Tanzania Water Wells Machine Learning Model

BY

James Mwaura Njoroge

Business Understanding and Introduction

Introduction:

- Tanzania, a country with a population exceeding 57 million, faces significant challenges in providing clean water to its residents. While numerous water wells have been established, many of these wells are either in disrepair or have failed entirely. Ensuring the functionality of these water wells is crucial for public health, agriculture, and overall quality of life. Predictive analytics can play a pivotal role in identifying which wells are likely to fail, need repair, or are functioning well, thus enabling proactive maintenance and efficient resource allocation

Stakeholders and Usage:

Government of Tanzania:

- Objective: Improve water supply infrastructure and resource planning.
- Usage: By analyzing patterns in well failures, the government can develop more effective strategies for constructing new wells, maintaining existing ones, and optimizing resource allocation. This can lead to better-informed decisions on where to invest in infrastructure improvements and preventative maintenance.

Non-Governmental Organizations (NGOs):

- Objective: Enhance the efficiency and impact of water-related aid programs.
- Usage: NGOs can use predictive models to prioritize wells that need urgent repairs or are at risk of failing. This enables them to deploy their resources more effectively, ensuring that their interventions have the maximum positive impact on communities reliant on these water sources.

Local Communities:

- Objective: Gain reliable access to clean water.
- Usage: By participating in data collection and reporting well conditions, local communities can contribute to the ongoing monitoring and maintenance efforts. This collaboration can help ensure that issues are addressed promptly, minimizing the time residents are without clean water.

Data Understanding

- Our data sources are :

Training Set Values:

- Description: Contains independent variables about each water well (e.g., type of pump, installation year, location).
- Usage: Used to train the predictive model.

Training Set Labels:

- Description: Contains the dependent variable (status_group) for each well, indicating its condition (functional, non-functional, needs repair).
- Usage: Provides target outcomes for training the model

Test Set Values:

- Description: Contains independent variables for wells needing predictions, similar to the training set values but without labels.
- Usage: The model predicts the condition of these wells.

Submission Format:

- Description: Template for submitting predictions, including well IDs and predicted status_group.
- Usage: Ensures predictions are submitted in the correct format for evaluation.

Conclusion on my final model- Random Forest

- The final model is a RandomForestClassifier with tuned hyperparameters trained on a dataset split into training and testing sets. The hyperparameters were tuned using an exhaustive search through various combinations to find the ones that maximize the balanced accuracy score on the test set. The best hyperparameters identified were as follows: 'n_estimators': 200, 'max_depth': 20, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'sqrt', 'bootstrap': False.
- The model achieved a decent level of performance, with a test accuracy of 78.70% and a balanced test accuracy of 71.61%. These metrics indicate that the model generalizes reasonably well to unseen data and is not overfitting excessively to the training set. The balanced accuracy score is particularly useful in scenarios where classes are imbalanced, as it takes into account the imbalance and provides a more reliable measure of overall model performance.
- The confusion matrix plot visualizes how well the model is predicting each class. It shows the number of true positives, true negatives, false positives, and false negatives for each class, allowing for a deeper understanding of the model's strengths and weaknesses in classification. Overall, the final model appears to be a solid choice for the given dataset and task.

Predictive Analysis of Tanzanian Water Well Conditions

- **Model Performance** The classifier built to predict the condition of water wells in Tanzania achieved a test accuracy of 78.70% and a balanced test accuracy of 71.61%. These metrics suggest that the model performs reasonably well in identifying the condition of water wells based on features
- **Important Features** The most important features identified by the model include the type of pump used, the installation year, and possibly other geographic or environmental factors. Understanding these key features can help stakeholders prioritize maintenance and repair efforts for water wells.
- **Useful Predictions** For an NGO focused on locating wells needing repair, the model's predictions can be highly valuable. By identifying non-functional or deteriorating wells accurately, the NGO can allocate resources more efficiently and effectively, ensuring that clean water access is maintained or restored where needed most.

Recommendations for Stakeholders:

- Modify Input Variables: Based on the model's insights, stakeholders could consider modifying certain input variables. For example, investing in newer pump technologies or improving maintenance schedules for wells installed in specific years could lead to better overall well conditions.
- Target Results: The model can help stakeholders set specific targets for well conditions. By analyzing patterns in non-functional wells, they can influence how new wells are built, ensuring they are more resilient and require less frequent repairs.
- Geographical Considerations: Considering geographic or environmental factors that influence well conditions can further enhance the model's predictive capabilities. For instance, areas with certain soil types or rainfall patterns may require different pump types or maintenance strategies.
- In conclusion, the predictive model offers valuable insights into the condition of Tanzanian water wells, aiding stakeholders in making informed decisions regarding maintenance, repair, and future well construction strategies.

Next Steps

- Validation and Deployment of Model: Validate the predictive model using additional datasets or real-time data to ensure its accuracy and reliability. Once validated, deploy the model for ongoing monitoring and prediction of water well conditions.
- Actionable Insights Implementation: Implement actionable insights derived from the EDA analysis, such as prioritizing maintenance in high-population areas, improving water quality monitoring, and investing in pumping infrastructure. Collaborate with stakeholders and authorities to translate these insights into practical initiatives.
- Continuous Improvement: Continuously evaluate and improve the model based on feedback and new data. Incorporate feedback from field teams, stakeholders, and ongoing data collection to refine the model's predictive capabilities and enhance decision-making related to water well management