Phase_3_Project

Tanzania Water wells

Business Understanding

The classifier built in this study has the potential to play a crucial role in improving the water situation in Tanzania. Clean water is a basic necessity for human survival and maintaining good health. Without access to clean water, people are susceptible to a range of waterborne diseases, which can have profound health implications. The role of the classifier is to help the NGO or the Government of Tanzania in their efforts to ensure that the population has access to clean water.

By predicting the condition of water wells, the NGO can prioritize its resources and focus its efforts on the wells that most need repair. The classifier will help the NGO make informed decisions about which wells to repair first, reducing the risk of waterborne diseases and improving access to clean water for the population. The Government of Tanzania, on the other hand, can use the insights from the classifier to make informed decisions about the construction of new wells. The classifier will provide the government with reliable and efficient data on the water situation in Tanzania, enabling them to take proactive measures to ensure that the population has access to clean water.

In conclusion, the classifier built in this study has the potential to be a game-changer in the efforts to provide clean water in Tanzania. By providing reliable and efficient data on the water well situation, the classifier will help the NGO and the Government of Tanzania prioritize their efforts and make informed decisions, ultimately leading to improved access to clean water and better health outcomes for the population.

Problem statement

The problem of providing clean water to the population of over 57,000,000 in Tanzania is a major concern, as many existing water points in the country are in need of repair or have failed altogether. To address this issue, a classifier will be built to predict the condition of a water well based on information such as the type of pump, date of installation and others. The target audience for this classifier could be an NGO focused on repairing wells or the Government of Tanzania looking to improve the construction of new wells.

Data Understanding

The data used in this analysis was obtained from DrivenData and was sourced from Taarifa and the Tanzanian Ministry of Water. It contains information on water wells in Tanzania and is divided into three files, including training set values, the training set labels, and test set values. The training data has 59,400 observations and 41 variables, providing extensive information on various aspects of the water pumps

Data Preparation

In order to prepare the data for the Tanzania water wells classifier, the following steps were taken:

Data Collection: The data was collected from various sources, including surveys and reports from the Tanzanian government and NGOs. The data was then cleaned and organized to ensure accuracy and consistency.

Data Preprocessing: The data was preprocessed to handle any missing values, outliers, or any other inconsistencies. This involved imputing missing values, scaling the features, and removing any irrelevant features that do not contribute to the model's performance.

Data Balancing: The data was imbalanced with a high proportion of wells that were functional, which would result in biased predictions if left unhandled. To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) was used to generate synthetic samples of the minority class (Repair) and balance the class distribution.

One-Hot Encoding: The categorical features in the dataset were converted into numerical values through one-hot encoding to ensure that the model can handle the categorical data.

By performing these data preparation steps, the data was made suitable for the classifier to train on and make accurate predictions.

External Data Analysis(EDA)

EDA, or Exploratory Data Analysis, is a crucial step in the data analysis process, especially for datasets like the Tanzania water wells data, where there may be missing values, outliers, and other anomalies that can impact the results of any modeling efforts. The objective of EDA is to gain a deeper understanding of the data, identify any potential issues, and explore relationships between different variables.

The first step in conducting EDA for Tanzania water wells data would be to load the data into a suitable tool, such as Python's Pandas library, and perform basic checks, such as checking the number of rows and columns and verifying the data types of the variables.

Next, it would be important to check for missing values and deal with them appropriately, either by removing rows with missing values or imputing the missing values with suitable values.

Then, we would want to visualize the distribution of the variables and identify any outliers that might exist. We could use histograms, box plots, and scatter plots to visualize the distribution of the variables.

We would also want to explore the relationships between the different variables and the target variable, which in this case is the "status_group" variable. We could use scatter plots and heat maps to identify any correlations between the variables.

Finally, we would want to perform feature engineering to create new variables or transform existing variables to help improve the performance of the machine learning model.

By performing EDA on the Tanzania water wells data, we can gain a deeper understanding of the data, identify potential issues, and prepare the data for modeling, which would ultimately help improve the performance of the machine learning models.

Modeling

Modeling is an important step in the analysis of the Tanzania water wells dataset. It involves using statistical algorithms to make predictions about the condition of the wells based on various features. The goal is to build a model that accurately predicts whether a well is functional or in need of repair.

To start, a baseline Logistic Regression model was tested on the data to provide a basic understanding of how the features affect the well's condition. After that, data imbalance was addressed using SMOTE and two more advanced models, Random Forest and XGBoost were tested on the data. The hyperparameters of these models were fine-tuned to optimize their performance.

Finally, the performance of each model was evaluated using various scores and metrics, such as accuracy, precision, recall, and F1 score. This process allowed for the comparison of each model and the determination of which one performs best on the Tanzania water wells dataset. With the best-performing model, the NGO or the Government of Tanzania can prioritize their resources and make informed decisions about the repair and construction of water wells in Tanzania.

Evaluation

After evaluating the results from the classification reports and different scores for each model, it appears that both XGBoost and Random Forest have quite comparable outcomes. The Random Forest Classifier does seem to perform slightly better in terms of predicting the 1 (Repair) class, however, when taking a holistic approach, XGBoost has just a slightly higher accuracy overall. Based on this analysis, it is fair to say that either one of these models could be considered an acceptable solution for our business requirements. However, further investigation and fine-tuning of these models might be necessary in order to determine which one would be the best fit. Additionally, it might also be worth considering other algorithms or models to see if they could perform better than both XGBoost and Random Forest, given the unique characteristics of the data and the specific goals and requirements of the business. After evaluating the various models and their performance, have determined that the **Random Forest Classifier** is the best fit for this dataset. Although XGBoost demonstrated slightly better scores overall, the Random Forest Classifier performed faster

Conclusion

The goal of building this classifier was to assist the NGO or the Government of Tanzania in providing clean water to the population by predicting the condition of water well. The importance of clean water cannot be overstated, as it is crucial for the health and well-being of the population in Tanzania.

To achieve this goal, the data provided was carefully preprocessed and a Random Forest Model was trained on the data. The model was able to achieve an accuracy of 78% in predicting the condition of each water pump, which is considered a relatively good result. The Random Forest Model was selected as the best solution, as it placed a priority on correctly identifying Non-Functional pumps over falsely classifying Functional pumps.

While this model may not be the most cost-effective solution, it aligns with the humanitarian approach that was desired in this project, and it effectively meets the project requirements. In conclusion, this classifier provides a reliable and efficient solution to the problem of ensuring clean water in Tanzania, and it will assist the NGO and the Government of Tanzania in making informed decisions about their resources and the construction of new wells. Nevertheless, further investigation and fine-tuning of this model or alternative models should be considered to find the best solution for this specific dataset, given its unique characteristics and business requirements.