TANZANIA WATER WELLS PREDICTION

AUTHOR : PAUL

NGATIA

TABLE OF CONTENTS

- This here shall contain a number of things, namely;

    1. Overview
    2. Business Understanding
    3. Data Understanding
    4. Modelling
    5. Evaluation
    6. Conclusion
    7. Recommendations

OVERVIEW

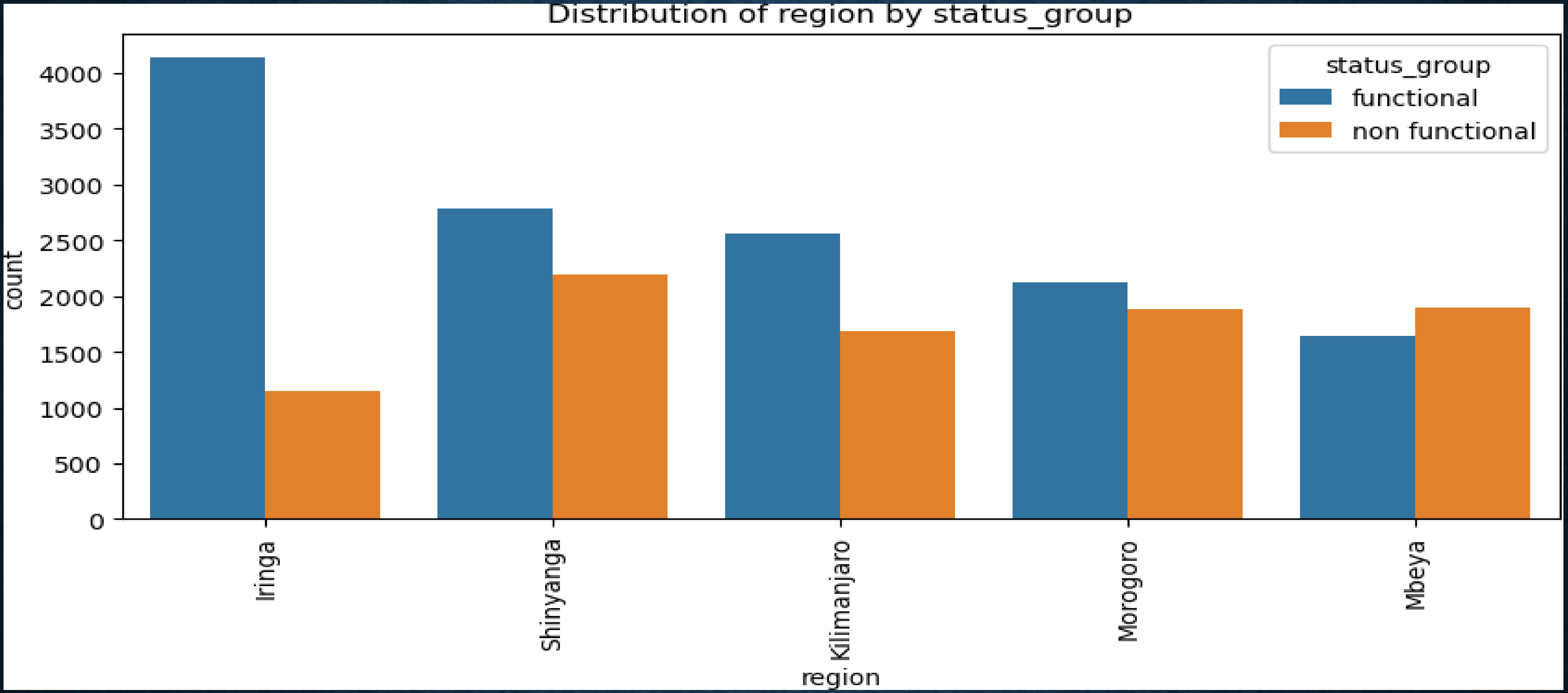
- Approximately 70% of the earth's surface is made up of water, one of the few necessities shared by all living things. Although there is some scarcity on land, regrettably, 96.5 percent of its coverage is made up of waters. Tanzania is among the nations that struggle with the lack of this essential resource for survival. Due to scarce resources for water extraction, this developing country, home to more than 57,000,000 people, struggles to meet the demand for clean drinking water. While there are currently some water pumps in the nation, some are sadly out of commission and others require maintenance.
- In the context of Tanzania, where this remains a persistent problem, predictive modeling emerges as a promising approach to enhance water resource management and optimize the drilling of new wells. By leveraging historical data, advanced statistical techniques, and machine learning algorithms, we aim to develop a robust predictive model that can accurately estimate the potential success and yield of water wells across different regions in Tanzania.

BUSINESS UNDERSTANDING

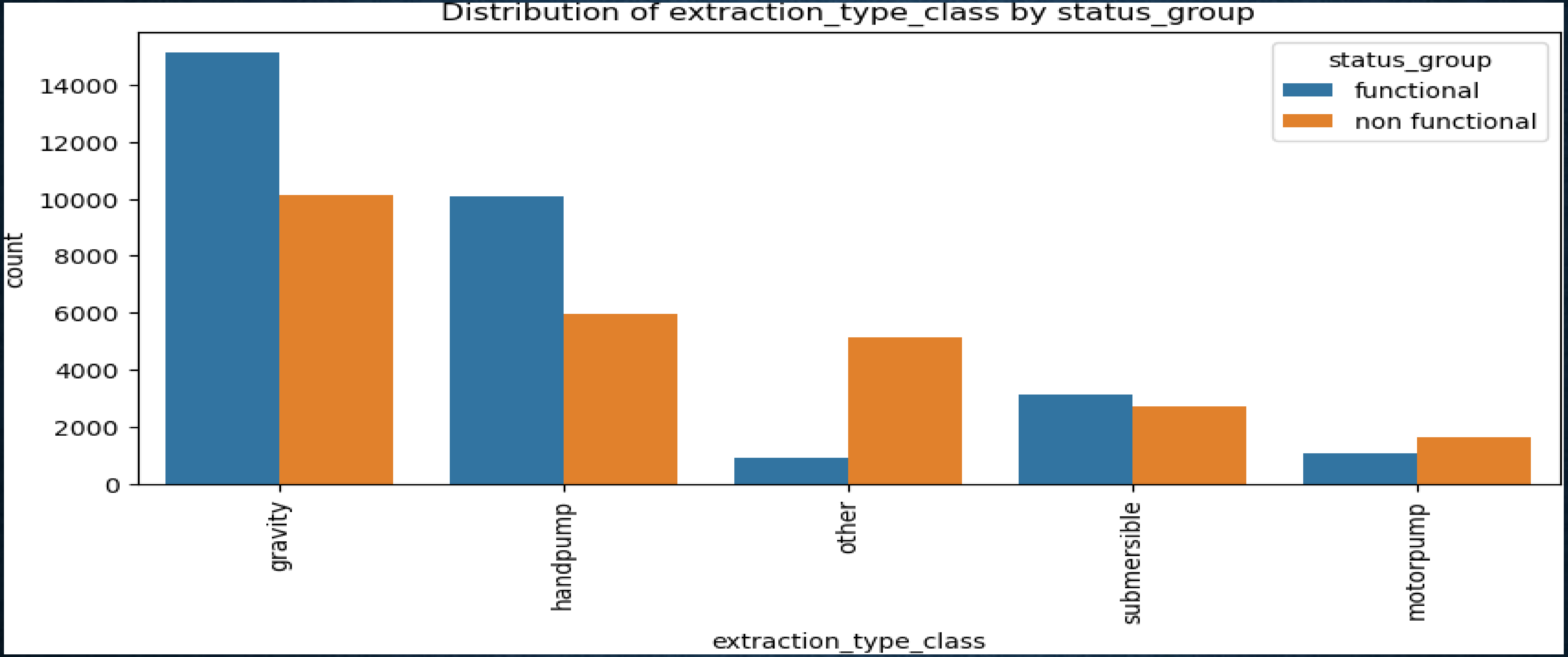
- Despite efforts to improve water access in Tanzania, significant challenges persist in ensuring sustainable and reliable access to clean water for all communities. The lack of accurate predictive models for water wells hampers efficient planning and resource allocation, resulting in suboptimal drilling locations, unreliable well yields, and inadequate maintenance strategies. As a result, communities continue to face water scarcity, health risks from contaminated water sources, and economic hardships.

DATA UNDERSTANDING

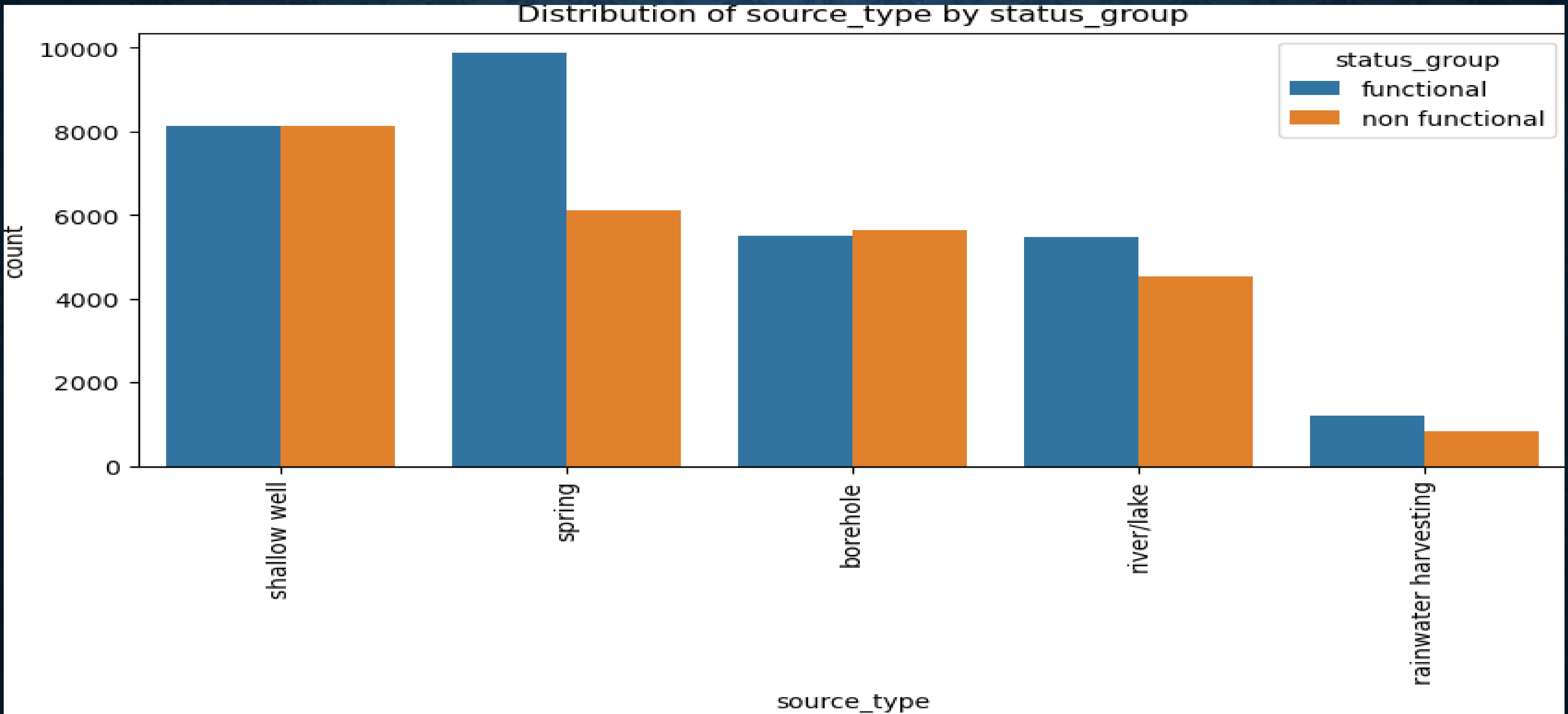
- In this project we shall use a dataset containing information about existing water wells in Tanzania sourced from an ongoing DrivenData competition.
- Four CSV files have been provided. One titled(Training set values) contains training set values with data on the independent features for the training set. The training set labels file (Training set labels) contains data on the dependent variable. The test set values (Test set values) contains values that will be used for prediction. A submission format (Submission format) has also been provided as this was a data science competition and the results of the analysis need to be in a specific format.
- The dataset contains 59,400 records and 40 columns. Of these columns, we identified 31 to be categorical, and 9 as numerical. We were able to further group the columns into the general features being captured.



From the distribution above, Iringa region has the most functional wells compared to the rest of the regions followed by Shinyanga, Kilimanjaro, Morogoro and Mbeya.



Based on the above distribution, gravity holds the highest indication for the type of extraction a waterpoint uses.



Springs are the most common source of functional wells.

MODELLING

Models Performed:

1. Dummy Classifier
2. Decision Tree Classifier
3. Random Forest Classifier
4. Support Vector Machine(SVM) Classifier
5. K-Nearest Neighbors(KNN)
6. KNN with Grid Search
7. Decision tree with Grid Search
8. Random Forest with Grid Search

- Our Final Model That performed best was the Random Forest Classifier wjich had an accuracy score of 0.78(78%)

EVALUATION

- Before fitting the data to the model, I scaled it using pipelines. Subsequently, to assest the effectiveness of our model, I additionally employed a confusion matrix. My algorithm was able to predict up to 78% of the functionality with a moderate level of success.
- I ran a total of 8 models, the baseline models and their hyperparameter tuned versions. The performance of the models varied with some overfitting on the training data and some not performing as well as expected. Notably, the performance of the some models reduced when I hyperparameter tuned the model. I attribute this to not having a wide enough search space for the best parameters. This can be improved upon in other iterations of the model
- The best performing model was Random Forest Classifier with an accuracy score of 78%.

CONCLUSION

- The accuracy of my Random Forest Classifier was 78%. While it is still a good predictive model, I would like to undertake further feature engineering to boost this accuracy score if I had more time. I achieved my objectives to be able to predict the functional wells and had a conclusive accuracy score.
- Even though the analysis may not provide a current solution to the current issue. I was limited by time among other things. More time would have allowed for the creation of better models. The machine used I used also put limitations to me i.e there are certain computationally expensive strategies such as hyperparameter tuning. Since this strategies have lengthy run periods, it is necessary to limit their application.

RECOMMENDATIONS

- The Lake Rukwa basin area, where there are considerably more non-functional wells than functional ones, may be worth considering for our stakeholders when they decide to build more wells in Tanzania.
- There are more non-functional wells than functioning ones in the Dodoma region; this situation has to be investigated.
- Over the course of time, wells with operating permission typically become more viable and useful than those without.
- Due to possible public misuse, unpaid wells frequently become inoperable; perhaps creating a reasonable payment plan will help stop this.
- Because wells without permissions also have a higher likelihood of becoming inoperable, our stakeholder must verify that they have permits to ensure that they are acceptable.

40
50
Thank You