Classification of water wells in Tanzania

Introduction

Access to clean water is a fundamental human need and a major challenge for many developing countries, including Tanzania. With a population of over 57 million, providing clean water to all citizens is a critical task. There are many water points already established in Tanzania, but some are in dire need of repair while others have completely failed.

In order to address this challenge, we aim to build a classifier to predict the condition of water wells based on information such as the type of pump used and the installation date. This tool will be instrumental in helping NGOs locate wells that need repair and the government of Tanzania to identify patterns in non-functional wells, which can inform the construction of new wells. This is a ternary classification problem, but can be designed to be binary. With this project, we hope to improve access to clean water for the people of Tanzania and help address this critical need.

Business Understanding

- Tanzania struggles to provide clean water to its population due to the failure or in need of repair of some of its existing water wells.There is a need to improve access to clean water in Tanzania. With a population of over 57 million, it is important to ensure that as many water wells as possible are functional and providing clean water to those who need it.

- The lack of access to clean water not only affects the health and well-being of the population, but also hinders the economic and social development of the country.. By building a classifier that can predict the condition of a water well, organizations such as NGOs and the government can more efficiently identify and repair non-functional wells and make informed decisions about the construction of new wells.The benefits of this project include improved water accessibility for the population of Tanzania, better management of existing wells, and data-driven decision-making for the government.

Main Objectives

- To build a classifier that can predict the condition of water wells based on information about the class the waterpoints belong to: functional, functional but need some repairs, and non-functional.

Specific Objectives

- Conducting a thorough exploratory data analysis of the dataset to understand the relationships between each feature and the labels.
- To develop a classifier that accurately predicts the condition of water wells, using the collected data.
- To evaluate the performance of the classifier and make improvements as necessary.
- To measure the impact of the classifier on the access to clean water in Tanzania, and to make improvements as necessary to ensure ongoing success.

Methodology

- Conduct an exploratory analysis of the dataset to understand the relationship between the data features and labels. After identifying a significant amount of missing data in the dataset, clean the data and impute the missing values. The missing data appeared to be a result of unavailable data during collection or insufficient collection methods.
- The dataset is imbalanced, requiring oversampling to generate synthetic data and balance the classes.
- Supervised learning classifiers were implemented, including Logistic Regression, Logistic Regression with SMOTE oversampling, Decision Tree with SMOTE oversampling, XGBoost with SMOTE oversampling, and Random Forest. The best performing classifier, Random Forest, underwent hyperparameter tuning.

Exploratory Data Analysis.

The dataset provided had almost 60,000 instances and 39 features, with most features being nominal categorical. During the examination of the features, it was discovered that some features were either duplicates or had a similar categorization. Furthermore, 7 features had NaN values, and 24 features had placeholder values such as 'unknown' or 0.

The data was cleaned by imputing the missing values, either by using the regional mean, median or mode, or by dropping the feature if it had a significant amount of missing data. Some categorical features required consistent capitalization. During the data cleaning process, feature selection was carried out, and the following categories were discovered: Useless, Missing Too Much Data, Redundant, Potentially Relevant, and Potentially Relevant but with high cardinality.

The Potentially Relevant features showed no clear correlation to the classes, and it was found that a combination of features would be necessary for successful prediction of the classes.

Modelling.

The following classification models were trained in the listed order.

- Baseline - Logistic Regression
- Logistic Regression with Oversampling
- Decision Tree with Oversampling
- Random Forest with Oversampling
- XGBoost with Oversampling
- Random Forest with Oversampling and Tuning

A basic Logistic Regression Classifier was selected as a baseline model and various classifiers tuned. The goal was to improve the Recall and Accuracy of the models for a dataset with imbalanced classes. The majority classes were well predicted, but the minority class was poorly predicted.

The sklearn StandardScaler and OneHotEncoder was used to scale and encode the continuous and categorical data, respectively. SMOTE oversampling was applied to the minority class for each classifier. The Random Forest Classifier had the most promising results, and I used RandomizedSearchCV to tune its hyperparameters.

The tuned model increased the Recall of the minority class, decreased the Recall of the majority classes, and slightly decreased the overall accuracy to the baseline level.

Conclusion.

- By utilizing a Random Forest Classifier, I obtained a 72% accuracy score and recall values of 0.75, 0.60, and 0.72 for the three categories, 'functional', 'functional needs repairs', and 'non-functional'. The minority class was the category with a recall of 0.60.
- The Ministry of Water could benefit significantly from this capability to predict water point statuses with a 72% accuracy since it would enable them to prioritize site visits.

Recommendations

- I recommend the Ministry of Water utilize the predictive model to develop a strategy for prioritizing waterpoint site visits. The model recommends that waterpoints predicted to be 'functional needs repairs' and 'non functional' be given priority status, while waterpoints predicted to be 'functional' be inspected on a routine basis and to check for any misclassified waterpoints.
- By creating a more effective maintenance program, the Ministry can reduce costs and allocate those savings towards expanding the water infrastructure.
- Additionally, the accuracy of the model and the Ministry's improved maintenance program can be used to showcase the government's progress and attract international aid.

Limitations and next step to take.

- To improve the model, maintenance records should be integrated to reflect repairs made to the waterpoints, to prevent the same waterpoints from being classified as needing repairs year after year. Historical maintenance records can also be integrated into the model.
- To investigate misclassifications in multiclass classification, it's important to recognize that some misclassifications are more problematic than others. For example, misclassifying a 'Functional Needs Repair' waterpoint as 'non functional' is less problematic than misclassifying it as 'functional', since the former only delays repairs while the latter results in wasted resources. The model's misclassifications can be categorized as correct, inconsequential, or problematic, and instances of problematic misclassifications can be reviewed and used to retrain the model with additional data.
- Additional classifiers can be used to determine which non-functional waterpoints to prioritize for maintenance, rather than simply assuming that all non-functional waterpoints should be visited. For instance, the Random Forest Classifier showed that a dry waterpoint is a strong indicator of non-functionality, and further classifiers could be trained to determine which non-functional waterpoints are worth visiting for repairs. This can help optimize the maintenance operation and reduce unnecessary costs.