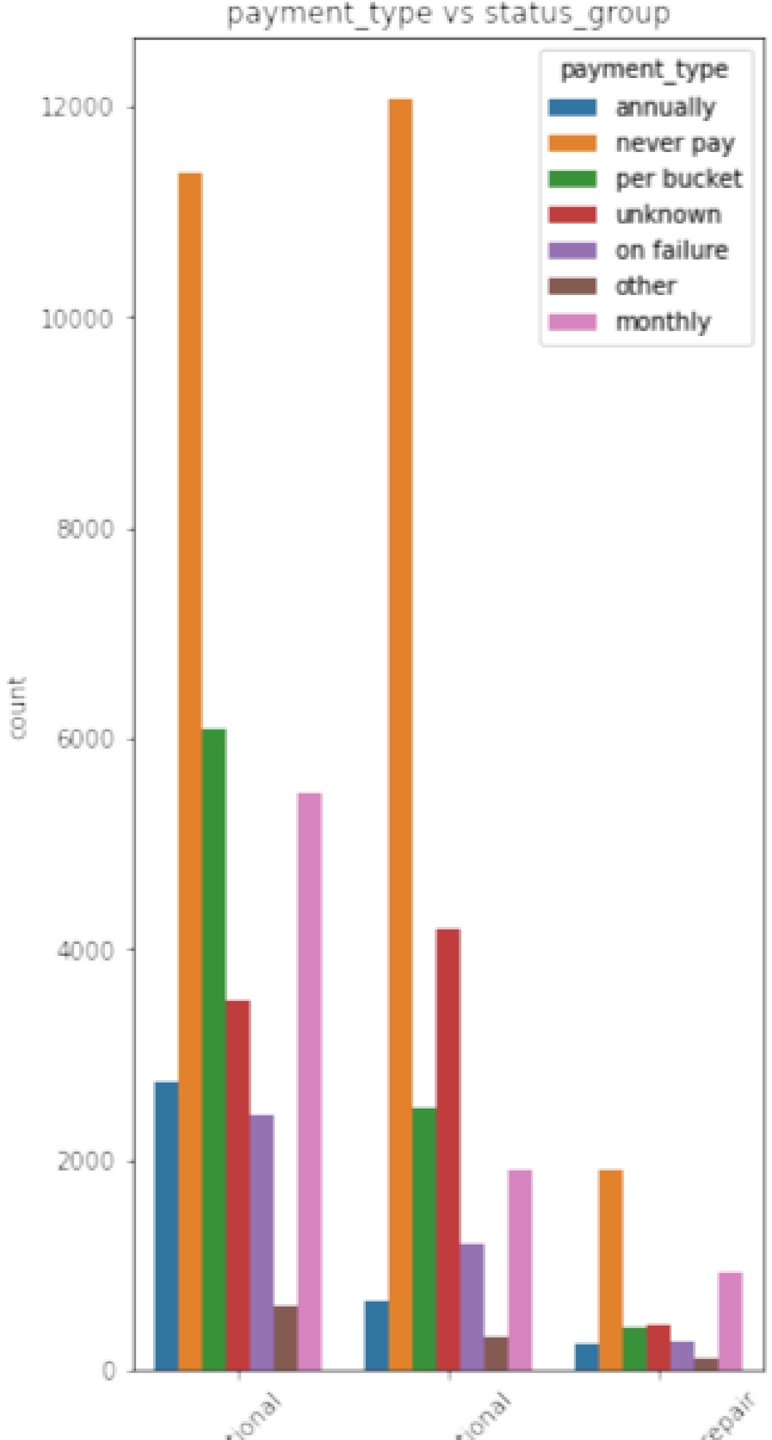
Water type group vs status group

- The above figure shows that simple pumps are the most common pumps found in Tanzania



Payment type vs status group

- The figure shows that most wells users don't have to pay to use the water point, but coincidentally most nonfunctional wells were used by users who did not have to pay. This means that those who managed the water points may have lacked funds to repair their wells leading to the wells they managed to cease from functioning



Model Performance

Logistic Regression with Standard Scaler and One Hot Encoding

- ***F1 Score:*** 0.8195
- ***ROC AUC Mean:***0.8474
- ***ROC AUC std:*** 0.0029
- ***Summary :*** This model showed good performance with a stable ROC AUC score showing consistent results across different folds

Logistic Regression with Min Max Scaler and One Hot Encoding

- ***F1 Score :*** 0.8195
- ***ROC AUC Mean:*** 0.8478
- ***ROC AUC std :*** 0.0027
- ***Summary :*** This model showed a similar performance with the previous with almost identical ROC AUC scores but was more consistent across different folds

Logistic Regression with robust scaler and target encoding

- ***F1 Score:*** 0.8077
- ***ROC AUC Mean:***0.8250
- ***ROC AUC std:*** 0.0043
- ***Summary :*** This model performed worse than the previous models with a wore F1 score and ROC mean. It showed that target encoding made the model performed worse than One Hot encoding. It also was more inconsistent in performance with a worse ROC AUC std

Logistic Regression with Robust Scaler and One Hot Encoding

- ***F1 Score:*** 0.8196
- ***ROC AUC Mean:*** 0.8479
- ***ROC AUC std:*** 0.0028
- ***Summary :*** This model showed similar performance with other linear regression models with one hot encoding further showing the strength of using one hot encoding for our categorical performance

Decision Tree with Robust Scaler and One Hot Encoding

- ***F1 Score:*** 0.8218
- ***ROC AUC Mean:***0.8179
- ***ROC AUC std:*** 0.0051
- ***Summary :*** This model showed the worst performance among our other models with the worse ROC AUC scores compared to other. The model also showed the least consistency on each fold

Model Choice

- From the above analysis of model performance, we have witnessed that logistic regression have performed better than the decision tree model.
- One Hot Encoding has proved to be the preferred encoding to deal with categorical columns. It improved the model scoring of model that used them.
- Choice of scaling showed no effect on our model metric scoring though robust scaling improved the runtime of our model.
- Therefore, the choice of model from my analysis will be the logistic regression with Robust Scaler and One Hot Encoding

Recommendations

1. Payment: Based on the findings it is recommended to priotise building wells that have a payment transaction. This ensures that scheme managers have money they can use to maintain wells in Tanzania

2. Improved data collection: The dataset had a lot errors that were noticed in data cleaning. This can affect the model accuracy so improvement in data collection might improve the accuracy of the model

3.Integration of external factors: Further information such as climate of the area would have been useful in our analysis

THANK YOU