

How well are the wells?

Predicting Tanzanian Water Wells Functionality

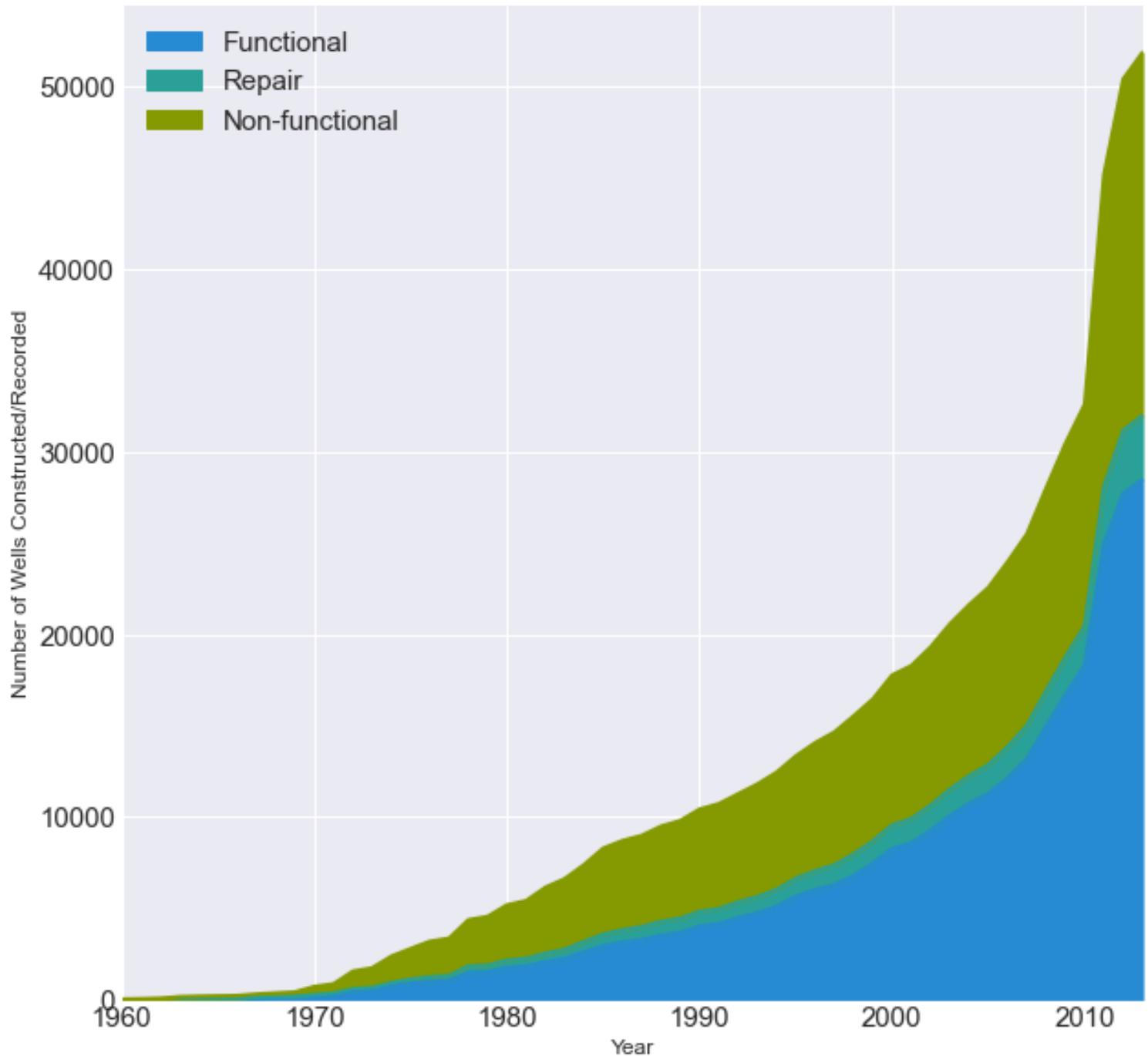
Sung Bae

Tanzanian Water Crisis Facts

- Importance of water
 - Source of life
 - Improve life quality
 - Increase school attendance
 - Empowers families
- 24 million people without basic access to safe water
- Numerous organizations have been working to provide safe and accessible water



Cumulative Number of Wells Over Time



Tanzanian Water Crisis Facts

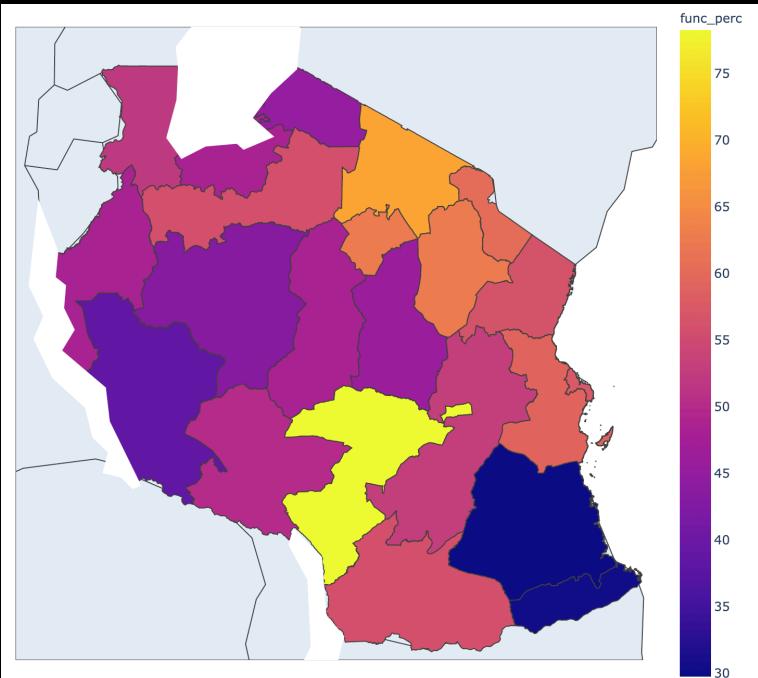
- More than 50,000 water wells are installed by 2013
 - 54.9 % Functional
 - 38.3 % Non-functional
 - 6.8 % Needs repair
- Imperative to determine which wells are not functioning or need repairing accurately

Goals and Objectives

- Construct a model that can classify
 - [1] functioning wells,
 - [2] functioning but need repairing wells, and
 - [3] non-functioning wells.
- Top Priorities
 - Recalls for [2] and [3]
 - Overall accuracy
- Provide features that affect functionalities

Geographical Factors

Percent Functioning Wells

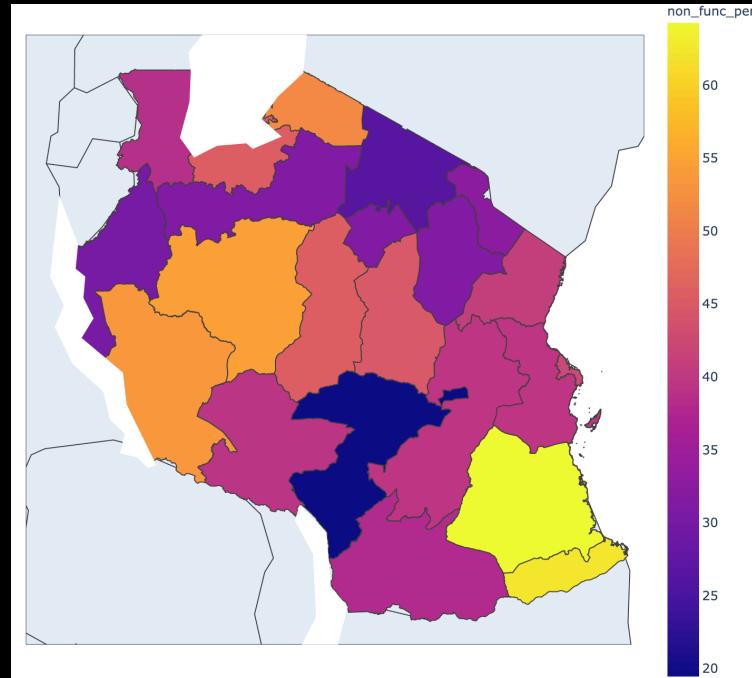


Region: Iringa

N = 5294

Percentage: 78.22%

Percent Non-Functioning Wells

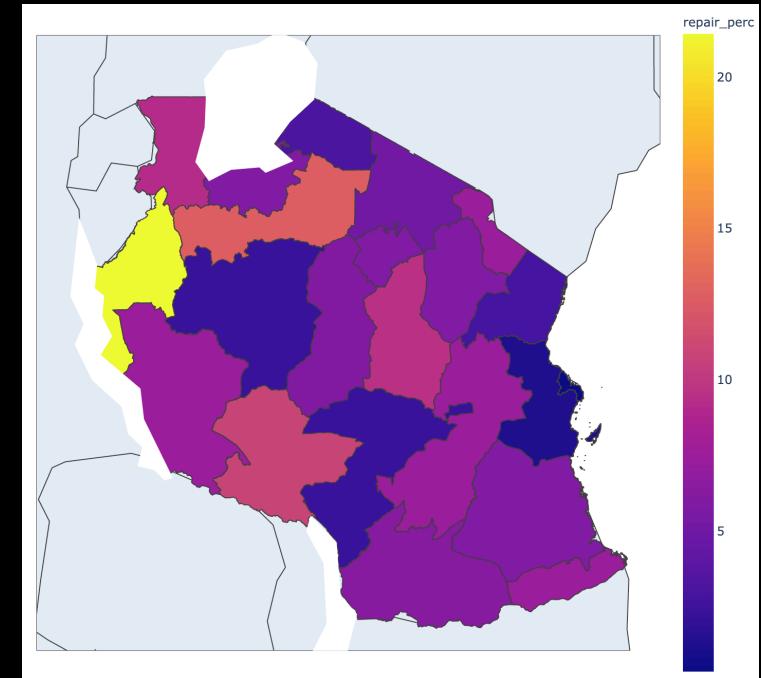


Region: Lindi

N = 1546

Percentage: 64.23%

Percent Need-Repairing Wells

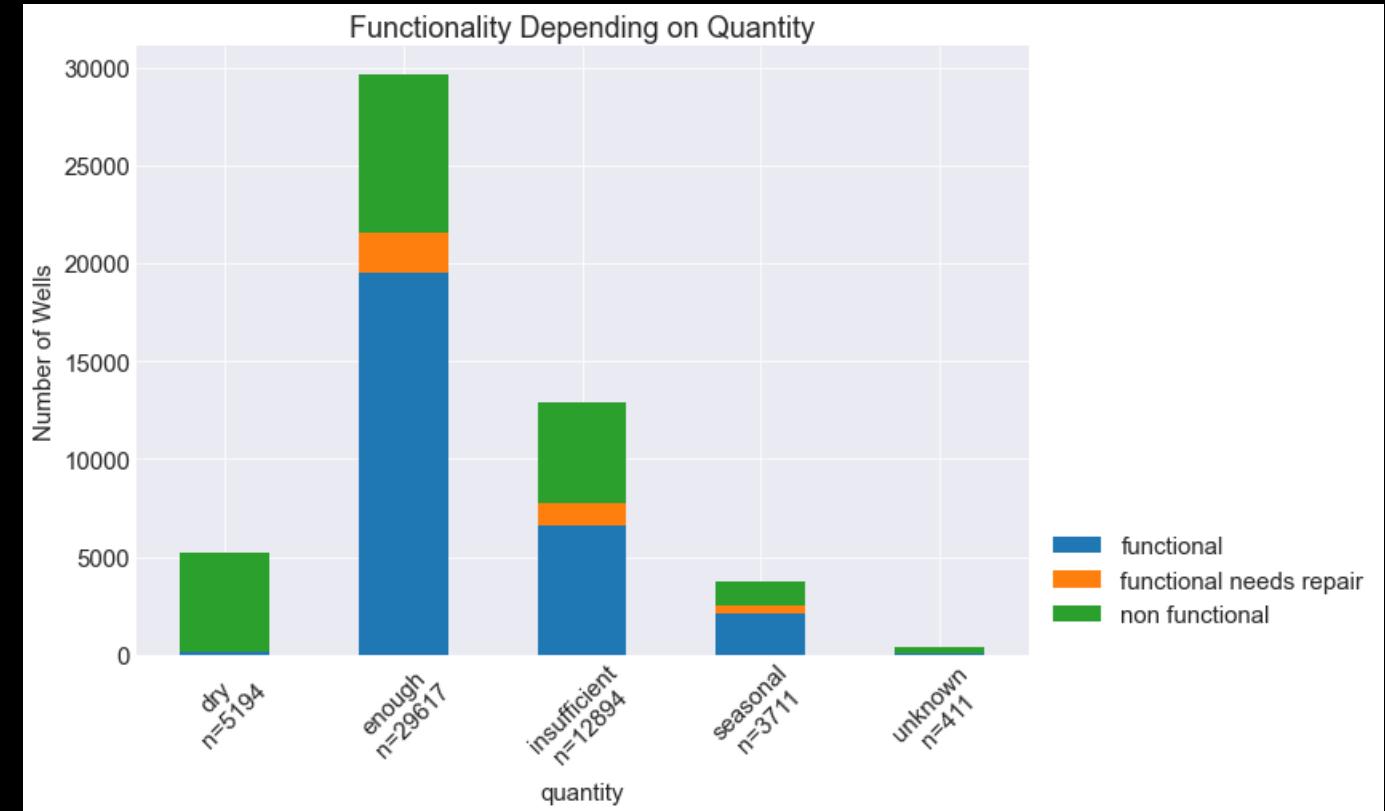
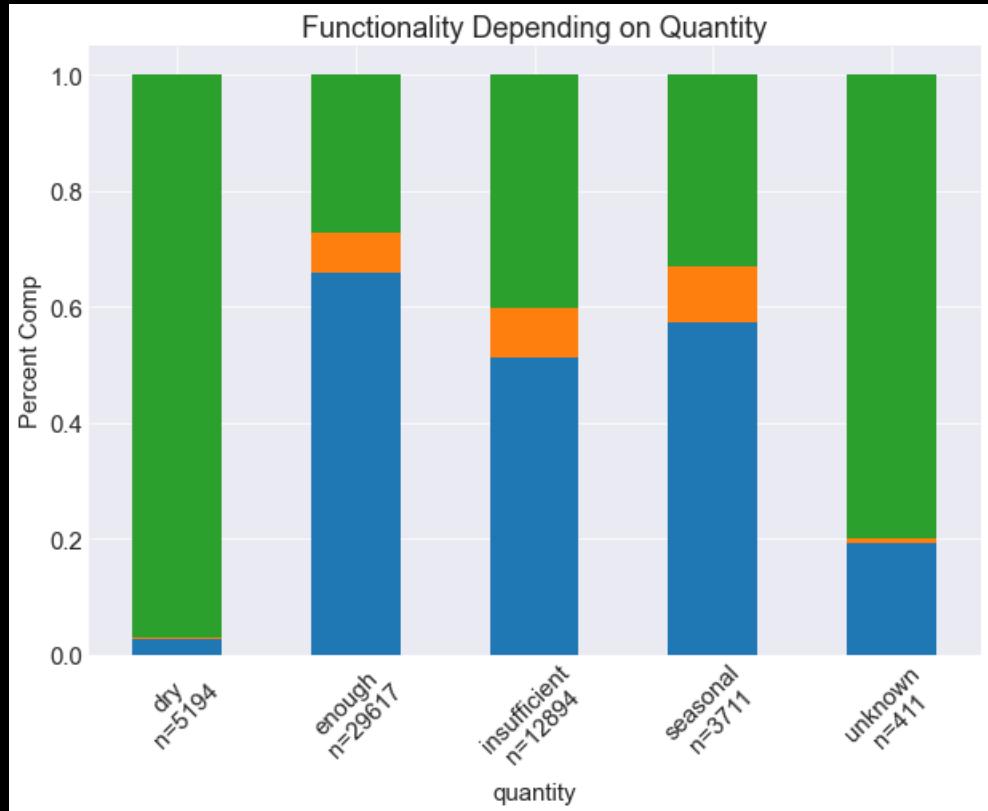


Region: Kigoma

N = 2816

Percentage: 21.41 %

Quantity Factors



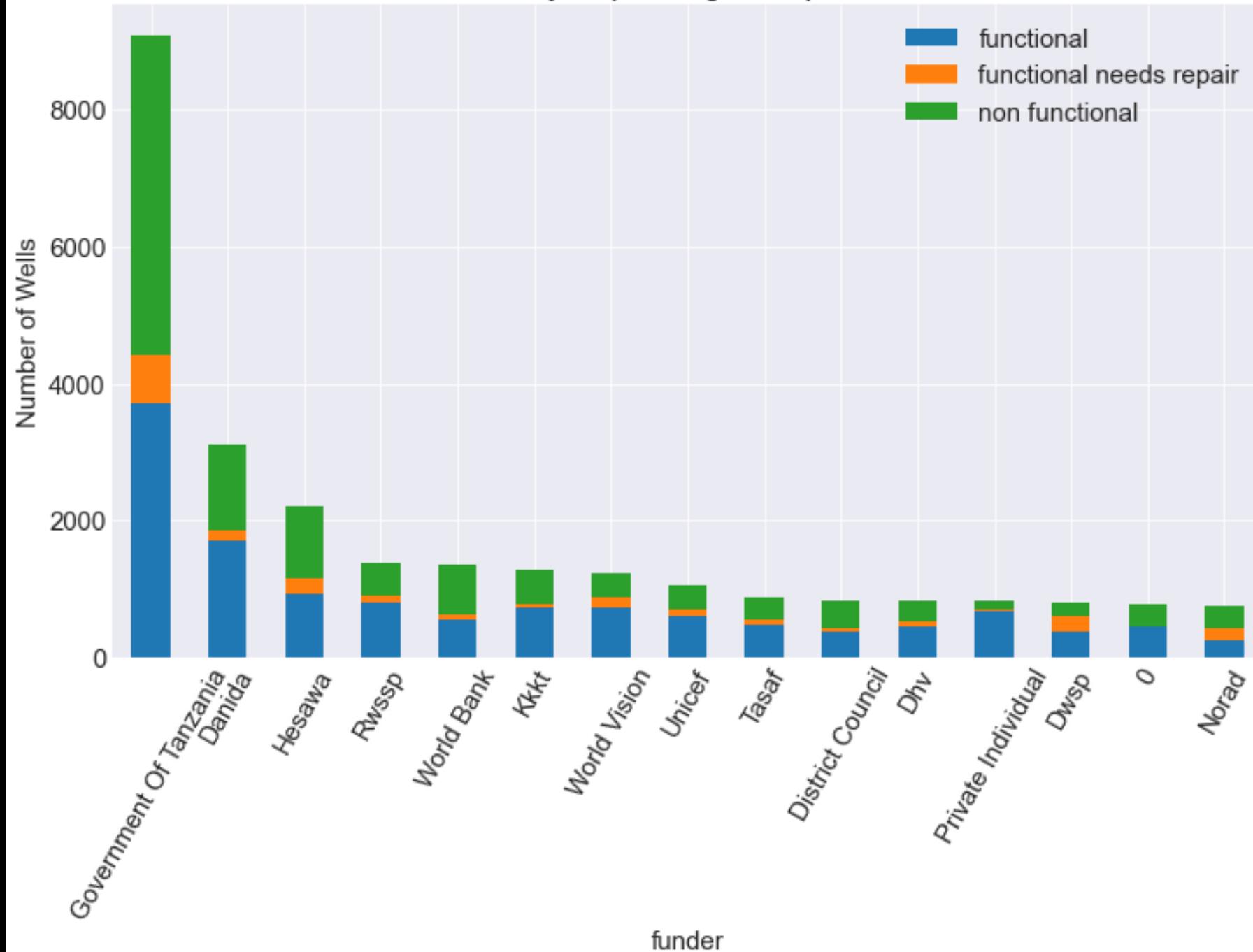
More than 90% non-functional wells for **"dry quantity"** wells.

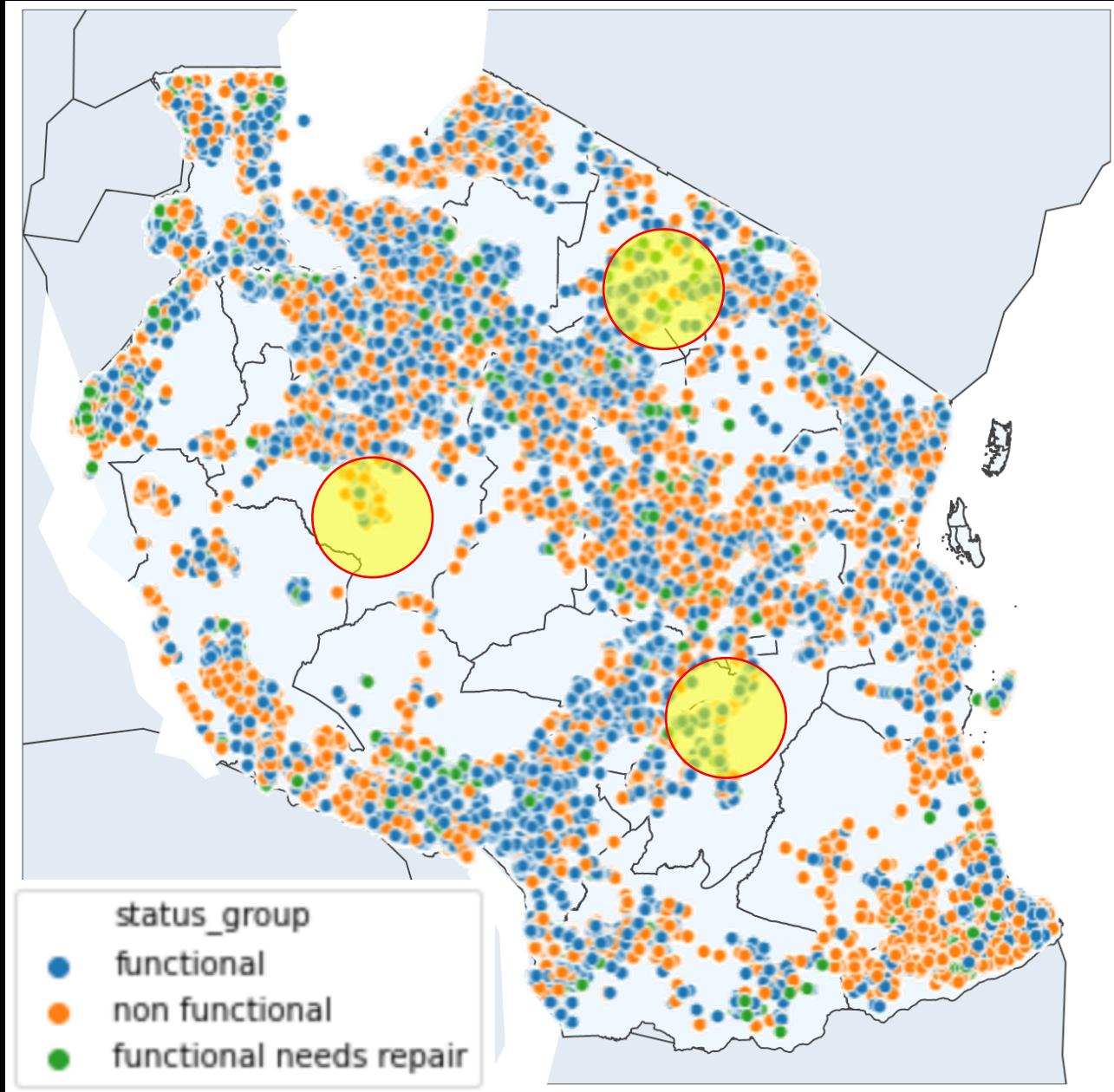
More than 80% non-functional wells for **"unknown quantity"** wells.

Funders Factor

- More than 50% of wells funded by the government are either not functioning or need repair

Functionality Depending on Top 15 Funders





Neighboring

Probability of finding

[1] functioning

[2] need repairing

[3] non-functioning

are calculated within 30 km of
each water well

Our Plan

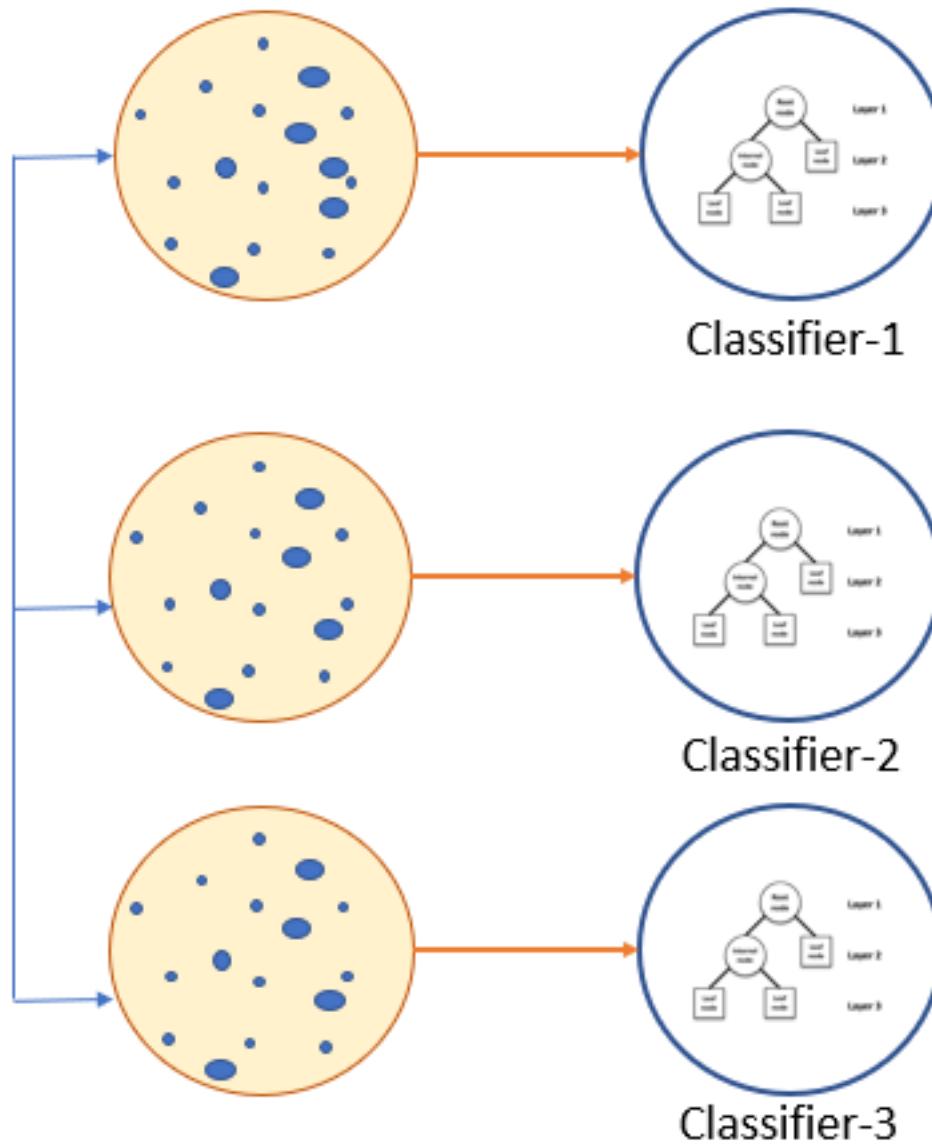
Bagging

- Decreases model's variance
- Examples
 - Random Forest
 - Extra Trees

Boosting

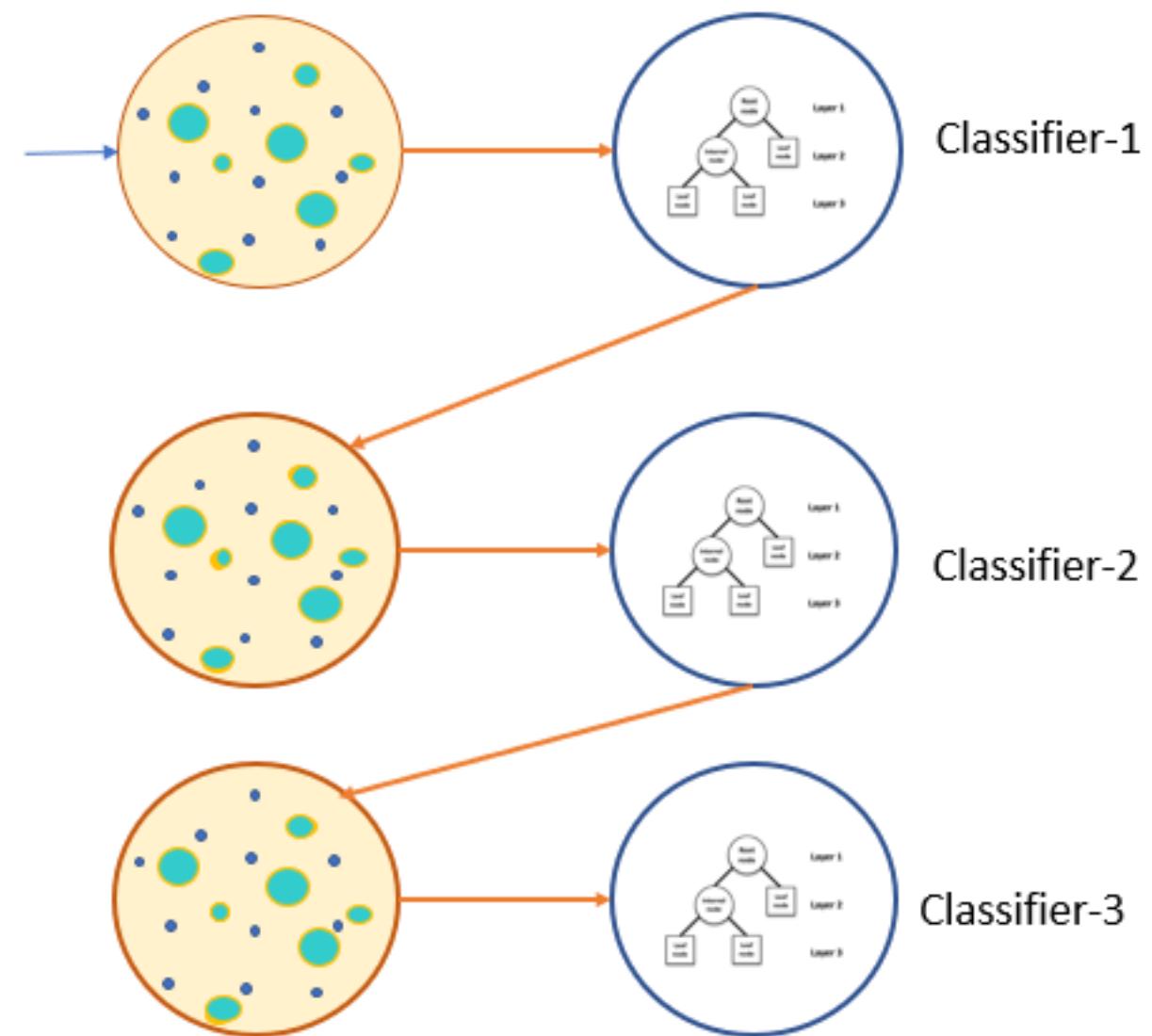
- Decreases model's bias
- Examples
 - XGBoost
 - Adaboost
 - Gradient Boost

Bagging



Parallel

Boosting



Sequential

Our Plan

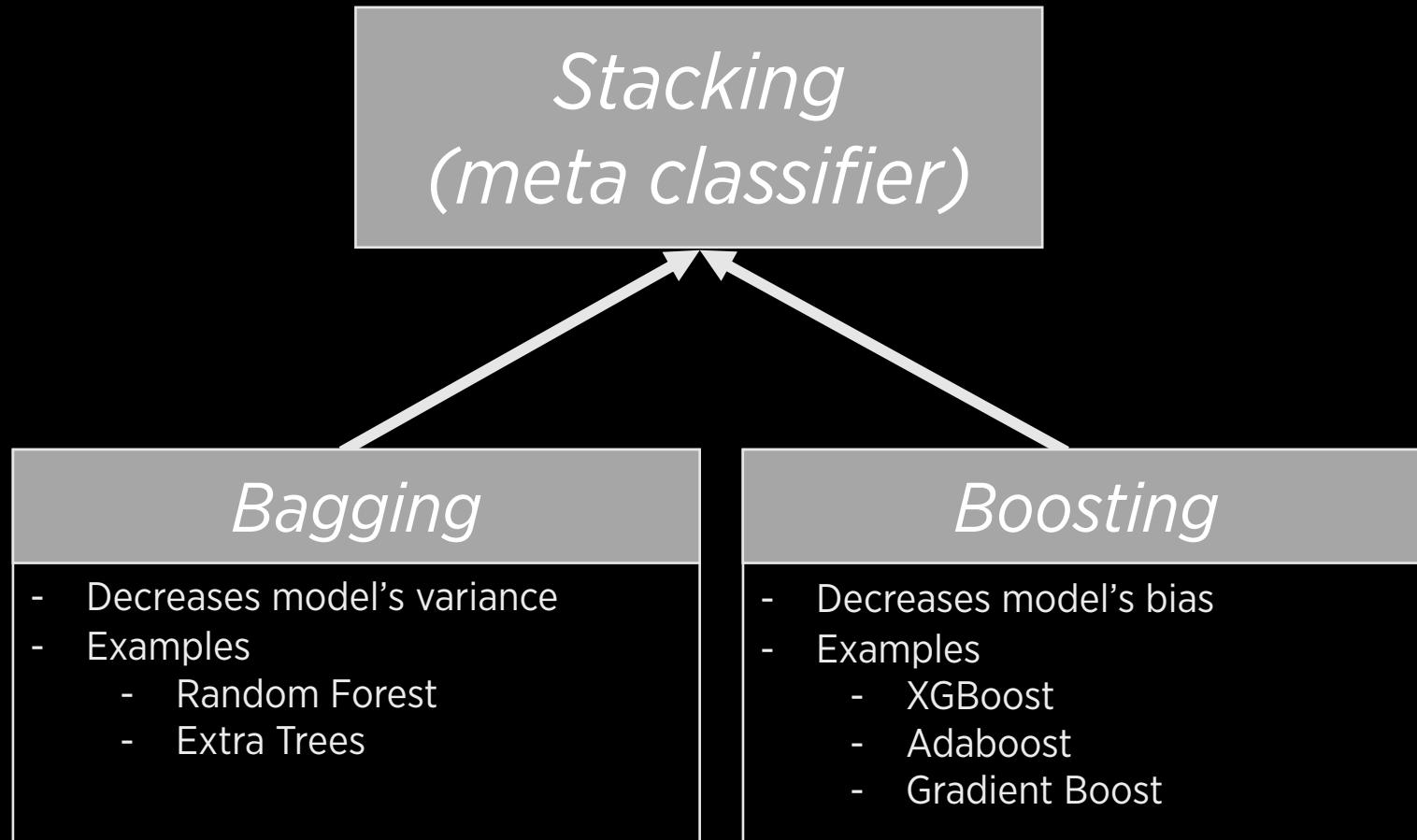
Bagging

- Decreases model's variance
- Examples
 - Random Forest
 - Extra Trees

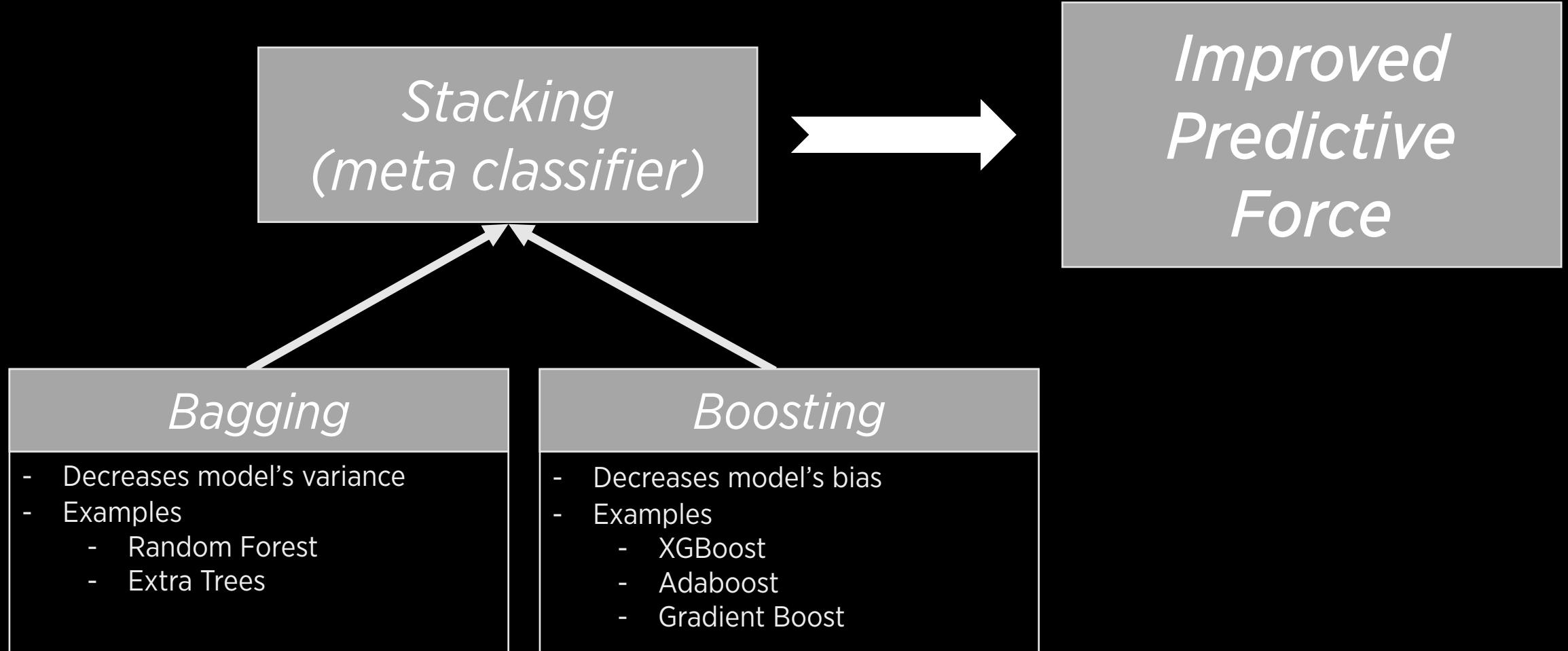
Boosting

- Decreases model's bias
- Examples
 - XGBoost
 - Adaboost
 - Gradient Boost

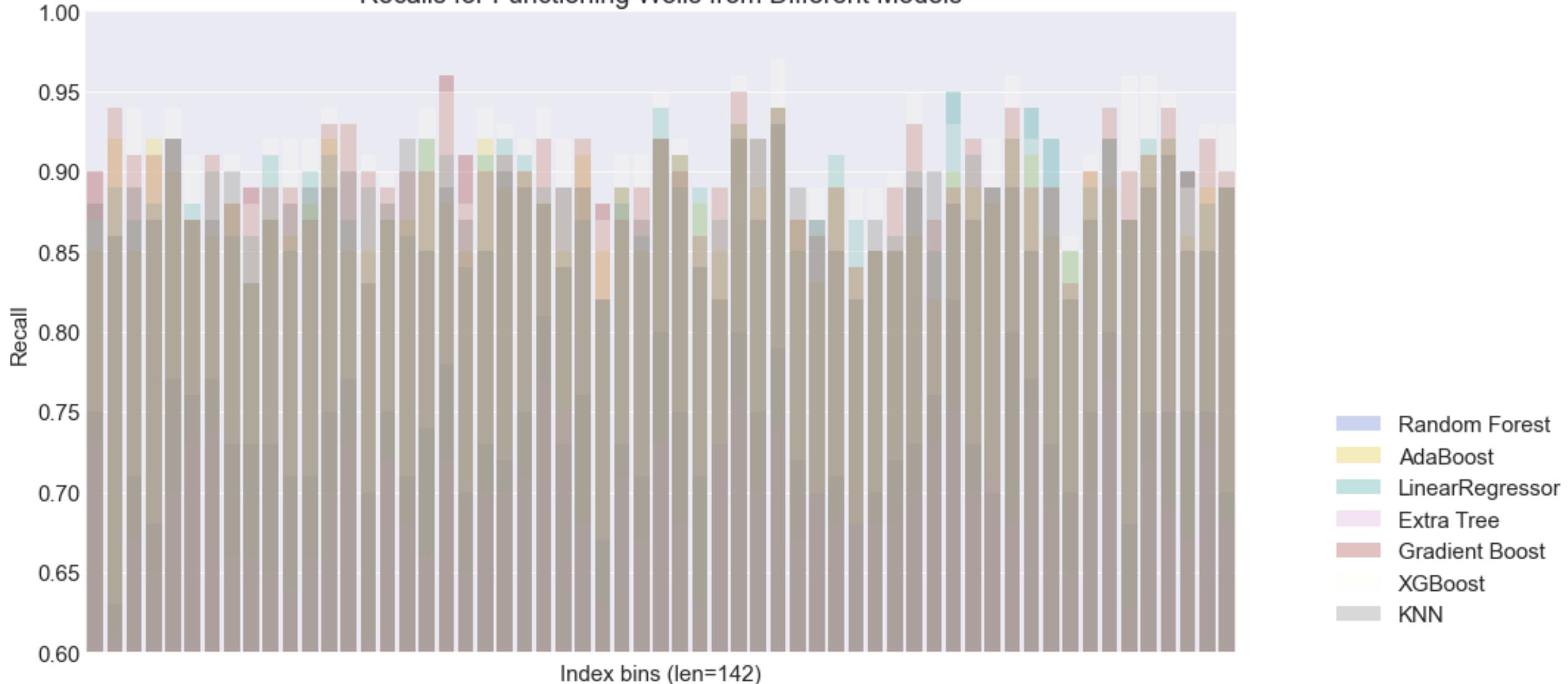
Out Plan



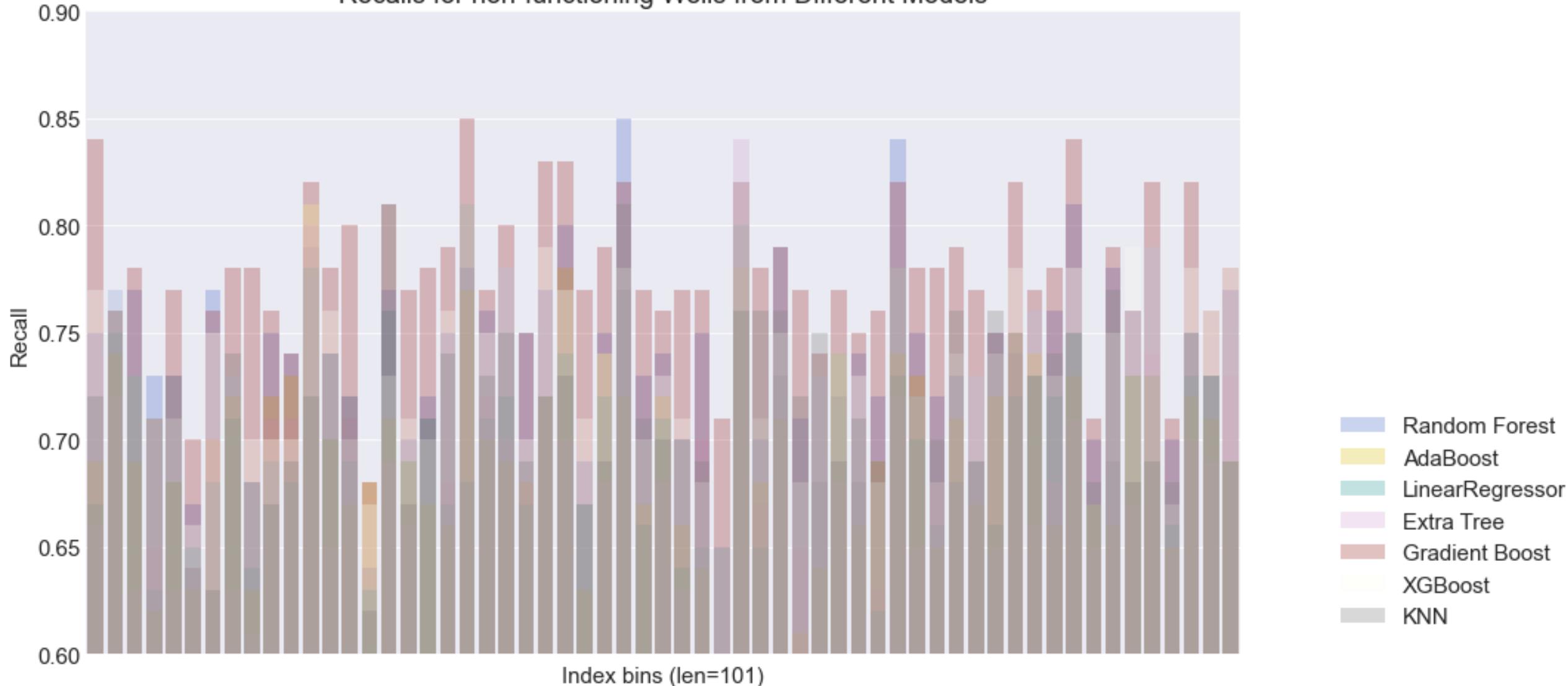
Out Plan



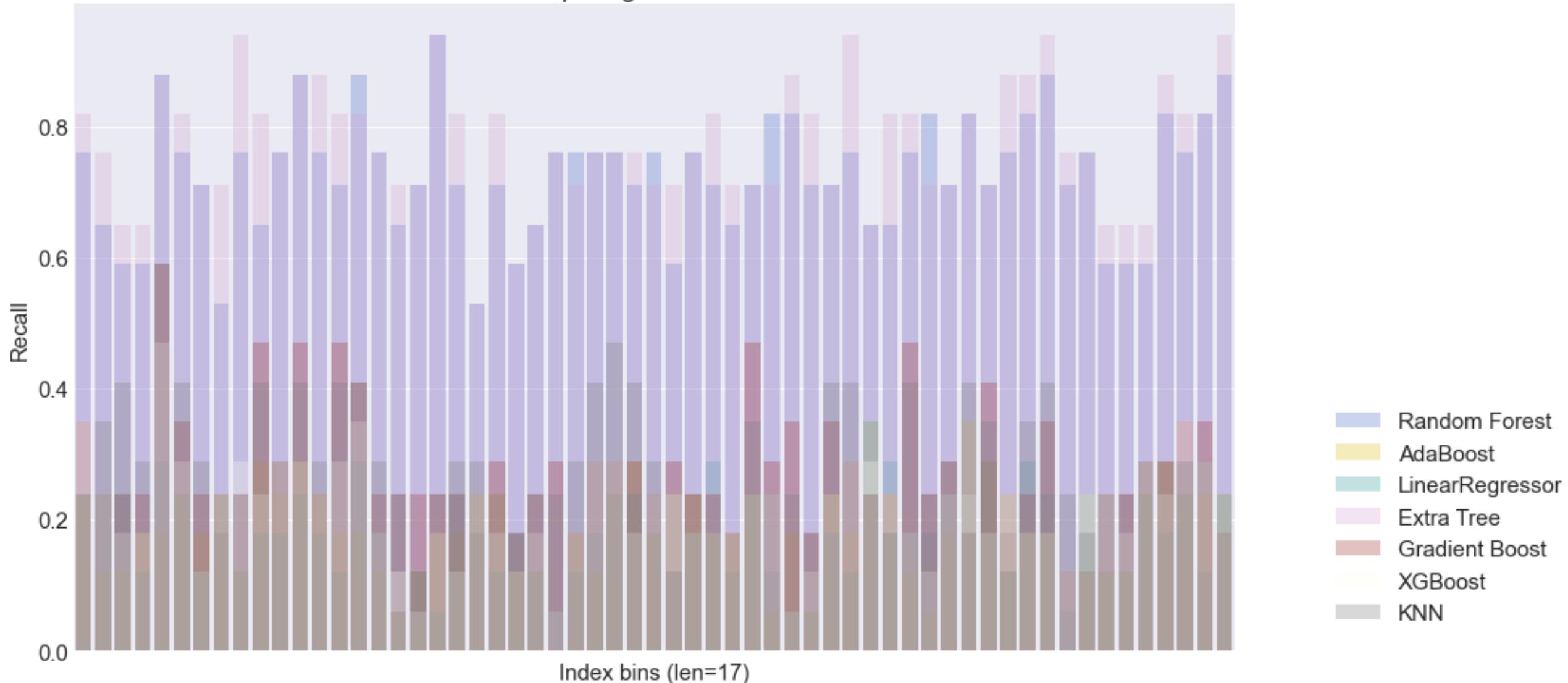
Recalls for Functioning Wells from Different Models



Recalls for non-functioning Wells from Different Models

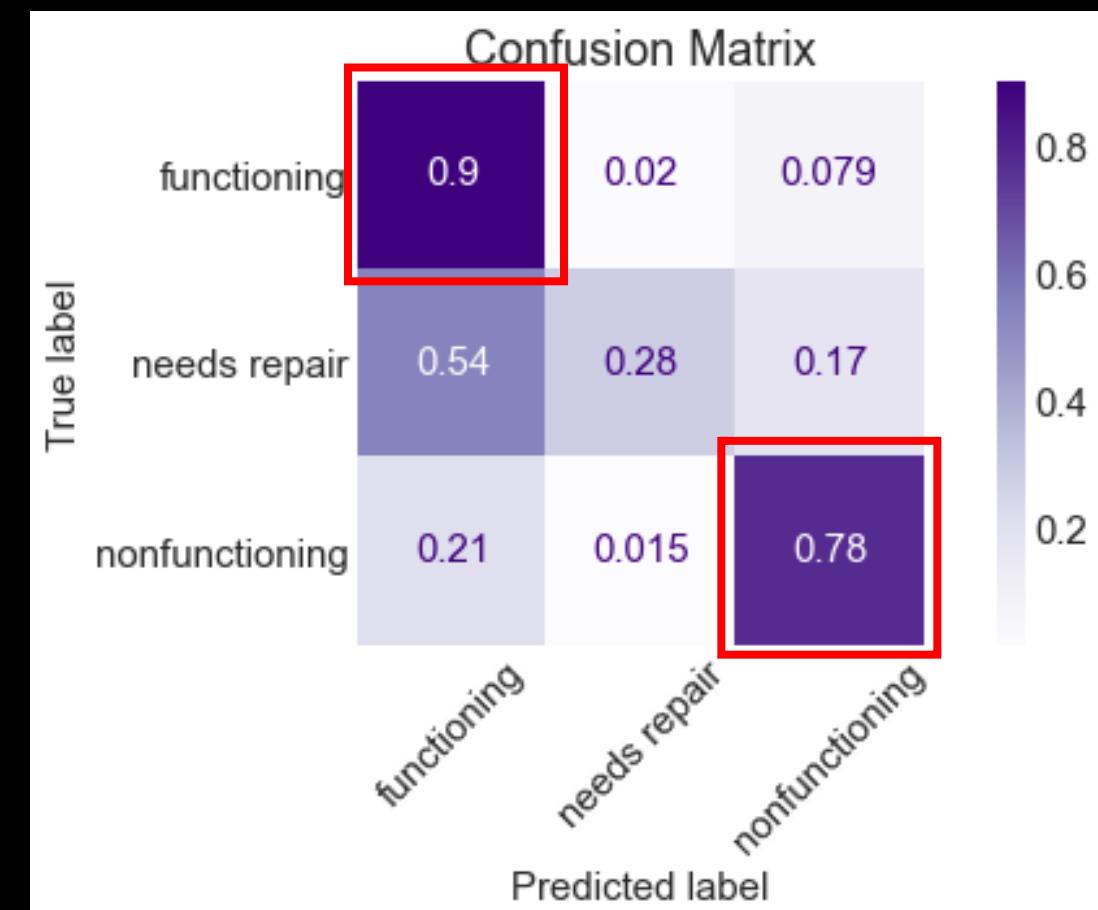


Recalls for need repairing Wells from Different Models



Best Accuracy Model: Adaboost

[i] CLASSIFICATION REPORT				
Train Accuracy : 0.934				
Test Accuracy : 0.8138				
Train AUC : 0.9501				
Test AUC : 0.8487				
CV score (n=3) 0.8971				
	precision	recall	f1-score	support
functioning	0.81	0.90	0.85	8490
needs repair	0.53	0.30	0.38	1019
nonfunctioning	0.85	0.78	0.81	6040
accuracy			0.81	15549
macro avg	0.73	0.66	0.68	15549
weighted avg	0.81	0.81	0.81	15549



[1] Possibly overfit [2] Accuracy = 81% [3] Test Accuracy = 79% [4] Low Recall for repair

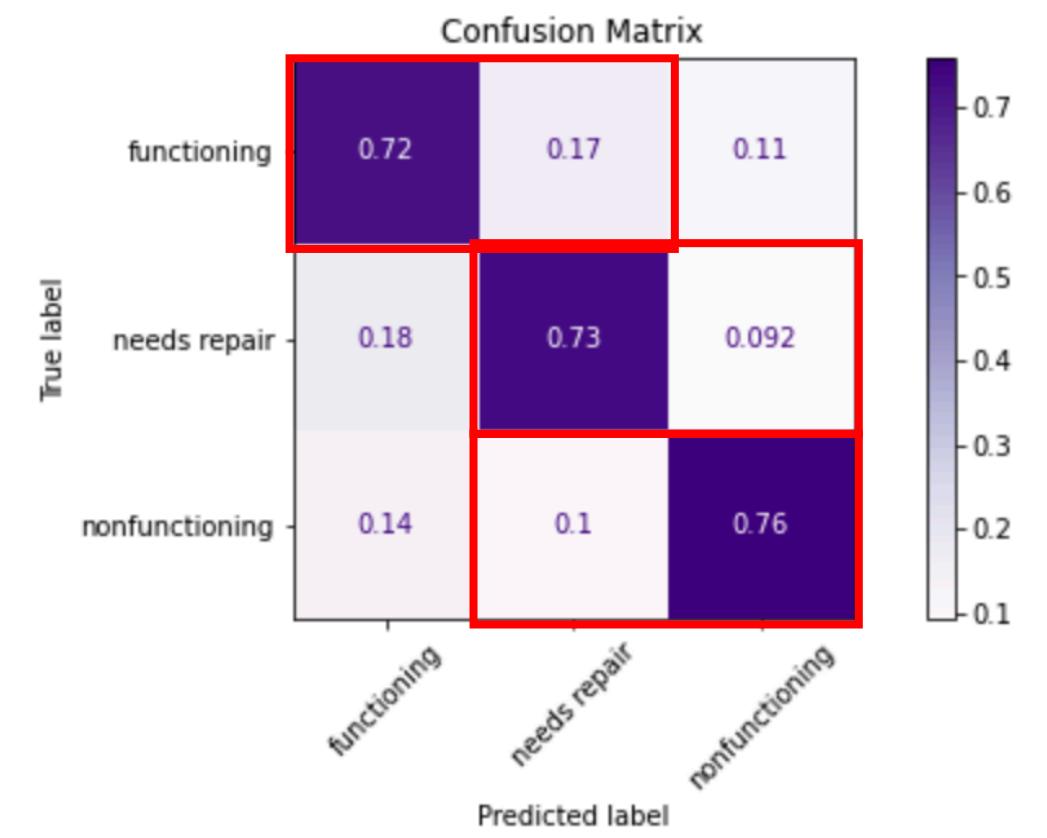
[5] 1st Layer Models: All the first layer models were used

Best Recall Model: Random Forest

[i] CLASSIFICATION REPORT

```
Train Accuracy : 0.7817  
Test Accuracy : 0.7343  
Train AUC : 0.9036  
Test AUC : 0.8624  
CV score (n=3) 0.7664
```

	precision	recall	f1-score	support
functioning	0.86	0.72	0.78	8490
needs repair	0.26	0.73	0.39	1019
nonfunctioning	0.81	0.76	0.78	6040
accuracy			0.73	15549
macro avg	0.64	0.73	0.65	15549
weighted avg	0.80	0.73	0.76	15549



[1] Recall = 73%

[2] Accuracy = 73%

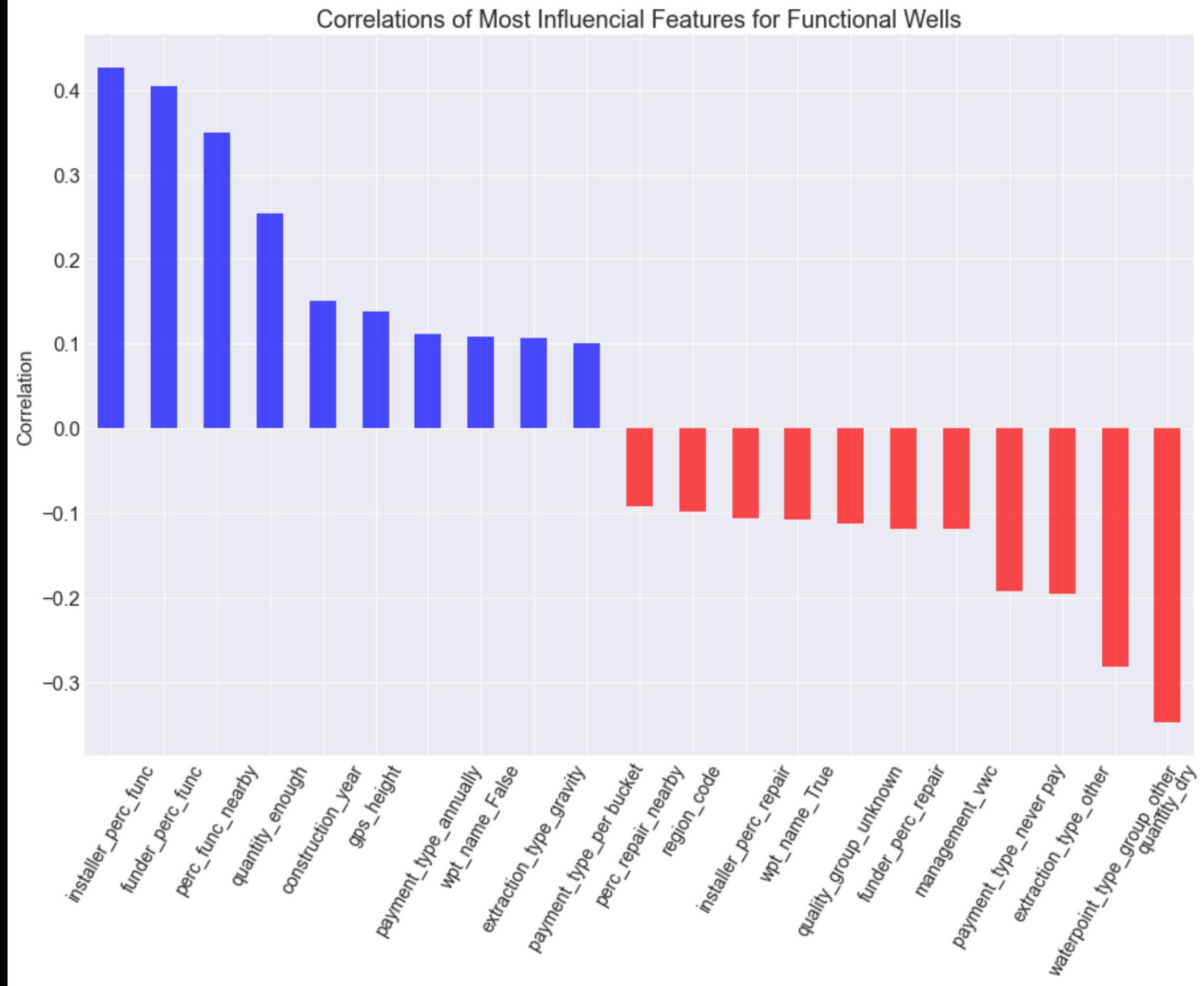
[3] Test Accuracy = ? %

[4] Effective Recalls = 89%, 82%, 86%

[5] 1st Layer Models: XGBoost, Random Forest, Extra Trees, KNN

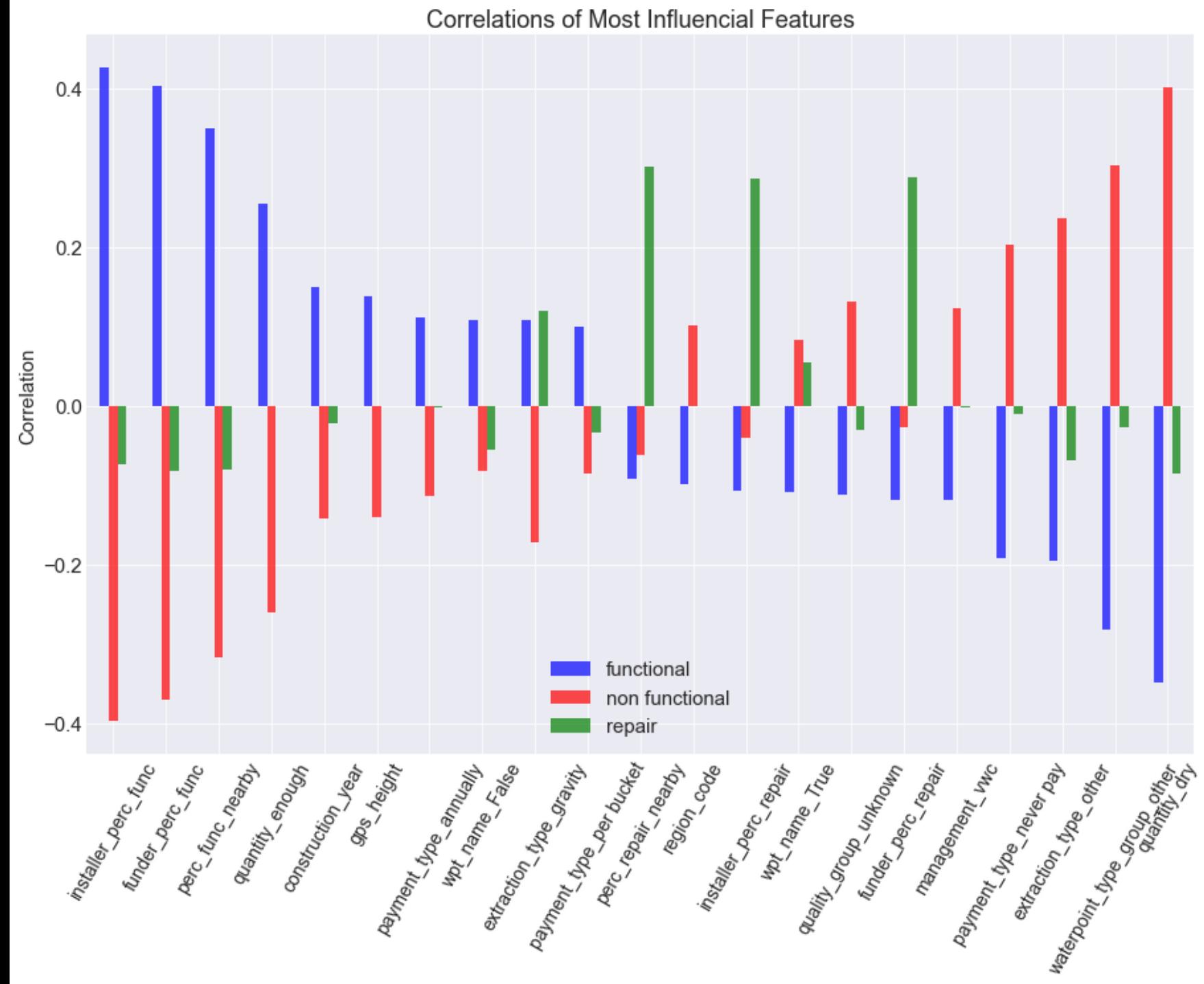
Interpreting the Model

- Most positive:
 - Installer
 - Funder
 - Neighbor
 - Payment
 - Extractor type
- Most Negative:
 - Quantity
 - Extractor type
 - Management
 - Neighbor



Interpreting the Model

- Expected
Functioning and non-functioning have opposite correlation
- New Finding
Repair is all over the places which makes the prediction extra tough



Conclusion

- Our final models can
 - Predict with 79% accuracy (Adaboost - 7 1st layers)
 - Predict with 85% recall (Random Forest - 4 1st layers)
- Important features
 - Positive
 - Installer, funder, neighbor- high functioning percentage
 - Payment - existence
 - Extractor type - gravity
 - Negative
 - Quantity - dry
 - Extractor type – other than gravity
 - Management – VWC (village water committee)
 - Payment – Lack of payment

Future Studies

- Further hyperparameter tuning can be done for each model used
- Different combinations of ensemble models can be tested
- Different methods of dealing with class imbalance can be used
 - under sampling
 - different class weights
- Research and estimate for the cost to best determine which model would actually benefit the situation

Thank you for listening

Appendix