Water pump distribution in Tanzania

Rahul Krishnan

Presentation

Business Objective

To build a model for the Govt of Tanzania that will help predict the status of a water pump based on certain input information.
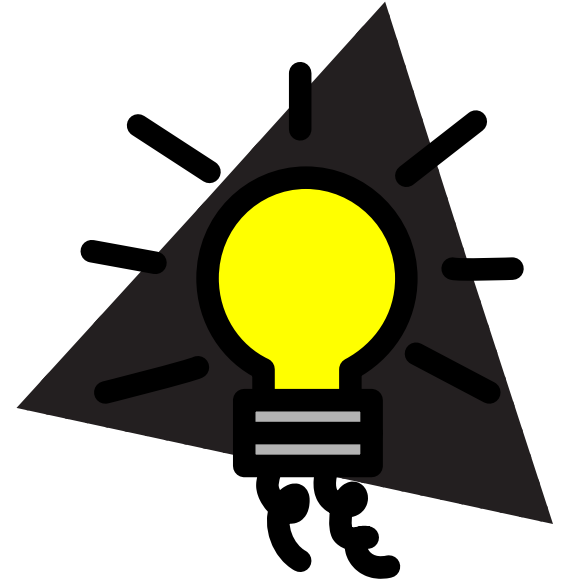
The water pump will be classified as follows:

1. Functional
2. Non-Functional
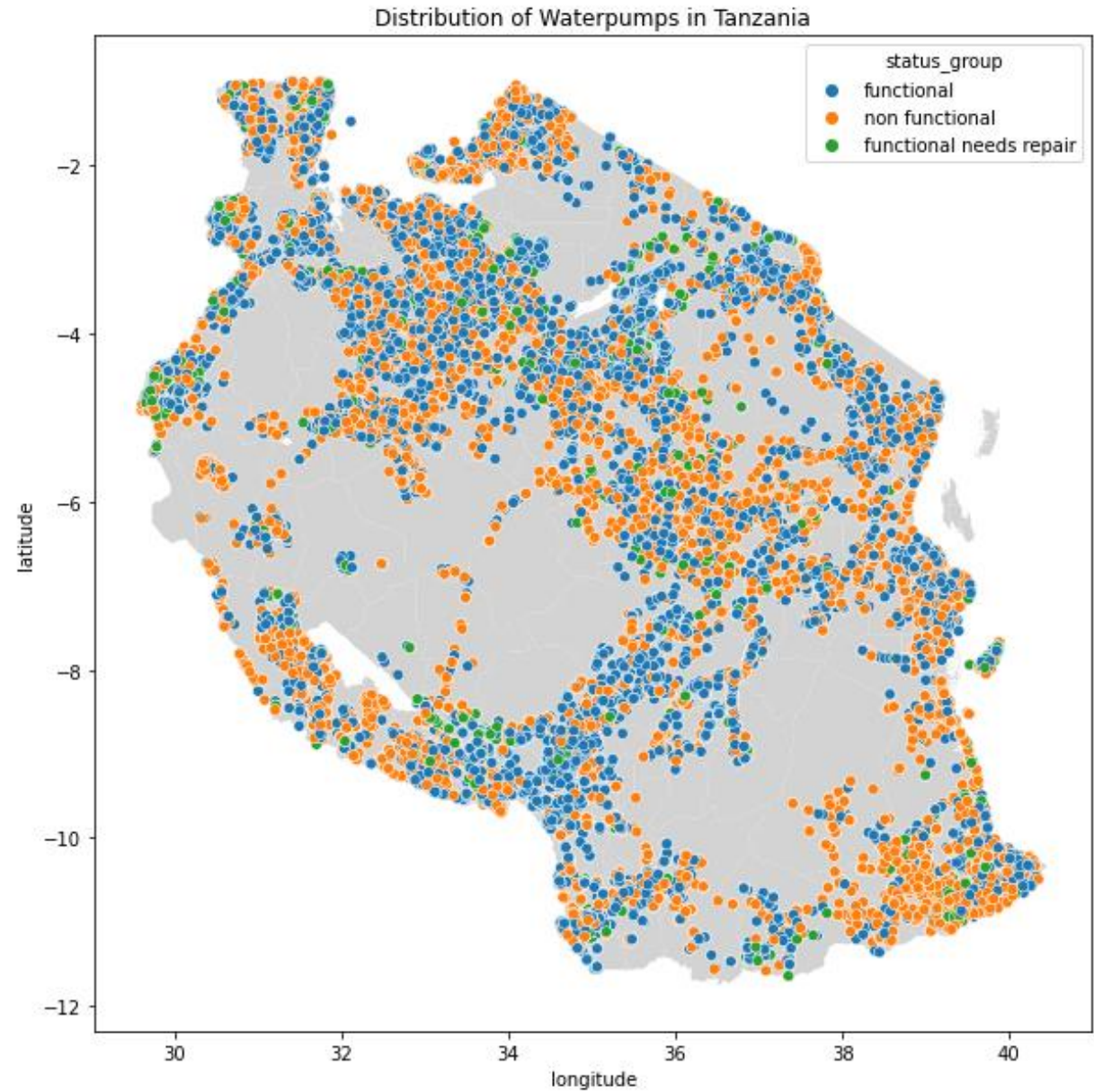3. Functional needs repair

Dataset

Dataset sourced from :
https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/
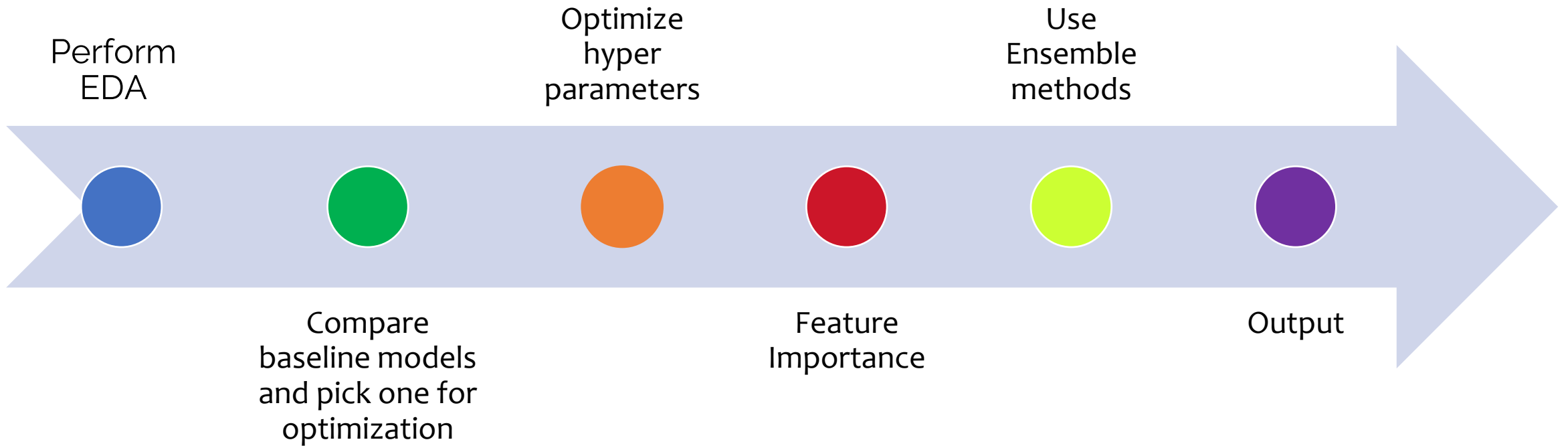
Water pump distribution by class

Distribution of Waterpumps in Tanzania

Process Steps

Perform EDA

Optimize hyper parameters

Use Ensemble methods

Compare baseline models and pick one for optimization
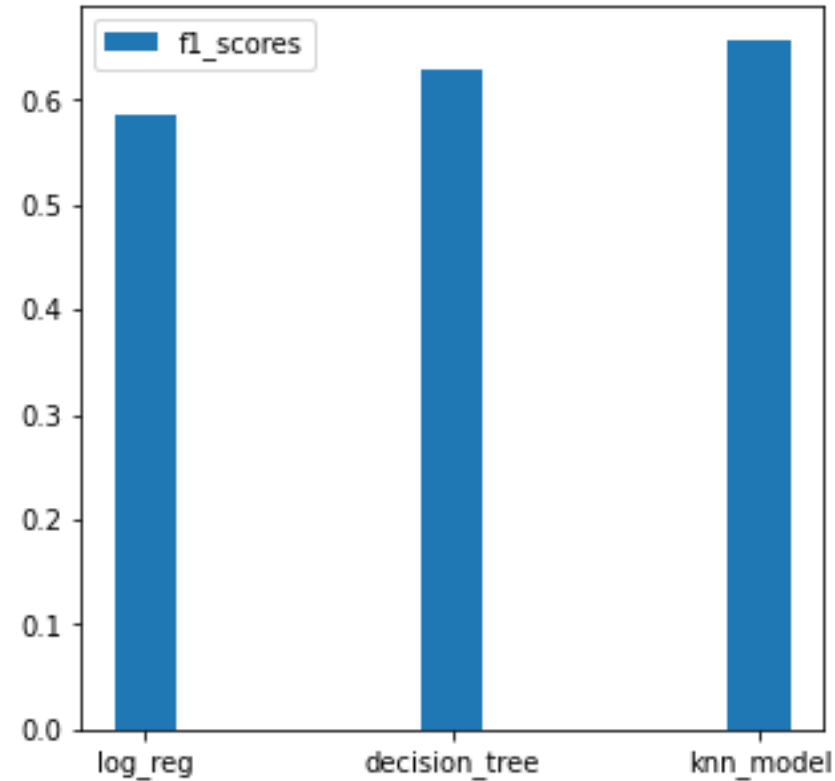
Feature Importance

Output

Baseline models stats

Since this is a multi-classification problem, we will look at the F1-score , which combines both accuracy and recall.

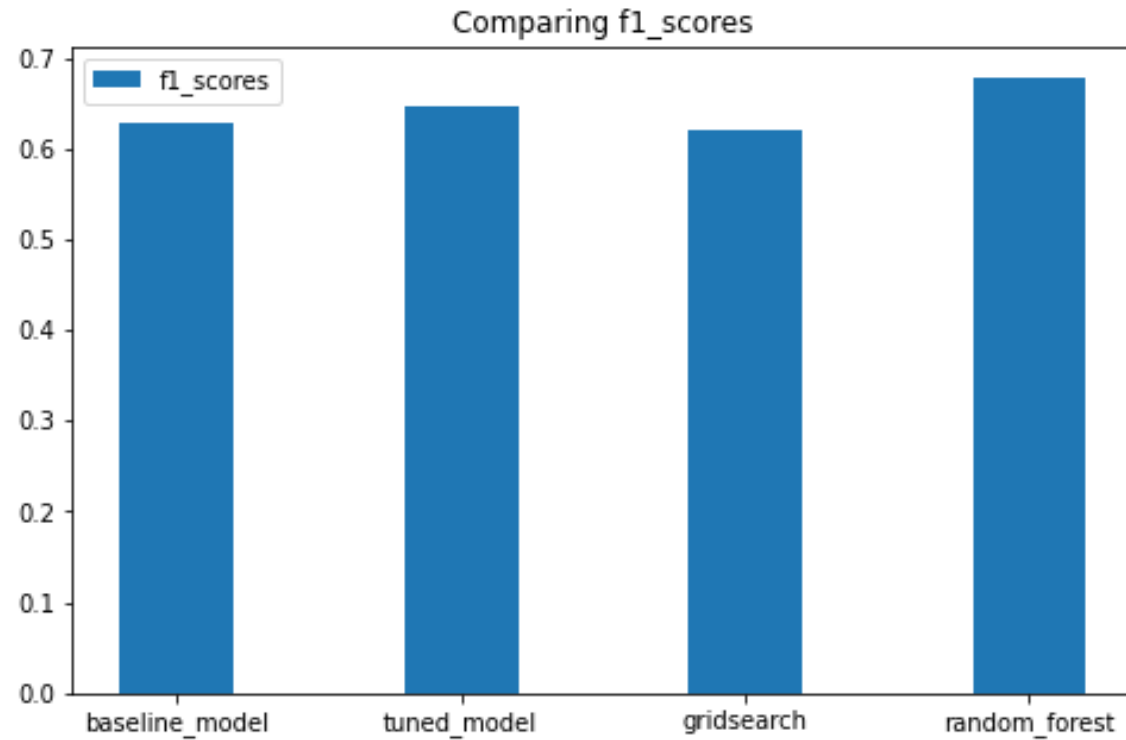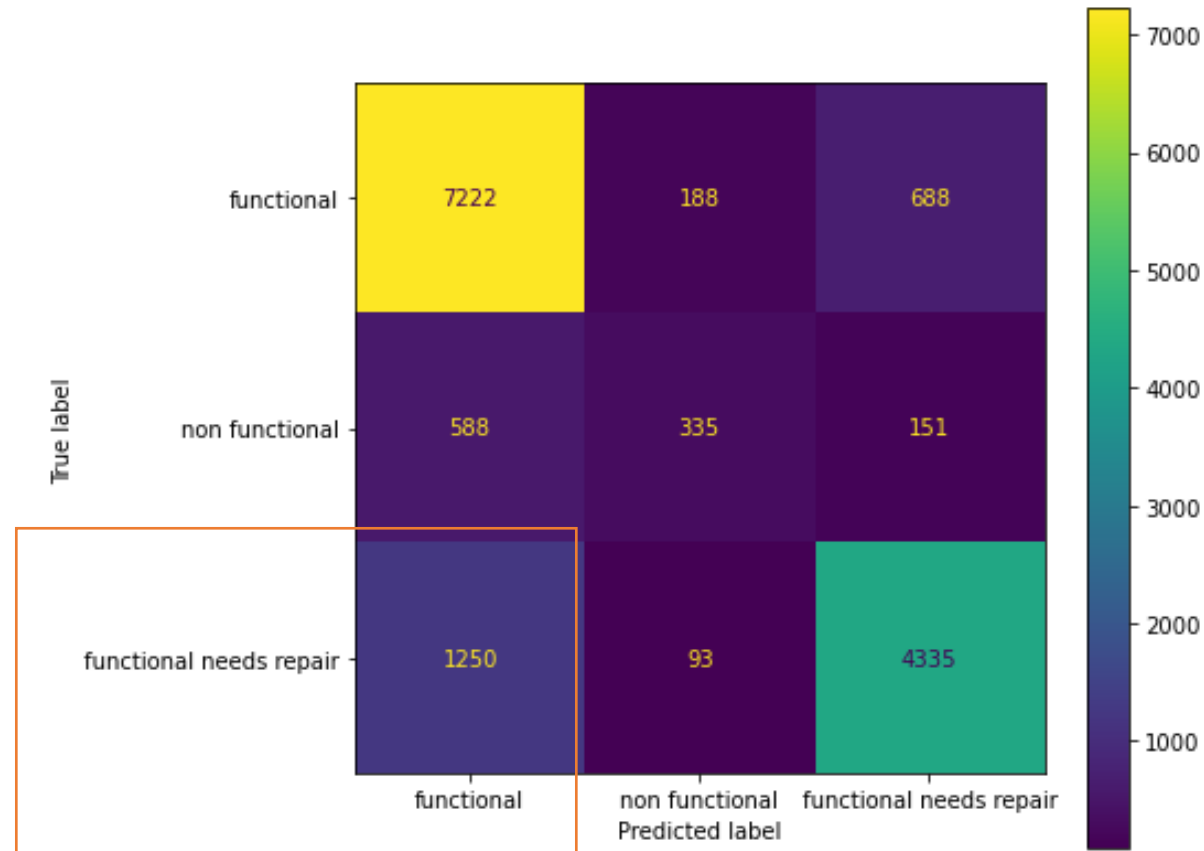|	Logistic Regression	Decision Tree	KNN
F1 - Score	0.59	0.63	0.66

Baseline model comparison

Decision Tree vs KNN:
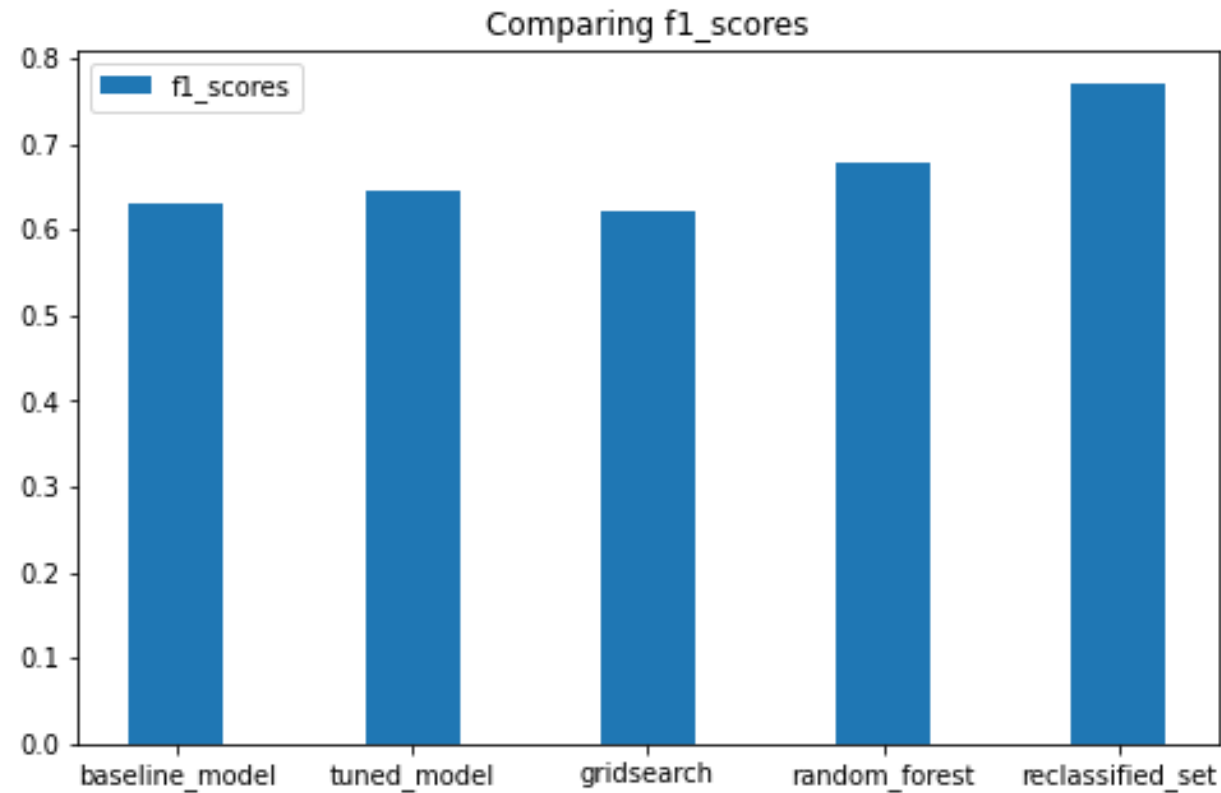1. Reduced computation time
2. More parameters to tune

Model comparison of Decision Tree

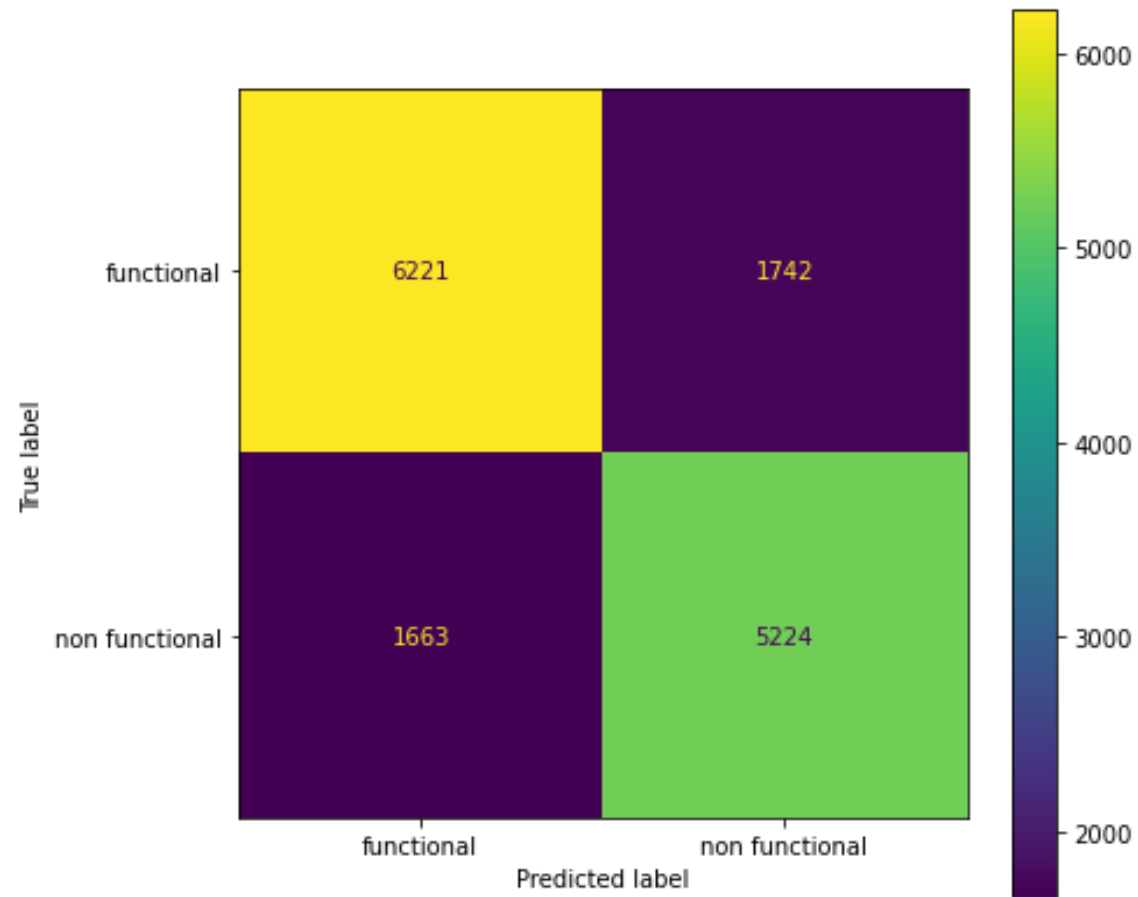# Confusion Matrix – Random Forest

Balanced vs Unbalanced

Comparing f1_scores

Final Results

Next Steps

- Hyperparameters should be tuned for the new dataset optimization.
- We can combine GrdSearchCV and RandomForest and evaluate performance
- Since we only used some the categorical features for our model, we can selectively add more features and check for performance.

Thank You!

This Presentation is Prepared by
Rahul Krishnan