JoelKy-coder / Phase_Three_Project

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

Phase_Three_Project

Public

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

JoelKy-coder Update README.md bf0f68f · now

Data	Push info	yesterday
.gitignore	Initial commit	2 days ago
IMG-20180422-W...	Push info	yesterday
Number of Functi...	Push info	yesterday
Presentation.pdf	Add files via upload	11 minutes ago
README.md	Update README.md	now
log_reg.pkl	Push info	yesterday
notebook.pdf	Add files via upload	12 minutes ago
water_wells.ipynb	Push info	yesterday

Phase_Three

Readme

Activity

0 stars

1 watching

0 forks

Releases

No releases published

Create a new release

Packages

No packages published

Publish your first package

Predicting the Condition of Water Wells in Tanzania

README

Languages

- ● Jupyter Notebook 100.0%

Business Understanding

Problem Description

Tanzania, with a population of over 57 million, faces significant challenges in providing access to clean and reliable water sources. While numerous water points (wells, boreholes, and pumps) have been established across the country, many are non-functional, in need of repair, or have failed altogether. This situation exacerbates water scarcity, particularly in rural and underserved areas, impacting public health, economic productivity, and overall quality of life.

The Government of Tanzania, in collaboration with NGOs and development partners, is committed to improving water infrastructure. However, limited resources and the vast number of water points make it difficult to prioritize repairs and maintenance. A data-driven approach to predict the condition of water wells can help the government and NGOs identify non-functional or at-risk wells, allocate resources efficiently, and inform future water infrastructure planning.

OBJECTIVES

The research seeks to meet the following objectives:

1. Analyze the Impact of Age, Technology, and Investment on Water Point Failure
2. Assess the Impact of Socioeconomic and Geographical Factors
3. Develop a Predictive Model for Water Point Failure

Data Understanding

This research utilized data from DrivenData about waterpoints. The dataset was split into three CSV files:

1. Training set values
2. Training set labels
3. Test set values

The training and test datasets contained similar columns, while the training set labels dataset included one column, which was the focus of the study.

Dataset

The dataset consists of three CSV files:

- `Training set values.csv` : Contains features for training.
- `Training set labels.csv` : Contains the target variable (well status).
- `Test set values.csv` : Contains features for testing.

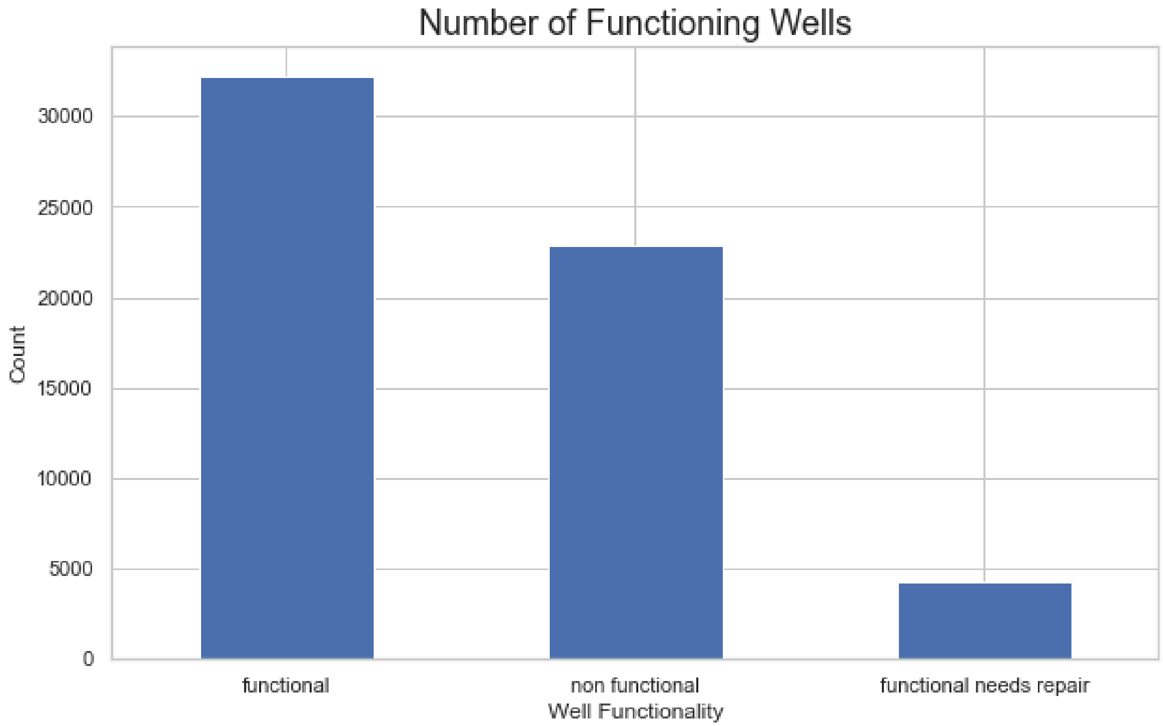
Exploratory Data Analysis (EDA)

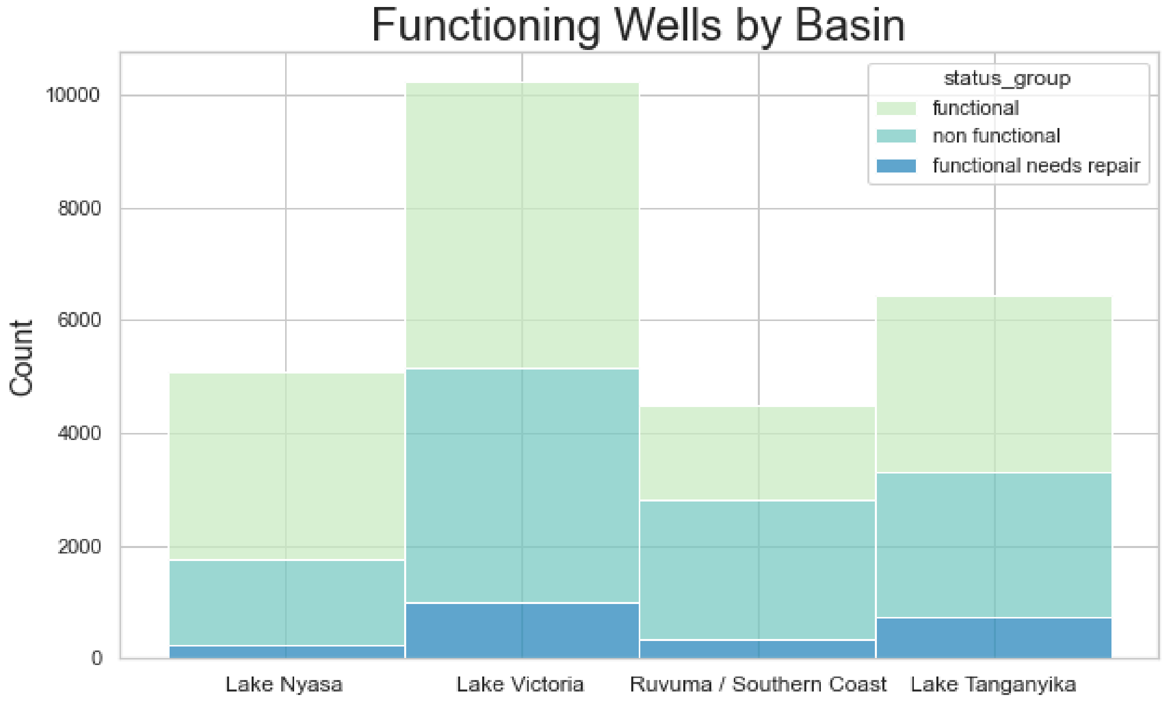
- Examined dataset size, missing data, and unique values.
- Visualized the distribution of well functionality.
- Analyzed correlations between numerical predictors.

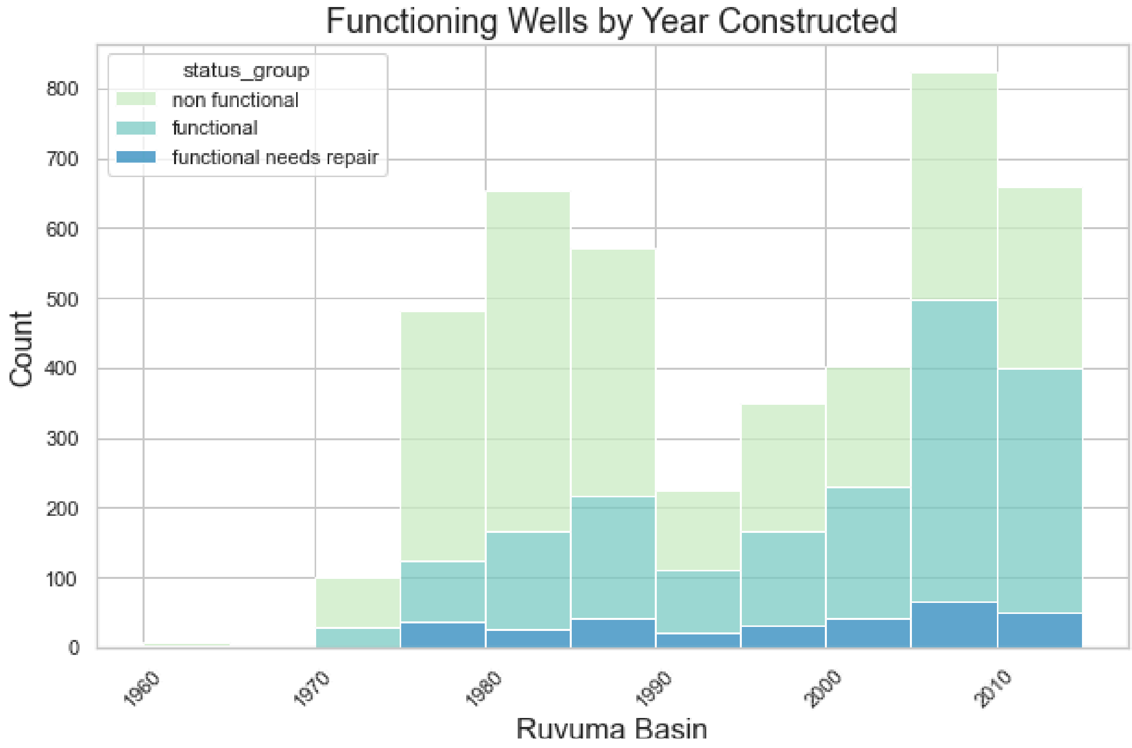
Model Development

- Preprocessed data by handling missing values and encoding categorical variables.
- Explored various classifiers: Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting.
- Used cross-validation and hyperparameter tuning to optimize model performance.

Data visualization




Functioning Wells by Basin
status_group
functional
non functional
functional needs repair
Count
0
2000
4000
6000
8000
10000
Lake Nyasa
Lake Victoria
Ruvuma / Southern Coast
Lake Tanganyika


Functioning Wells by Year Constructed
status_group
non functional
functional
functional needs repair
Count
0
100
200
300
400
500
600
700
800
1960
1970
1980
1990
2000
2010
Ruvuma Basin