Pump it Up: Data Mining the Water Table hosted by DRIVENDATA

By: Marissa Bush

Outline

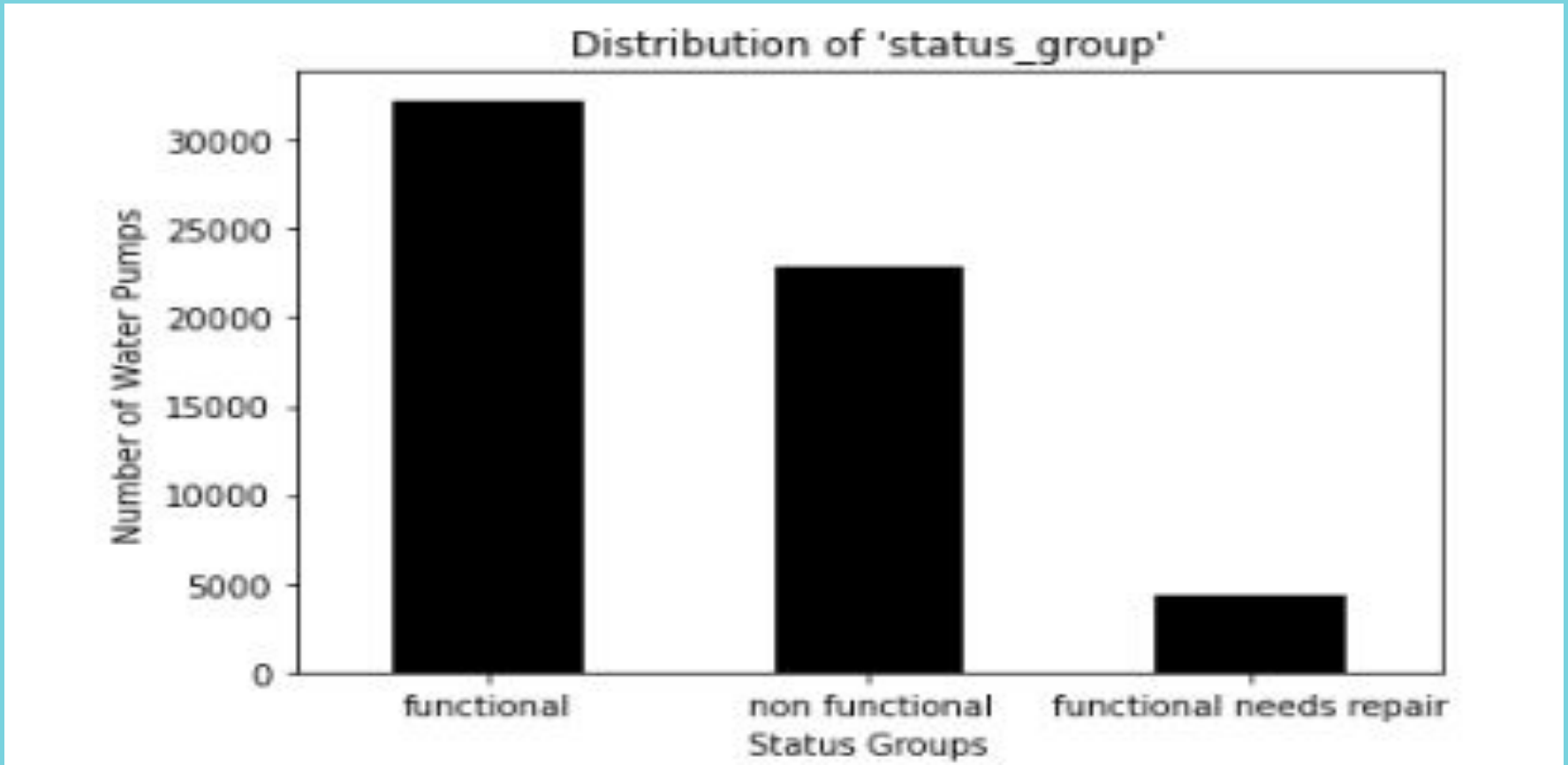
- Business Problem
- Data
- Methods
- Findings
- Conclusion + Future Work

01

Business Problem

This project attempts to find which water pumps are in need of repair based on data from Taarifa and the Tanzanian Ministry of Water.

Imbalanced Class Problem



02

Data

Began my work with two csv files:

1. training_set_labels.csv
2. training_set_values.csv

Data highlights

41 Columns, 59,400 Rows

- **Mostly categorical columns and significantly less numerical columns.**
- **The target column, “status_group”, had three classes, one with a notable difference in values. I decided to drop the third class, “functional needs repair”.**
- **A lot of the columns were redundant.**

03

Methods

OSEMN method for data analysis

Forward selection

My approach

OSEMN method for data analysis

- Obtain
  - Data collection
- Scrub/Explore
  - Clean data and feature selection
- Model
  - Ran through logistic regression, decision tree, and random forest models, iteratively.
- Interpret
  - Chose the model with the best accuracy score

04

Findings

Feature Engineering

Forward Selection

- Numerical Columns
- Basin
- Region
- Extraction type class
- Construction year
- Gps height

Feature Importance

Used this to determine which direction to go with my next selection for columns.

Feature Engineering

Categorical Columns

After the initial model of just numerical columns, I looked at feature importance and chose three categorical columns, iteratively

With all three, I used “pd.get_dummies”, which widened the dataset, but not to an overwhelming amount.

Feature Engineering

Numerical columns - “construction_year” + “gps_height”

- **Two numerical columns had a lot of zeros, over 18,000 for each column**
- **Changed those zeros to the median**

Conclusions + Future Work

RandomForestClassifier(max_depth = 20, random_state = 11)

The random forest classifier gave the best results for accuracy.

Overall score: 0.7525

Current rank: 4,557 out of 14,106

Future Work

For future work, I'd be interested in adding more categorical variables and studying how those affect the models.

Also, I'd be interested in learning more about the water pumps themselves, and applying any new background knowledge to the feature selection process.

Thank you!

Do you have any questions?

Marissabush.02@gmail.com

GitHub:
https://github.com/Marissa841/phase_3_project