Tanzania Water Wells Project

Author--Karanja Gakio

CONTENT

project overview

This project aims to significantly improve access to clean water in Tanzania an East African country with a population of 63.59 million, where only 57.1% currently have access to basic water supply.

BUSINESS PROBLEM

Karanja Gakio

1

Identify the patterns in functional and non-functional wells

2

Predict the functionality of water pumps based on the features provided.

3

Ascertain features that greatly affect water pump functionality

Data source

_There are 4 different data sets; training set, test set , target feature set and feature descriptions set . With the given datasets , the requirement is to build a predictive model and apply it to determine the status of the wells .

_The data sets was forked and scarpped from the web sites shown below ;

kaggle

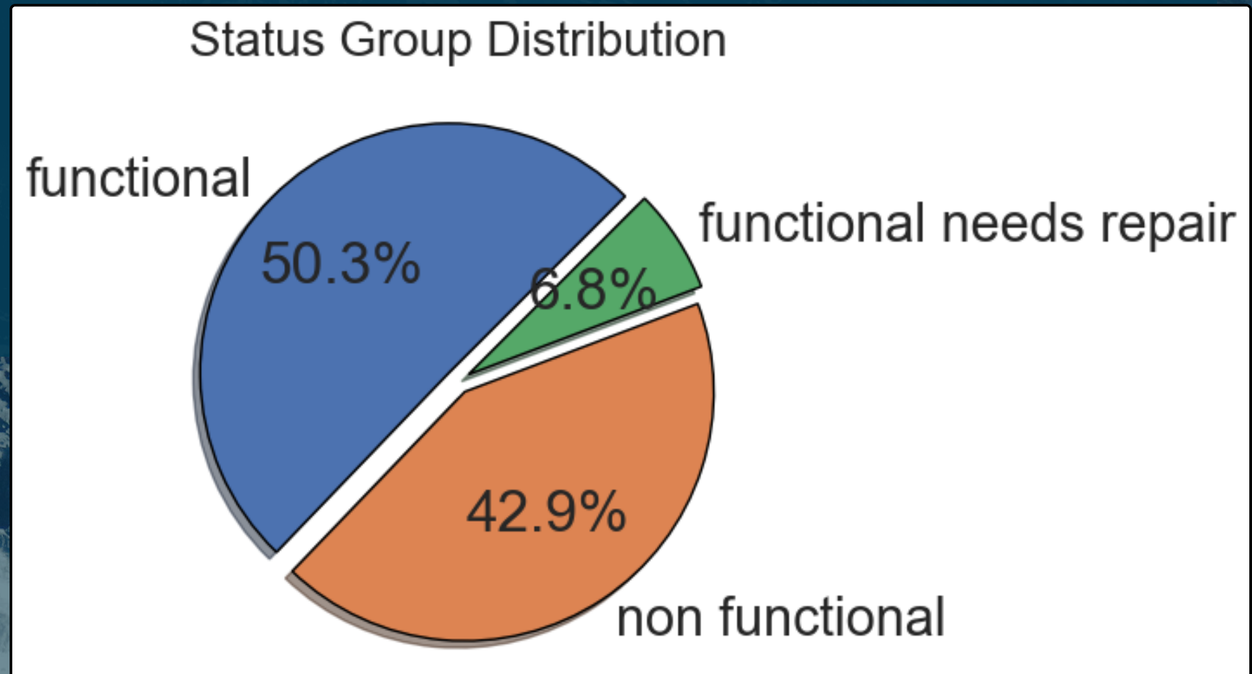
GitHub

Data Understanding

_The dataset has 10 numeric columns , 31 categorical columns and a total of 59,400 rows

_Functional waterpoints ,non functional and functional needs repair waterpoints are distributed as shown below

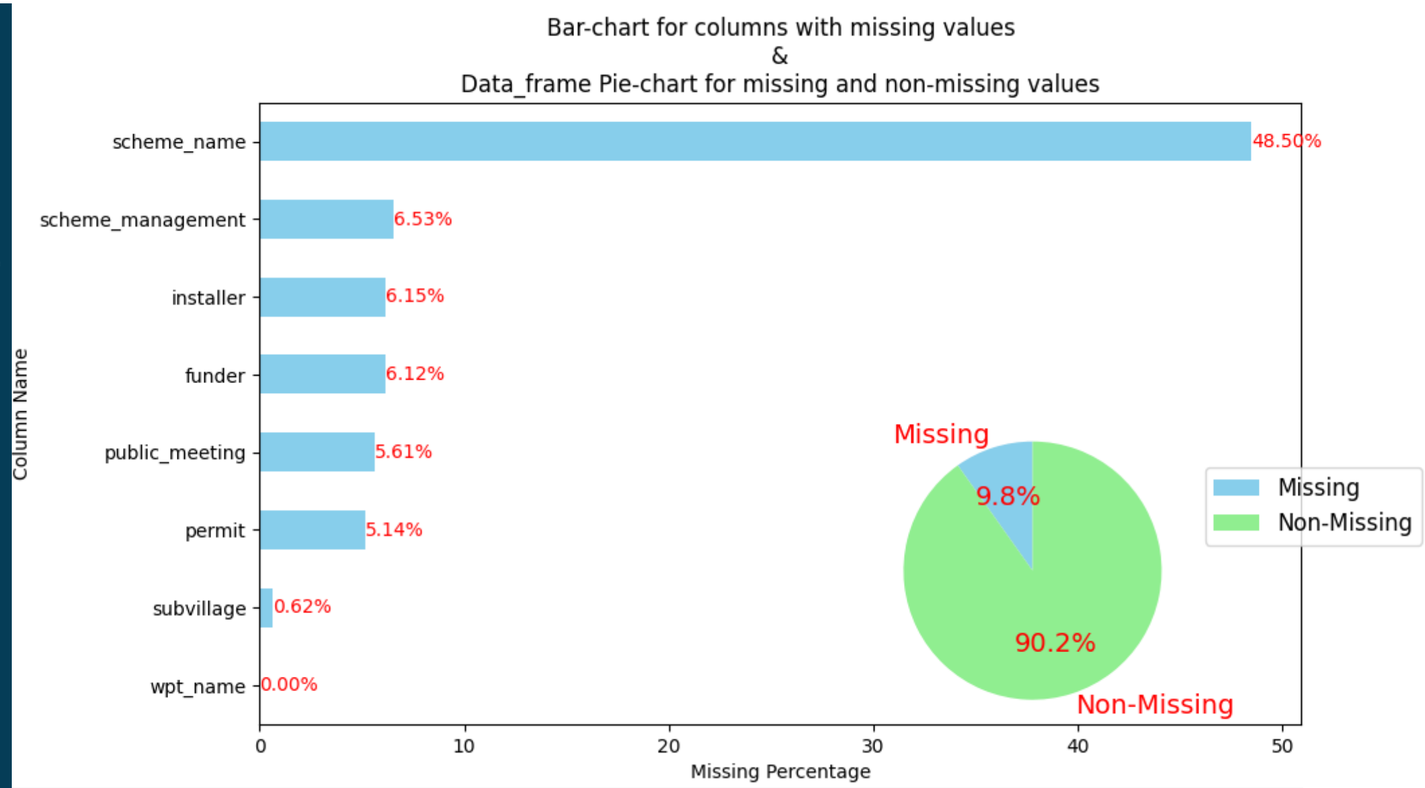
_



Data Cleaning

-After checking both missing values and outliers in the DataFrame, i found there are 8 columns with missing values and only 9.8% of the whole dataset were null values.

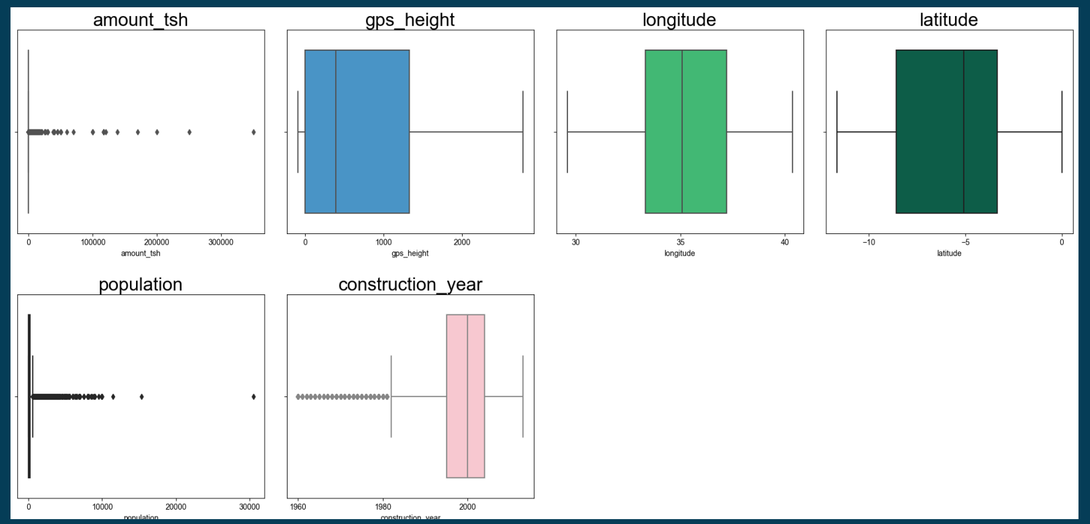
_The missing values are visualized in the adjacent image .



DATA CLEANING

as for the outliers ,6 columns were found to have outliers but only 2 columns had actual outliers the 4 others were substantiated and elaborated as to why their values were different

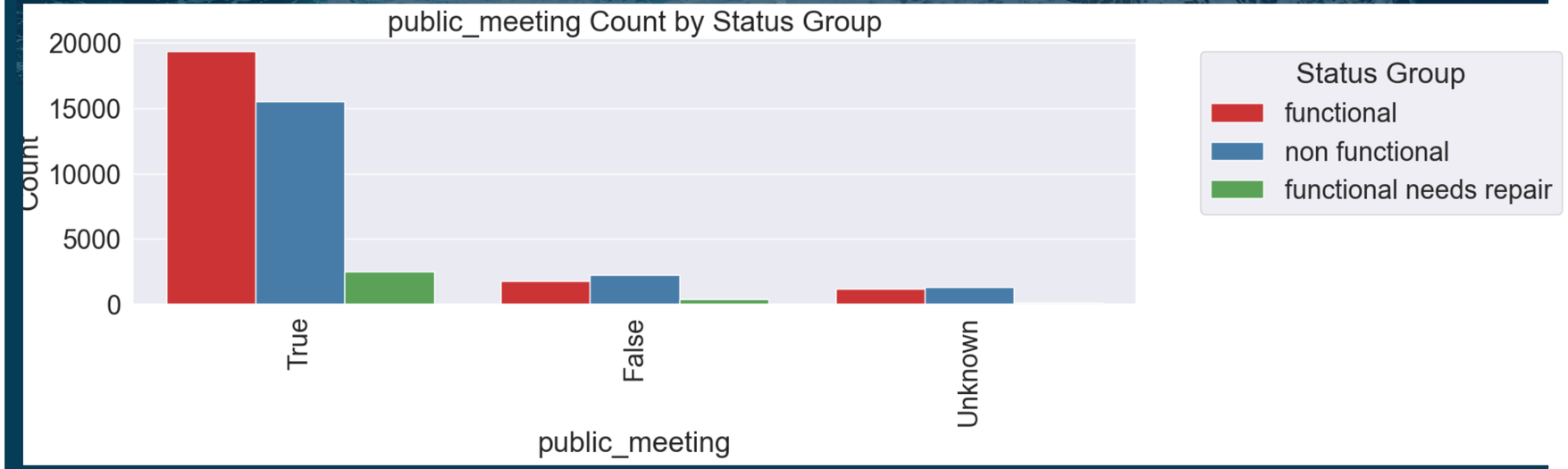
_The outliers are visualized in the adajacent image



Exploratory Data Analysis (EDA)

_Status Group being the target variable i performed several Bivariate Analysis by grouping it with 18 feature columns with an intention to get significant correlation of feature columns that can be used to solve the business problem

_one of the columns is shown below where functional water points are rampant when public-meetings are held



Modeling

_The project requirement is a binary classification model but The Target variable is in Ternary format i turned the values into a binary value through Binarization then used MultiClassifier that will enable modeling

_I only used 7 feature columns that proved significance in solving the business problem

_Three models perfomed the task and their slides are in next 3 pages.....

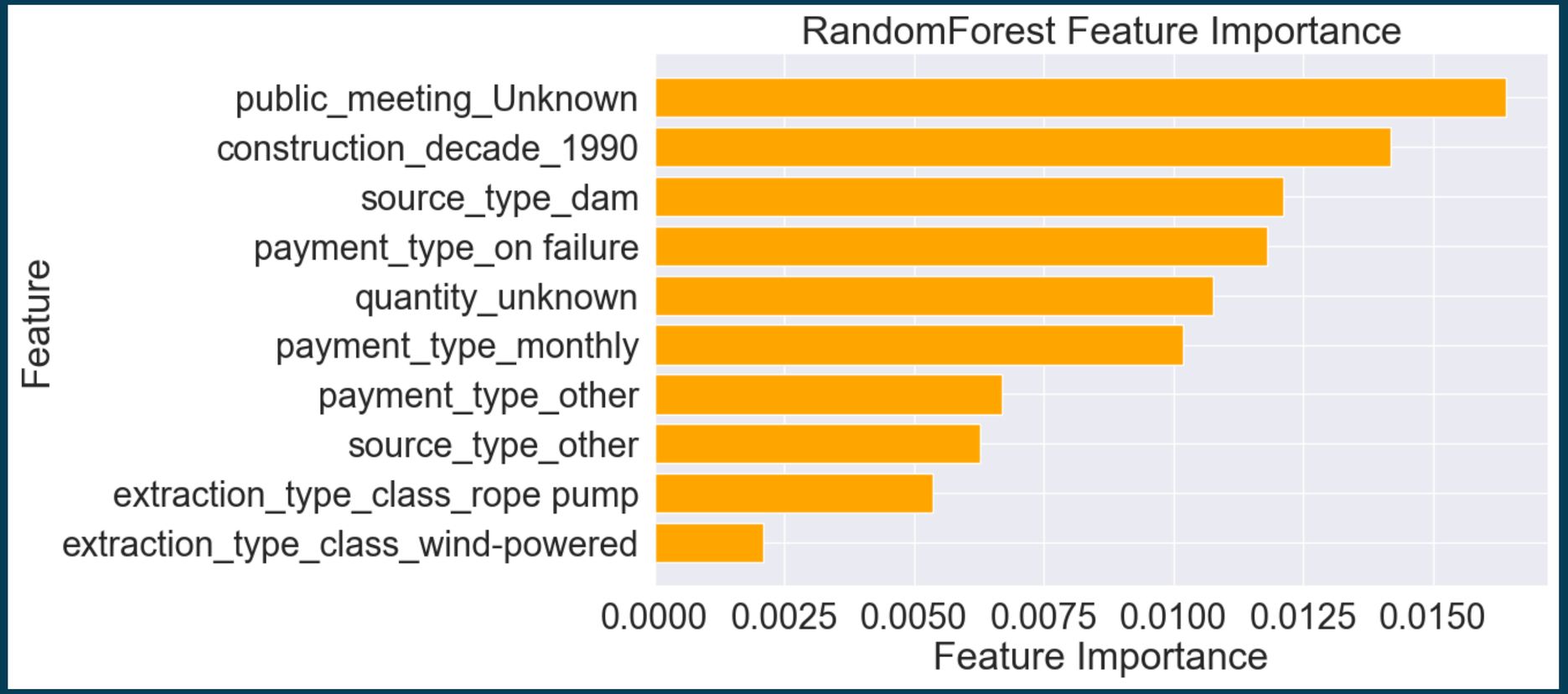
Random forest Model

Precision Score: 0.75

Recall Score: 0.71

Accuracy Score: 0.71

F1 Score: 0.71



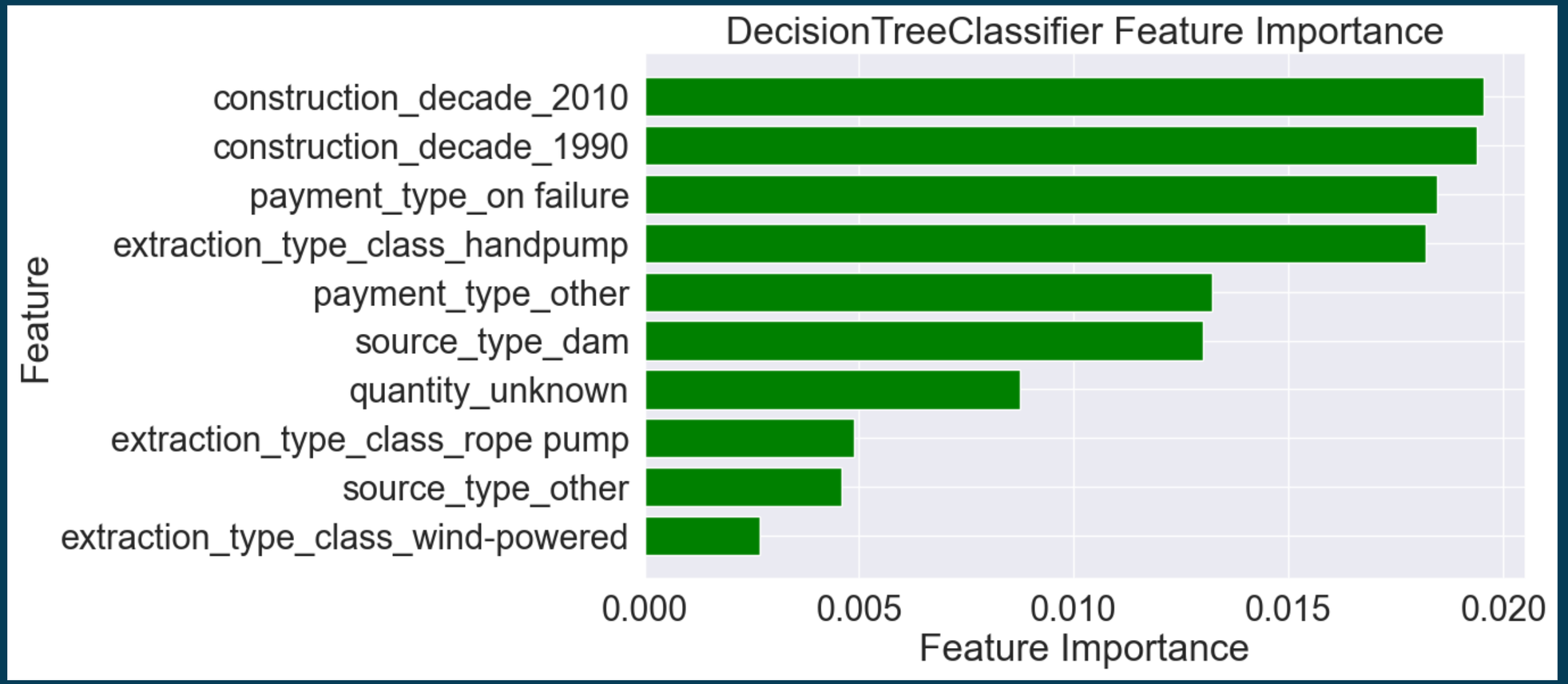
_Decision Tree Classifier Model

Precision Score: 0.75

Recall Score: 0.70

Accuracy Score: 0.70

F1 Score: 0.70



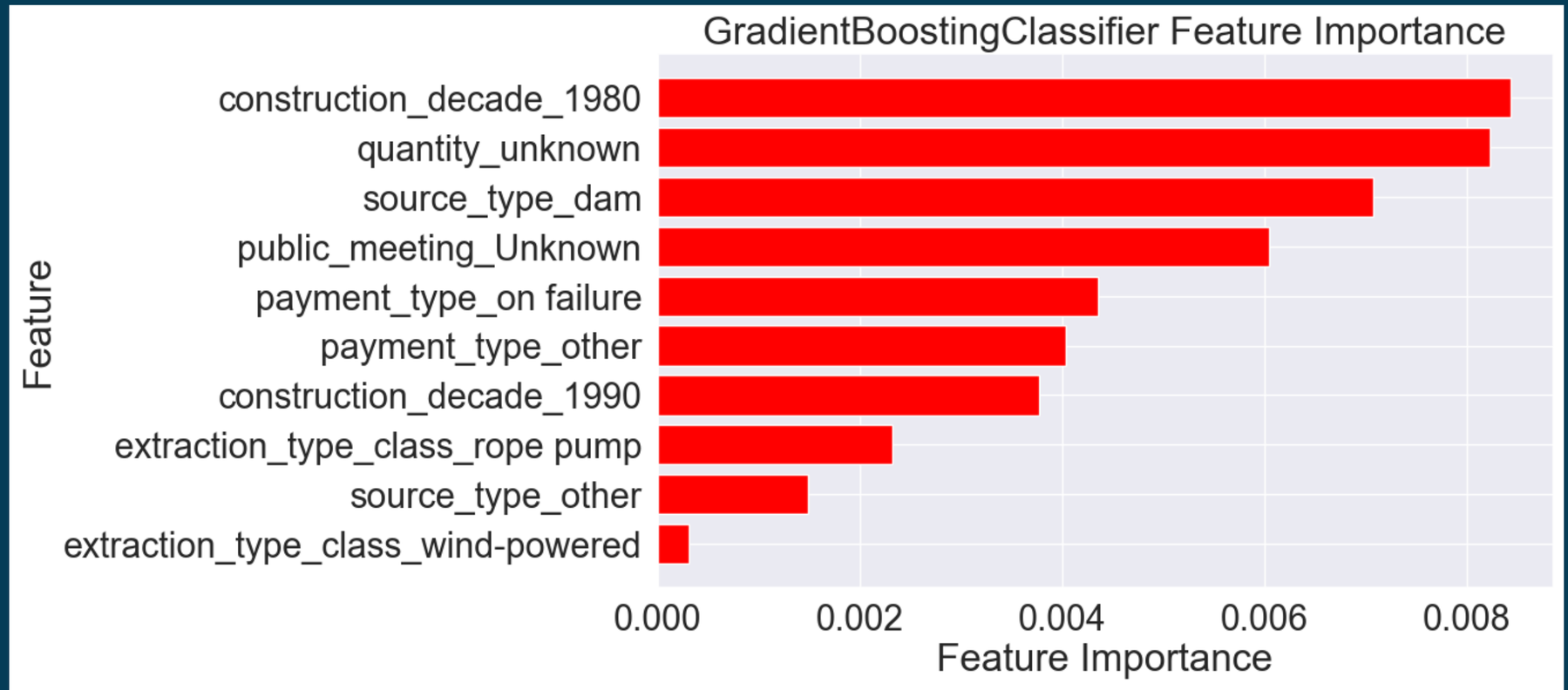
_Gradient Boosting Classifier Model

Precision Score: 0.76

Recall Score: 0.69

Accuracy Score: 0.69

F1 Score: 0.69



Recommendation

__Key features like "construction deacde","permit","public meeting","payment type","quantity","source type" serve as vital indicators of water pump functionality. The organization should prioritize these features when determining the operational status of a water pump.

conclusion

_The model demonstrated promising performance during continuous training. However, there's ample room for improvement by incorporating more recent and extensive data. With a richer dataset, I am confident that the model can yield superior predictions and enhance its overall performance. Additionally, updating the data is likely to address the existing imbalance within our dataset, leading to more accurate and reliable results.