George-Chira /
ml_water_project

Code Issues Pull requests Actions Projects Security Insights Settings

0 stars 0 forks 1 watching 1 Branch 0 Tags Activity

Private repository

1 Branch 0 Tags Go to file t Go to file + Add file Code ···

George-Chira Presentation upload 0aa63ce · 1 minute ago

data	Remove .gitignore file	2 days ago
Project_Notebook_PDF_copy....	Notebook pdf uploaded	7 minutes ago
README.md	updated README	yesterday
final_predictions_smote.csv	finished predictions	yesterday
index.ipynb	finished project	yesterday
presentation.pdf.pdf	Presentation upload	1 minute ago

README

README.md

Overview

This project aims to build a machine learning model that accurately classifies the operational status of water wells in Tanzania. The classification includes three categories: `functional`, `non-functional`, and `functional but needs repair`. By predicting these statuses, the model assists NGOs and government agencies in prioritizing well repairs and resource allocation, ensuring consistent access to potable water. This initiative is crucial for meeting Sustainable Development Goal 6: Clean Water and Sanitation.

Business and Data Understanding

Stakeholder Audience

The primary stakeholders are NGOs and government agencies responsible for managing water resources in Tanzania. These organizations need actionable insights to direct their efforts toward repairing and maintaining water wells, rather than unnecessary new constructions. The predictive model helps in the strategic allocation of resources by identifying wells that are non-functional or require repair, optimizing maintenance schedules, and reducing operational costs.

Dataset

The dataset used for this project comes from DrivenData and consists of over 59,000 records of water wells across Tanzania. The dataset includes various features such as geographic coordinates, well management details, and construction attributes. These features are instrumental in identifying the factors that contribute to a well’s functionality, allowing for the development of an effective predictive model.

Data Descriptions:

- **amount_tsh**: Total static head (amount of water available).
- **date_recorded**: The date when the data was recorded.
- **funder**: Organization that funded the well.
- **gps_height**: Altitude of the well.
- **installer**: Organization that installed the well.
- **longitude/latitude**: Geographic coordinates of the well.
- **basin**: Geographic water basin.
- **region**: Region where the well is located.
- **population**: Population around the well.
- **public_meeting**: Whether a public meeting was held (True/False).
- **scheme_management**: Entity managing the water point.
- **permit**: Whether the water point is permitted (True/False).
- **construction_year**: Year the well was constructed.
- **extraction_type**: The extraction method used.
- **management**: How the water point is managed.
- **payment**: Type of payment required.
- **water_quality**: The quality of the water.
- **quantity**: Quantity of water available.
- **source**: The source of the water.
- **waterpoint_type**: Type of water point (e.g., communal standpipe, hand pump).

The full dataset can be accessed on the DrivenData website.

Modeling

Initial Baseline Model

A baseline logistic regression model was chosen as the initial approach due to its simplicity and interpretability in multiclass classification tasks. The baseline model serves as a benchmark to evaluate the impact of subsequent improvements, such as hyperparameter tuning and class imbalance handling.

Hyperparameter Tuning

GridSearchCV was used to optimize hyperparameters, specifically focusing on regularization strength, solver types, and penalty. Hyperparameter tuning aimed to balance bias and variance, improving the model's generalization to unseen data.

Addressing Class Imbalance with SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) was applied to address class imbalance, ensuring better recall and precision for minority classes, which is crucial for accurately identifying wells in need of repair.

Model Variants

- **Baseline Logistic Regression Model**: Provided a foundational benchmark.
- **Hyperparameter-Tuned Logistic Regression Model**: Explored optimized parameters for better accuracy.
- **SMOTE Logistic Regression Model**: Addressed class imbalance, aiming to improve recall for minority classes.
- **Decision Tree Model**: Explored a non-linear and interpretable alternative.

Evaluation

Cross-Validation

Cross-validation was used to evaluate the models’ performance, reducing the risk of overfitting and ensuring robustness. The mean cross-validation accuracy was the primary metric for comparison.

Final Model Selection

The SMOTE Logistic Regression model was selected for final predictions due to its ability to handle class imbalance and its alignment with the project’s goals of improving well maintenance and resource allocation.

Final Model Evaluation and Interpretation

The final model evaluation was performed using the following metrics:

- **Accuracy:** 73%
- **Precision:** 74%
- **Recall:** 73%
- **F1 Score:** 73%

These scores reflect the model's effectiveness in correctly classifying the functionality of wells across all three categories. The balanced precision and recall indicate robustness in handling both majority and minority classes, critical for the stakeholders’ goal of prioritizing well repairs and resource allocation.

Business Implications

The final model's predictions empower NGOs and government agencies to focus on repairing wellss needing repair and the non-functional wells, maximizing the lifespan of existing infrastructure and optimizing resource utilization. This aligns with the broader objective of improving water access and ensuring sustainable water resource management in Tanzania. However, the lower recall for non-functional wells indicates that some wells in need of urgent repair might be missed, suggesting an area for further improvement. Enhancing the model's ability to identify these wells would further align it with our ultimate goal: ensuring that every community in Tanzania has access to clean and reliable water sources.

Conclusion

- This project developed a model that predicts the functionality status of water wells in Tanzania with a high degree of accuracy, precision, and recall. The SMOTE Logistic Regression model, in particular, provides a balanced approach to handling class imbalance, which is critical for making informed decisions about well maintenance and resource allocation.
- Although the model performs well, there is room for improvement, particularly in identifying non-functional wells. Future work could explore additional data sources, advanced modeling techniques, or further tuning to enhance the model's performance and impact.

Running the Notebook

To run the notebook, follow these steps:

1. Clone the repository to your local machine.
2. Open a terminal and navigate to the cloned repository.
3. Install dependencies using `pip`.
4. Open the notebook from the Jupyter interface.
5. Run the cells in sequential order to execute the analysis.

Contributing

Contributions are welcome! To contribute, follow these steps:

1. Fork this repository.
2. Create a new branch.
3. Make your changes and commit them.
4. Push to the branch.
5. Create a pull request.

Navigating the Repository

The repository contains the following key files:

- **index.ipynb**: The Jupyter Notebook containing all data preprocessing, model training, and evaluation steps.
- **README.md**: This file, providing an overview and detailed explanation of the project.

- **final_predictions.csv**: The final output file containing predictions for the test set.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● **Jupyter Notebook** 100.0%