PREDICTING WATERPOINT FUNCTIONALITY 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

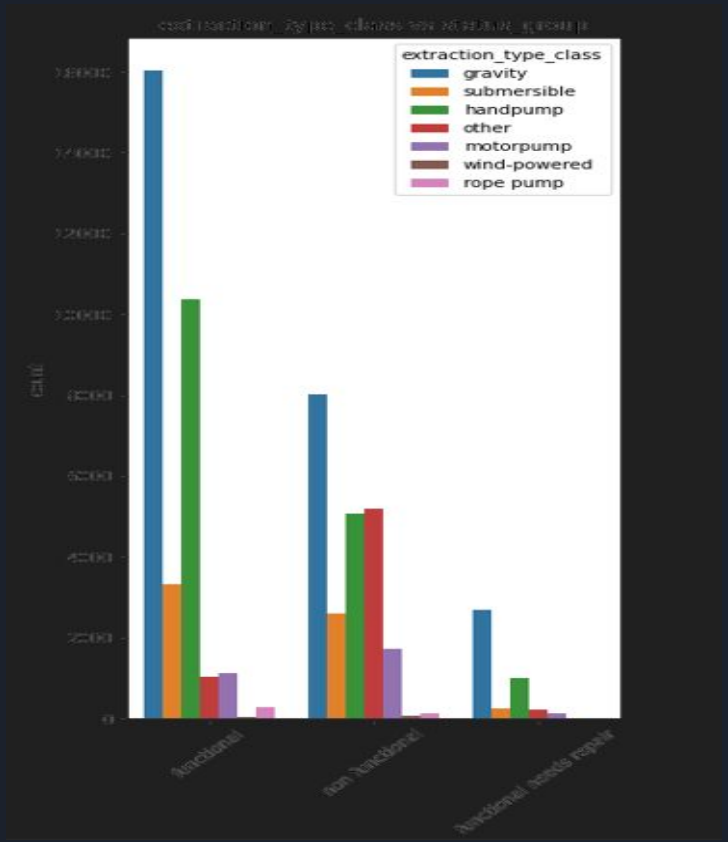
After gaining independence, the Tanzanian government introduced a policy to provide free potable water to all rural residents by 1991. Formalized in 1971, this policy made the government responsible for developing, operating, and maintaining water supply systems without implementing cost recovery measures. During the 1970s, many projects were funded by donors, particularly from Sweden, leading to the construction of numerous waterpoints. By the time the dataset was collected, some of these waterpoints remained fully operational, others required repairs, and some had ceased functioning altogether.

Objective

The main goal of this project is to develop a model capable of predicting whether a waterpoint is functional or non-functional based on a set of independent variables. This insight can assist the Tanzanian Government and other stakeholders in identifying waterpoints that may require repairs based on their specific characteristics.

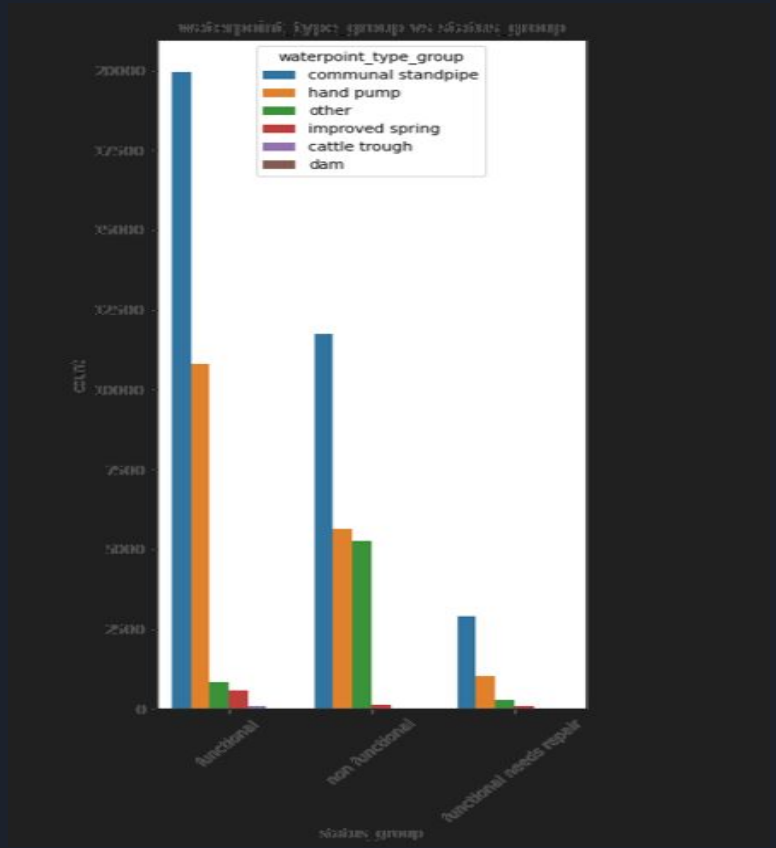
Water Type vs Status Group

This figure shows that simple pumps are most common pumps found in Tanzania



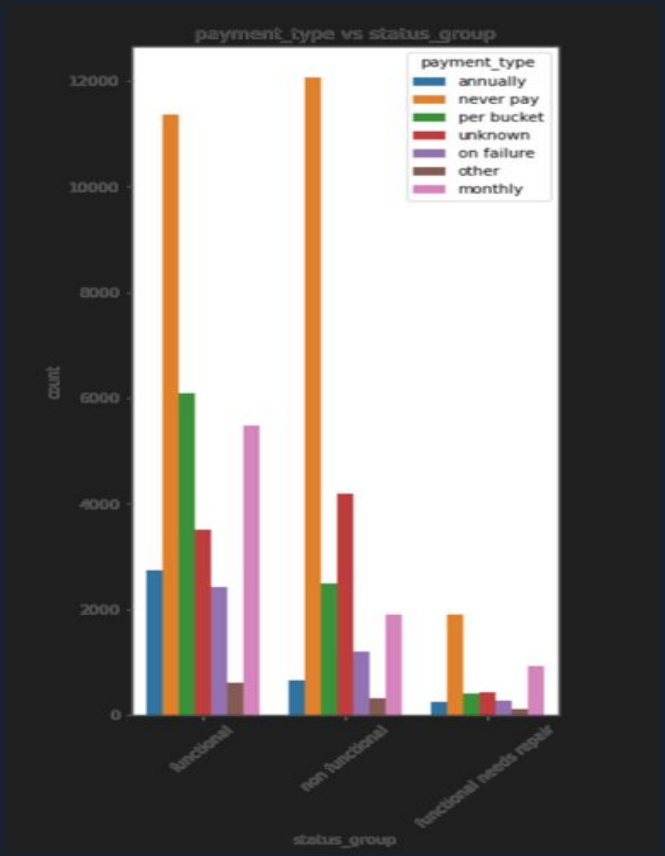
Extraction type vs Status group

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



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 Perfomance

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: 0.8077

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

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