JONNY HOFMEISTER

Tanzanian Waterpoint Classification

Predicting Waterpoint Functionality

Presentation Overview

Discussion topics for today

01
Business Probelm

02
Data

03
Modeling Results

04
Moving Forward

Business Problem

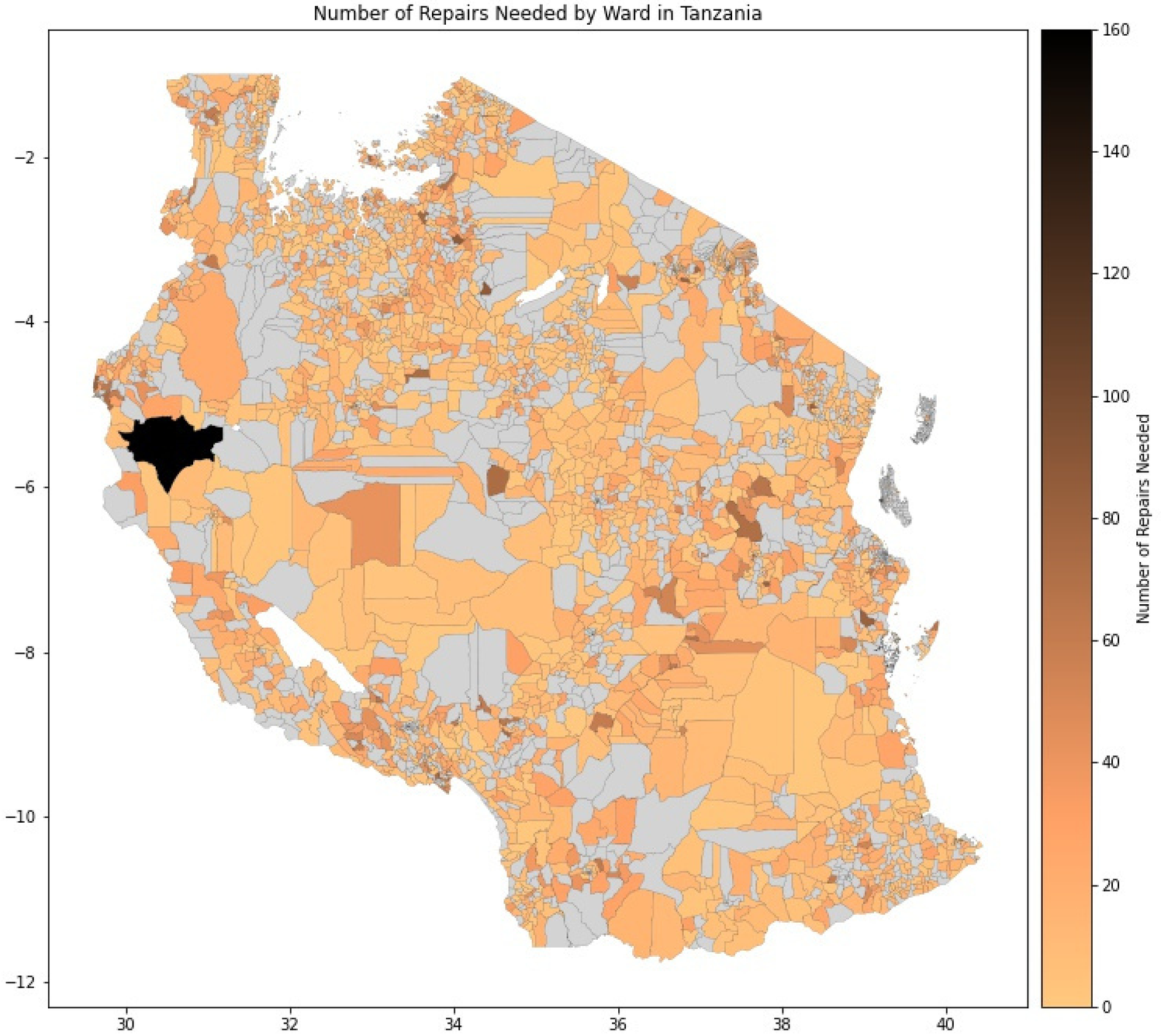
These are our main goals:

Predict Functional Status of Waterpoints

- Train a model on waterpoints with known functionality.
- Predict the functionality of unknown waterpoints across Tanzania.
- Then, determine best practices for addressing repair.

Map

Shows the extent of missing data



Data Understanding

Target: Functional Status Group

Ternary Classification

- 'Functional'
- 'Non Functional'
- 'Functional Needs Repair'

Metrics:

How to analyze classification success

In the context of efficiently performing repairs, a FN isn't any worse than a FP. So balance Precision and Recall

Goal: Maximize F1 score

Confusion Matrices and Accuracy will also be used in evaluation.

ROC-AUC isn't a good score for multi-class probelms

Data Understanding:

Features

Scalers and Encoders:

- MinMax Scaler
- Simple Imputer
- One-Hot Encoder
- Count Encoder

Numericals:	Total Static Head, Population, Year Created, etc
Booleans:	Public Meeting, Permit
Categoricals:	Can be OHE or Freq Encoded
OHE:	Scheme Management, Extraction Type, Quantity, Source
Frequency:	Location based columns - Region. LGA, Ward, etc

Pipeline Process

End to End modeling with Sklearm →

01

Separate features

Sort columns by datatype

02

Impute missing values

Impute NaNs and zeros in categorical and numerical columns. Fill accordingly

03

Scale, One-Hot and Count Encode

Scale Numericals

OHE and Count (freq) encode categorical columns

04

Fit Models and Score

Fit the different model types.

Score them with desired metrics.

Examine confusion matrix.

05

Iterate Modeling

Iterate through modeling process.

Search for best parameters.

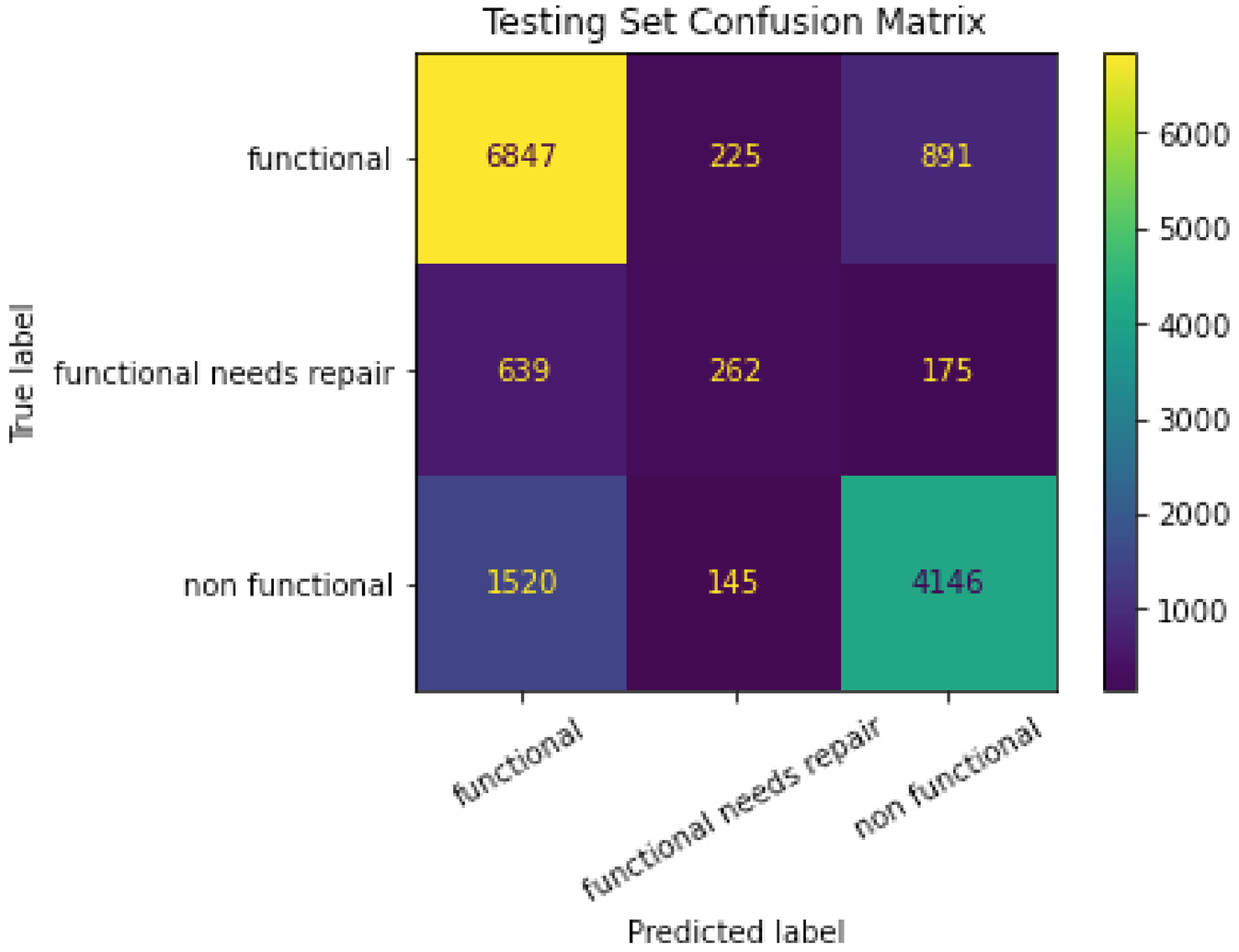
Optimize the model for target metrics.

Repeat.

K-Nearest Neighbors Model

Determines similarity between samples using distance metrics

- Training F1: 0.8
- Testing F1: 0.75
- Training Acc: 0.81
- Testing Acc: 0.76



Random Forest

SCORES BEFORE PARAMETER TUNING

Training F1: 0.96
Training Acc: 0.96

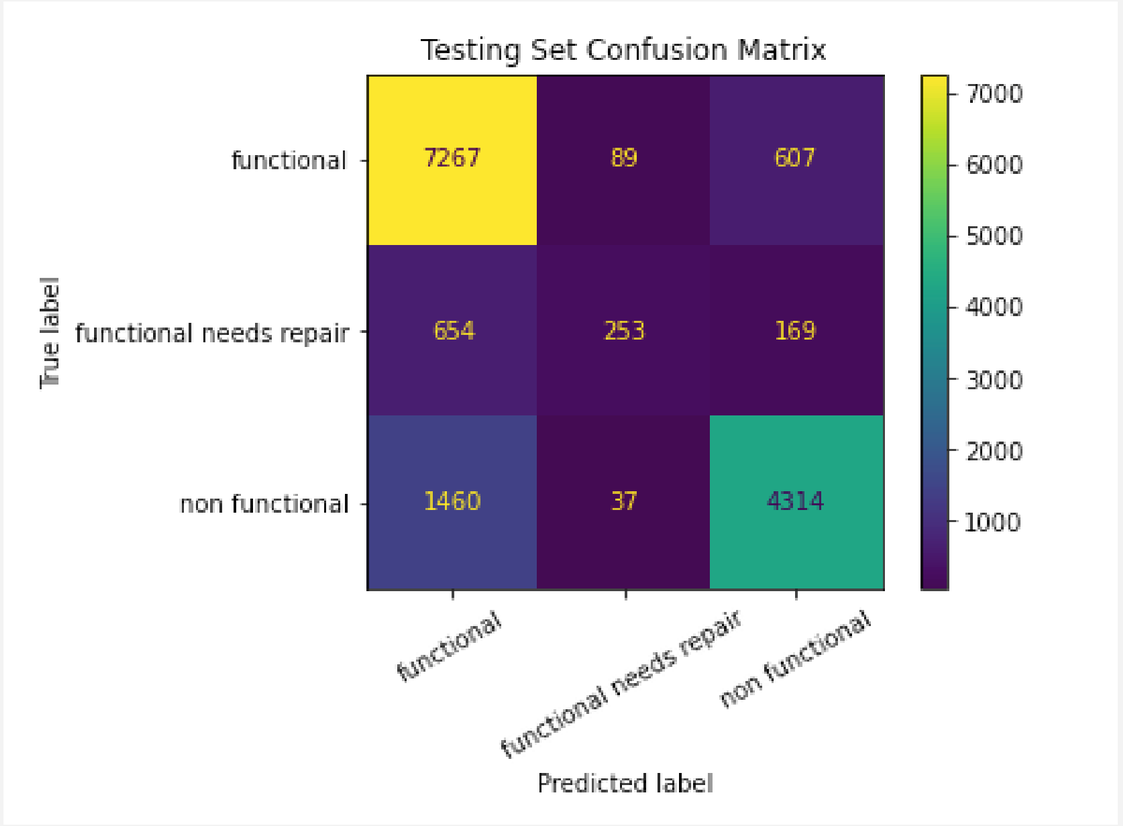
Testing F1: 0.76
Testing Acc: 0.78

GRID SEARCH

Improved testing scores by 2% and reduced overfitting by 10%.

PARAMETERS

n_estimators = 160
max_depth = 30
min_samples_split = 2
min_samples leaf = 2



Decision trees created to classify samples with splits/if statements

8 out of 10

Pumps predicted with correct functionality

0.78

Best testing F1

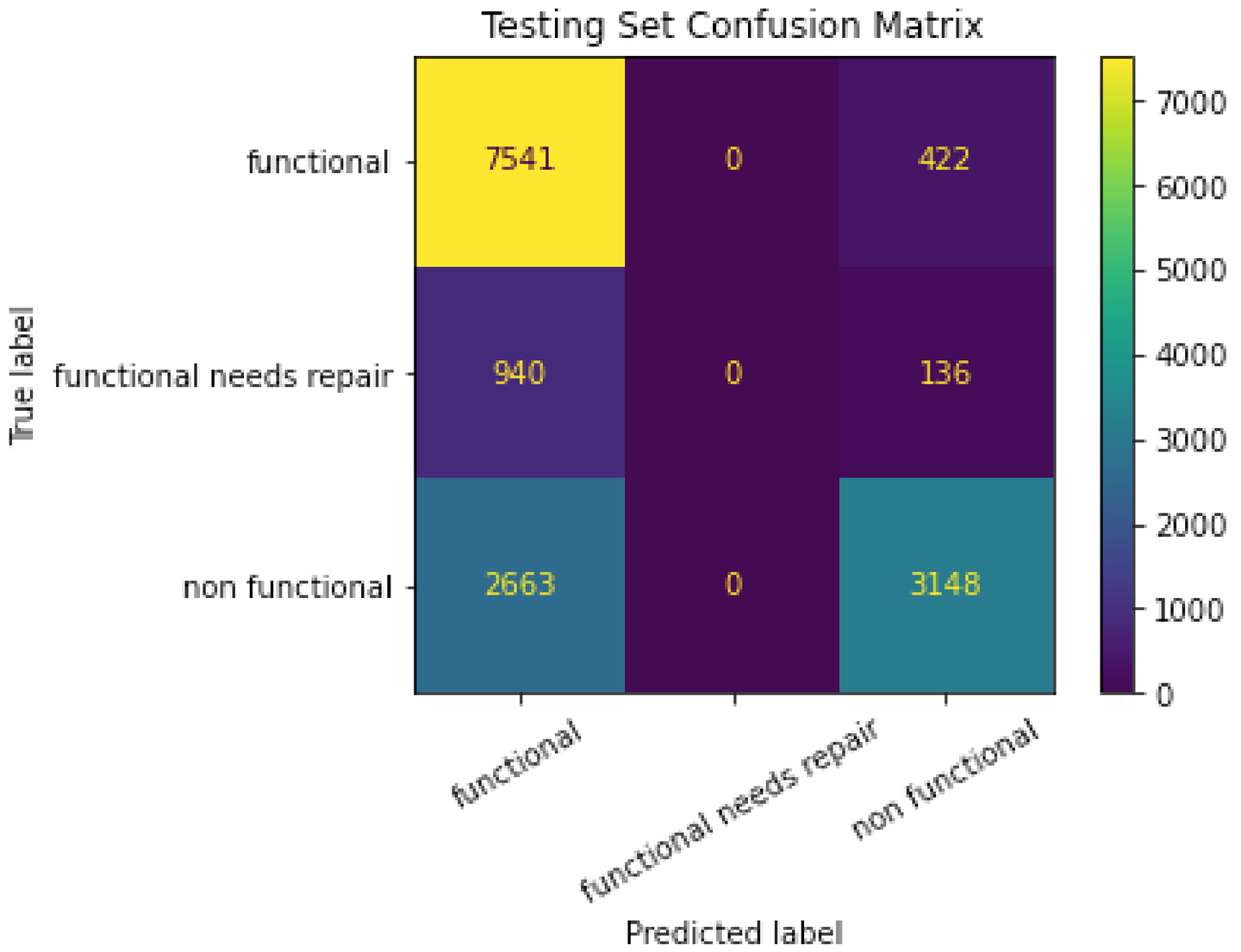
80%

Best testing accuracy

Support Vector Classifier

Manipulates parameter space to create optimized decision boundaries for classes

- Testing F1: 0.68
- Testing Acc: 0.72
- Failed to predict 3rd class



Results

Current Random Forest model can fill missing data with 80% accuracy and balanced Precision and Recall

Next Steps

Rigorous Feature Selection

Search for features that produce signal and not noise.

If we want to predict functionality in wards with no data, perhaps location features wont add any information to the model, only overfit it.

Better Imputation Strategies

Rather than impute features like construction year with the median value, impute with a random selection from a probability distribution representing how often each year shows up in the data.

Repairs

Use the model to begin to fill in missing pump statuses in Tanzania.

Address the TSMs goals and constraints for repairs, begin to generate actionable insights for most efficient methods.

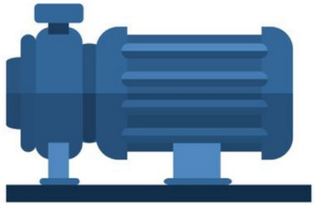
Prioritize populated areas? Certain types of pumps? Etc.

Resource Page

Sources and Citations

Dataset on DrivenData competition:

DRIVENDATA



Pump fundamentals

Sklearn Documentation

JONNY HOFMEISTER

Thank you for your time!

GitHub Link and email:
jonny.hofmeister@gmail.com