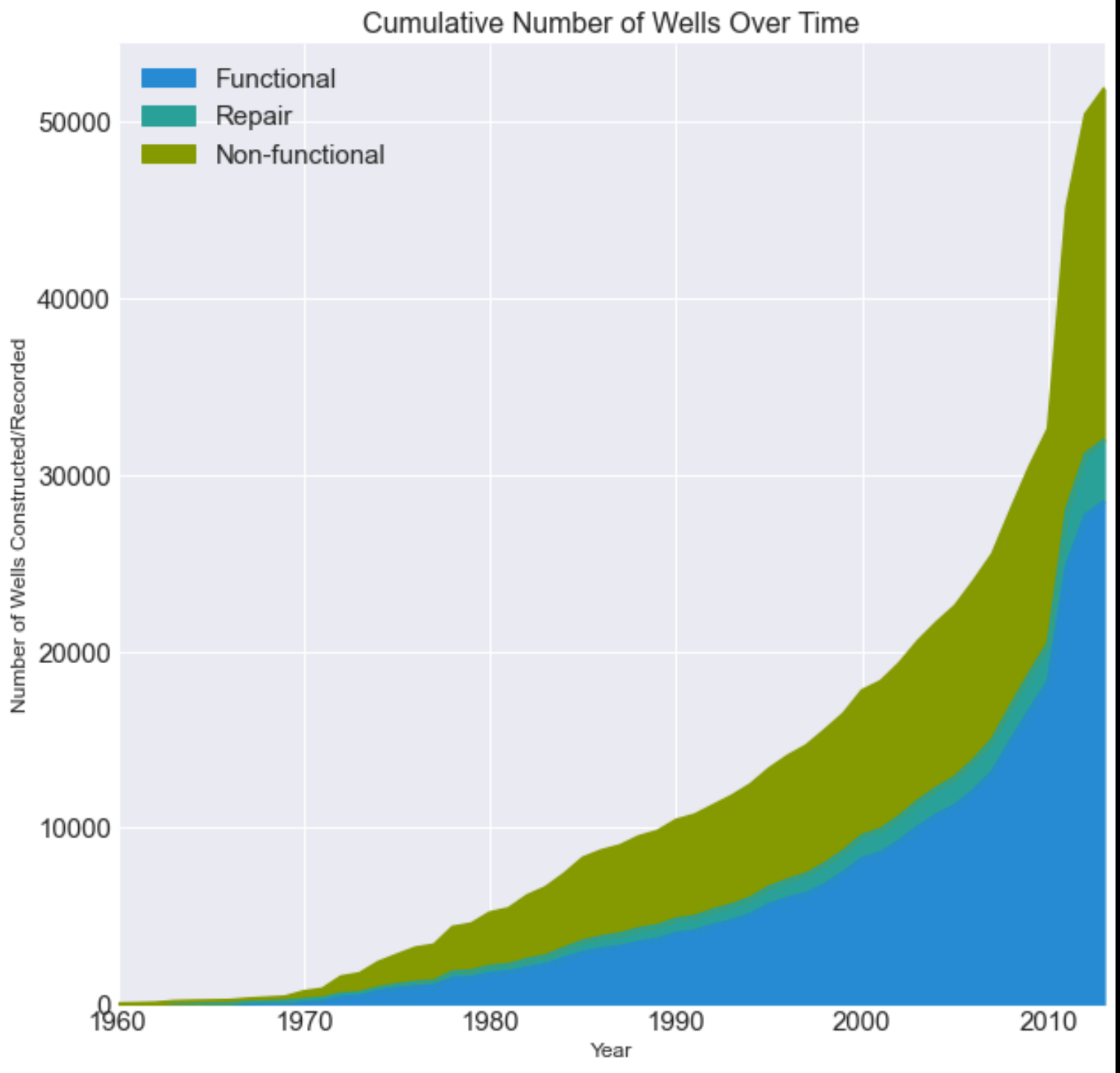
How well are the wells?

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

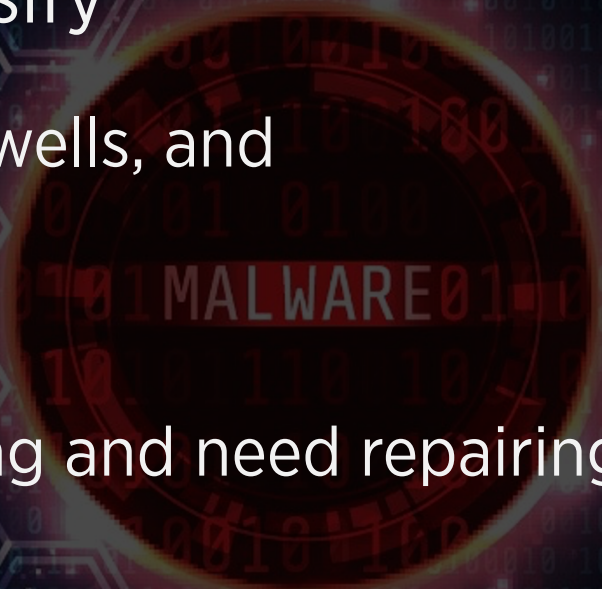


Tanzanian Water Crisis Facts

- 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

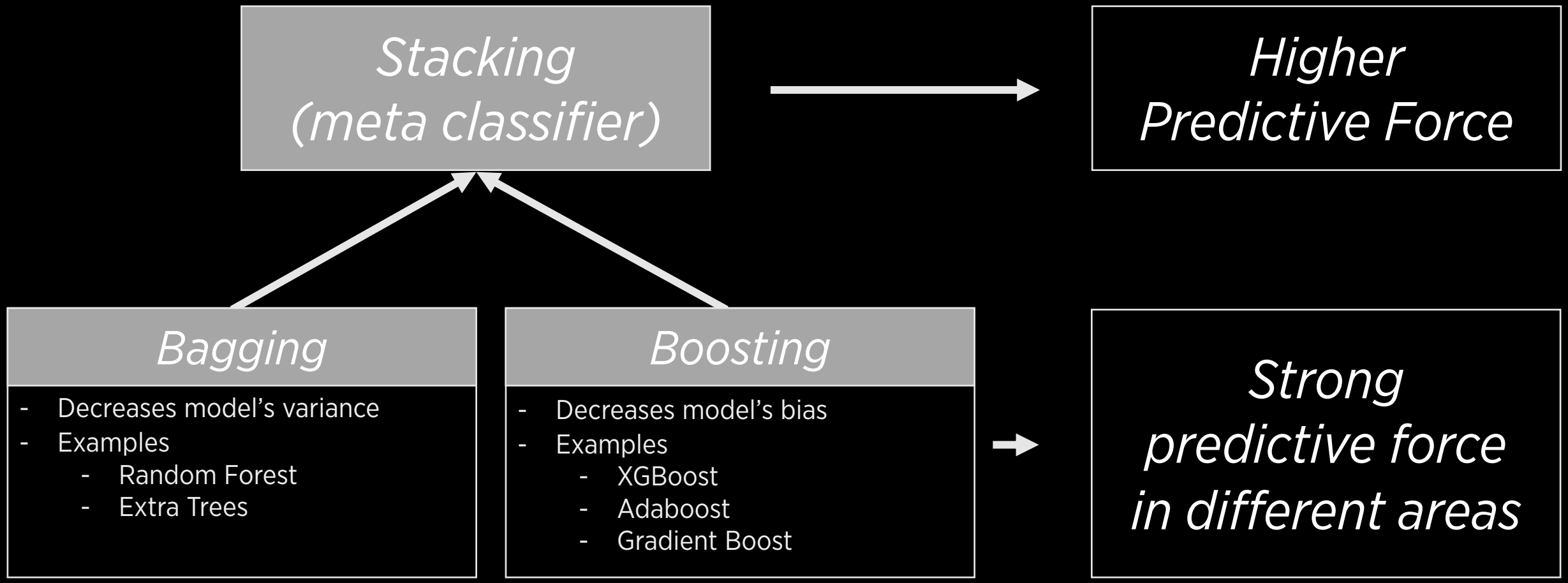
- Construct a model that can classify
  [1] functioning wells,
  [2] functioning but need repairing wells, and
  [3] non-functioning wells.

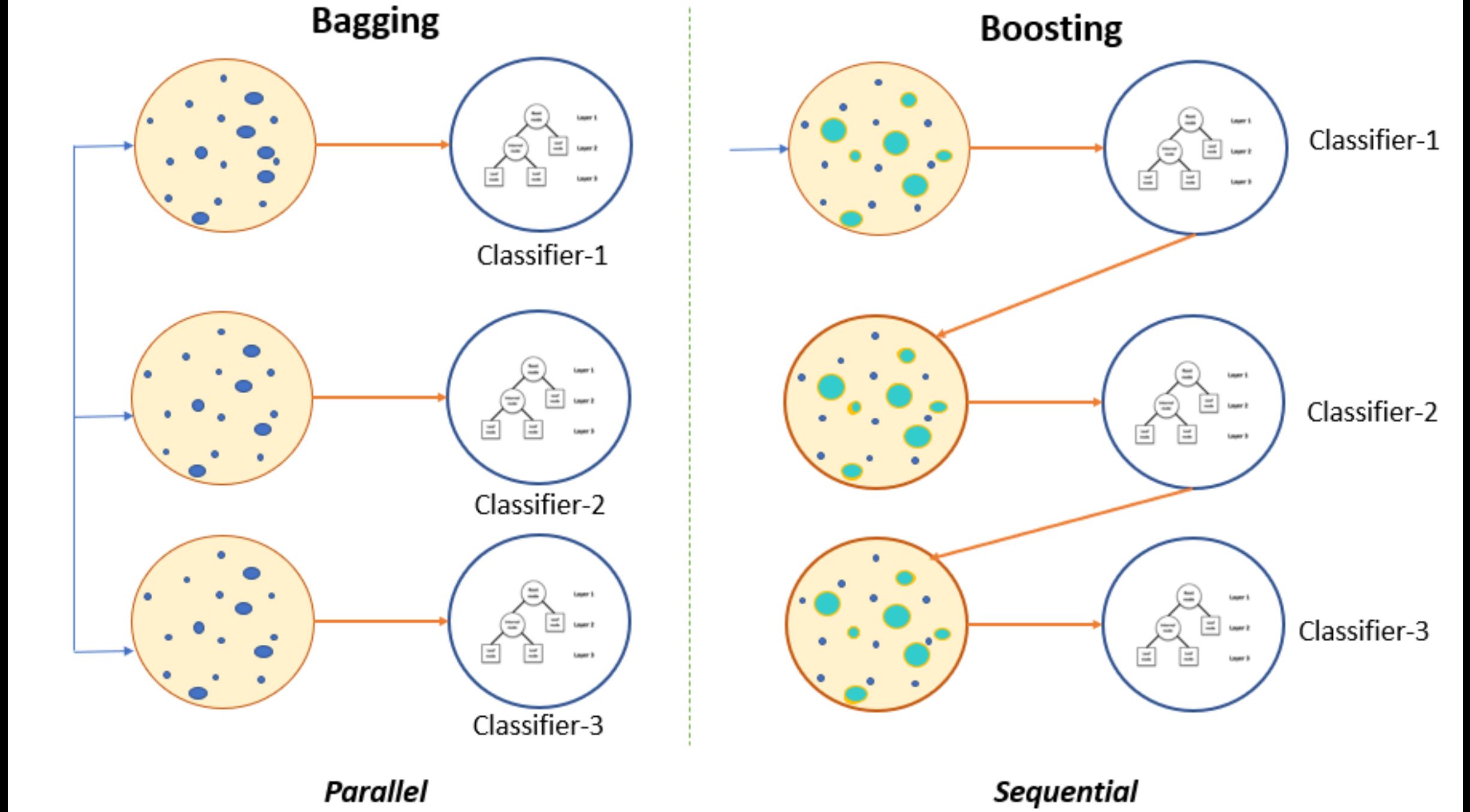


- Top Priorities
  → Correctly identify non-functioning and need repairing wells (high recalls)
  → Overall accuracy

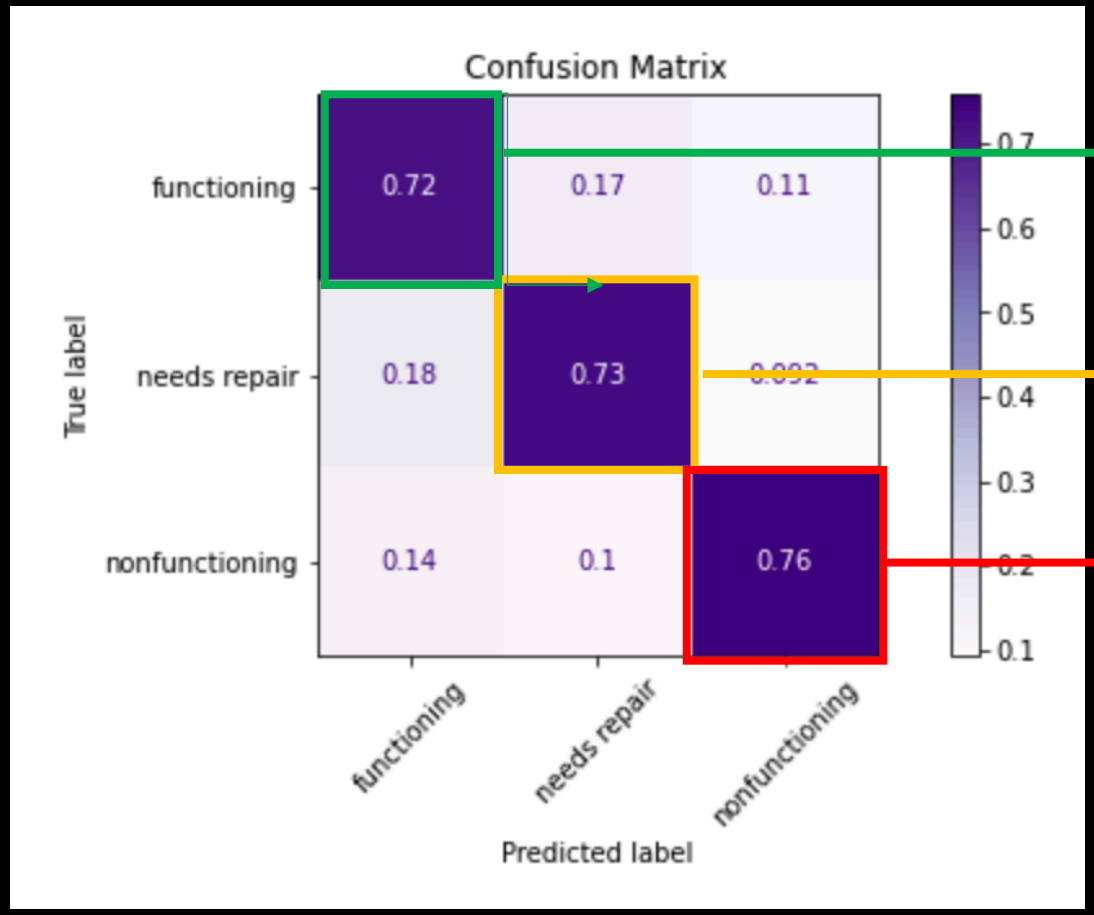
- Provide features that affect functionalities of water well

Modeling Plan




Bagging
Classifier-1
Classifier-2
Classifier-3
Parallel
Boosting
Classifier-1
Classifier-2
Classifier-3
Sequential

Best Model: Recalls



72% of functioning wells were correctly identified.

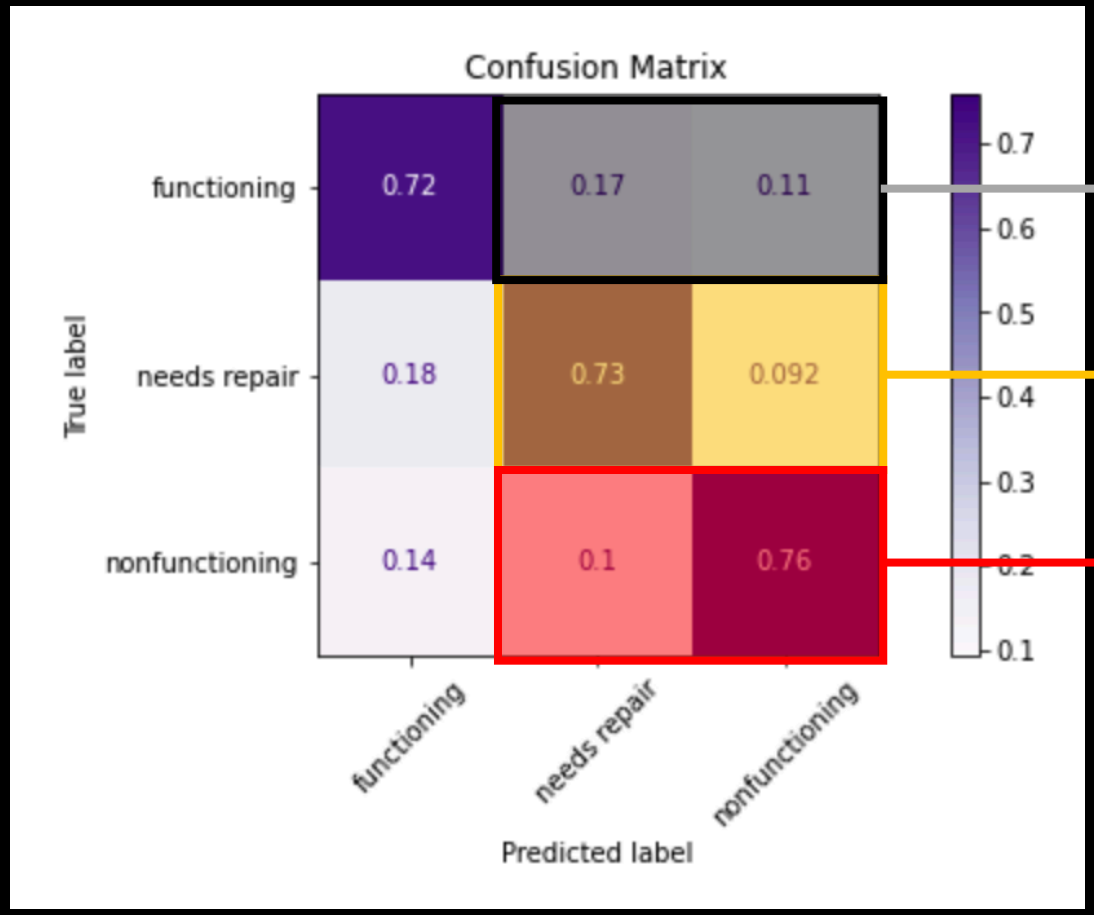
73% of need-repairing wells were correctly identified.

76% of non-functioning wells were correctly identified.

[1] Recall = 73% [2] Accuracy = 76.6% [3] Effective Recalls = 72%, 82%, 86%

[5] 1st Layer Models: XGBoost, Random Forest, Extra Trees, KNN; 2nd Layer: Random Forest

Best Model: Effective Recalls



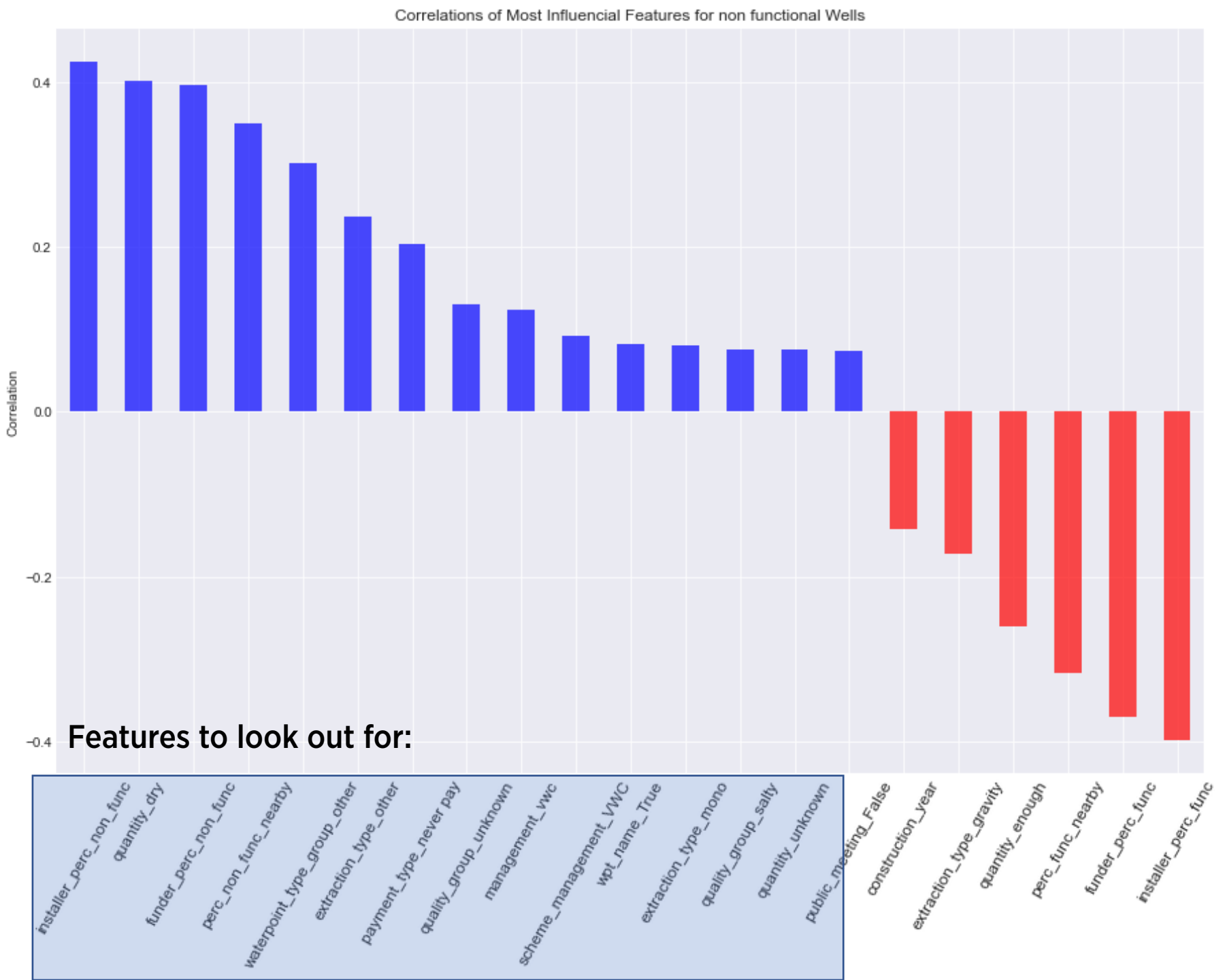
Out of all the well-functioning wells, 28% will be identified as in a need of functioning
- wasteful cost

Out of all the wells that need repairing, 82.2% them would be flagged to be in a need of repairing.

Out of all the wells that are not functioning, 86% them would be flagged to be in a need of repairing.

Important Features

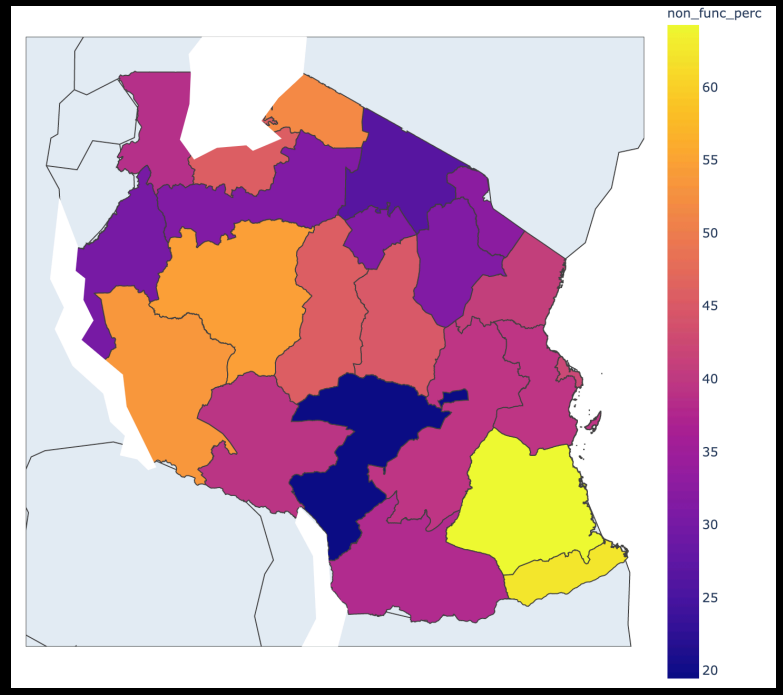
- Installer
- Funder
- Neighboring
- Water Type
- Payment
- Management
- Extraction type
- Waterpoint name



Action 1: Watch out for certain regions!

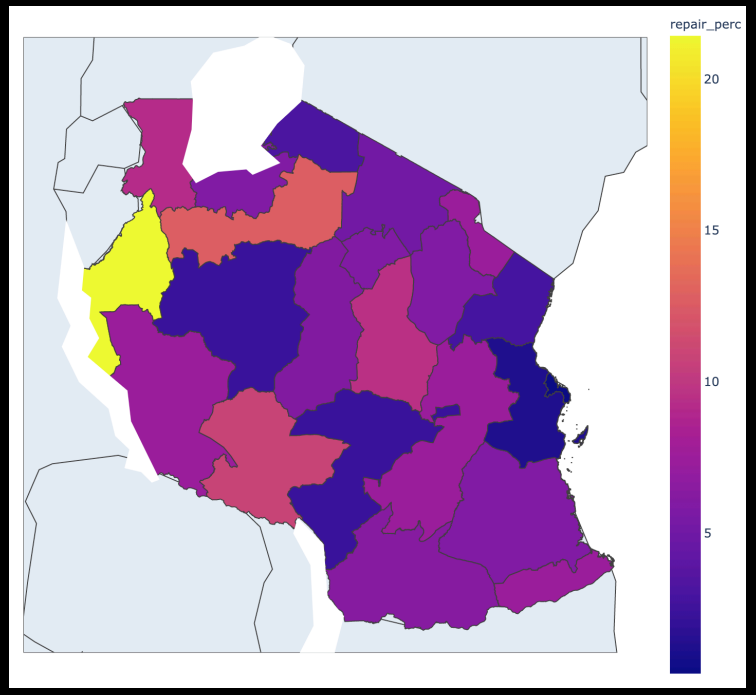
- Regions with most non-functioning wells:
  - Lindi (64.23%)
  - Mtwara (62.43%)
  - Tabora (54.42%)
- Regions with most need repairing wells:
  - Kigoma (21.41%)
  - Shinyanga (12.75%)

Percent Non-Functioning Wells



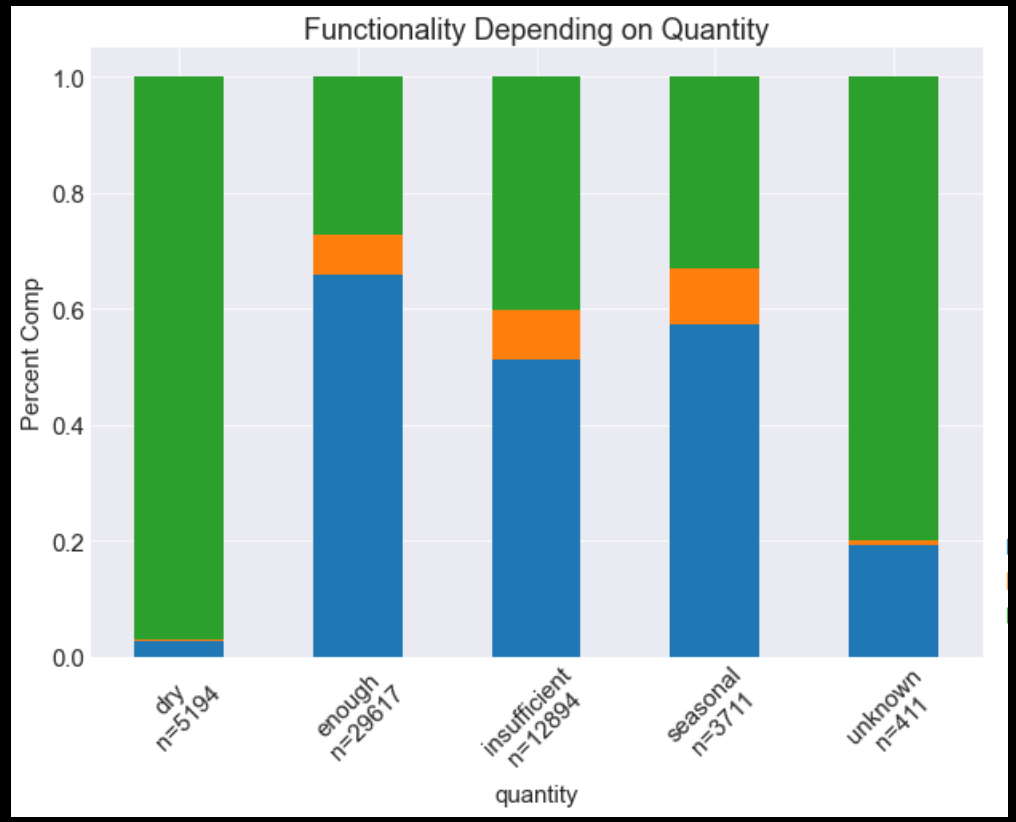
Region: Lindi
N = 1546
Percentage: 64.23%

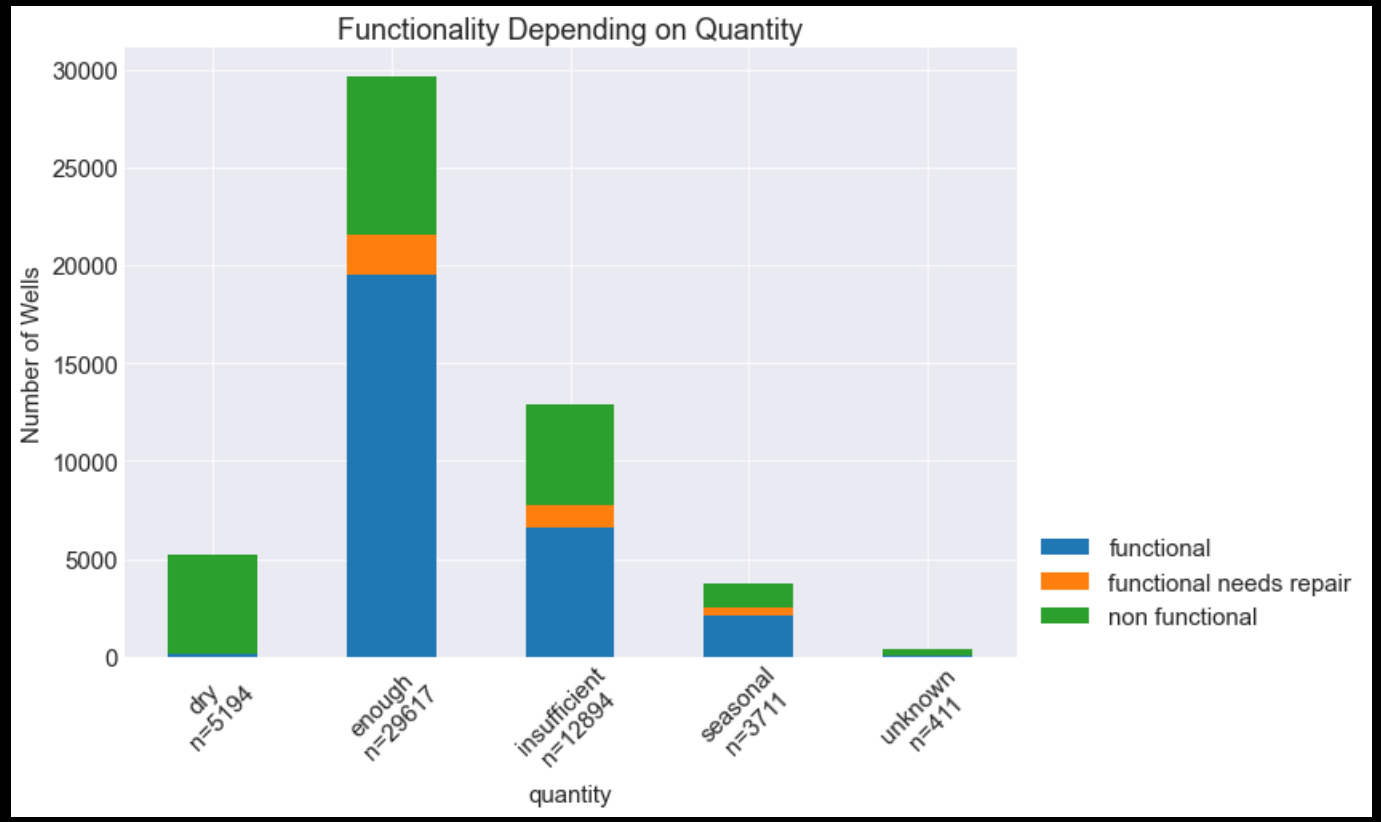
Percent Need-Repairing Wells



Region: Kigoma
N = 2816
Percentage: 21.41 %

Action 2: Watch out for ”Dry” and “unknown”

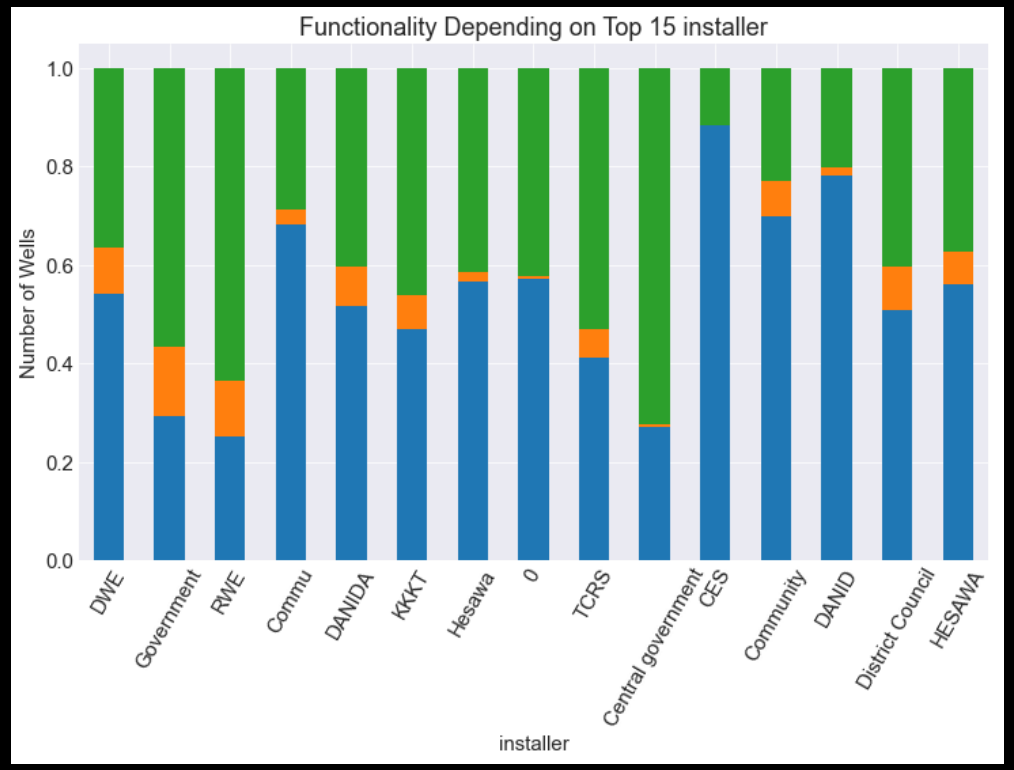


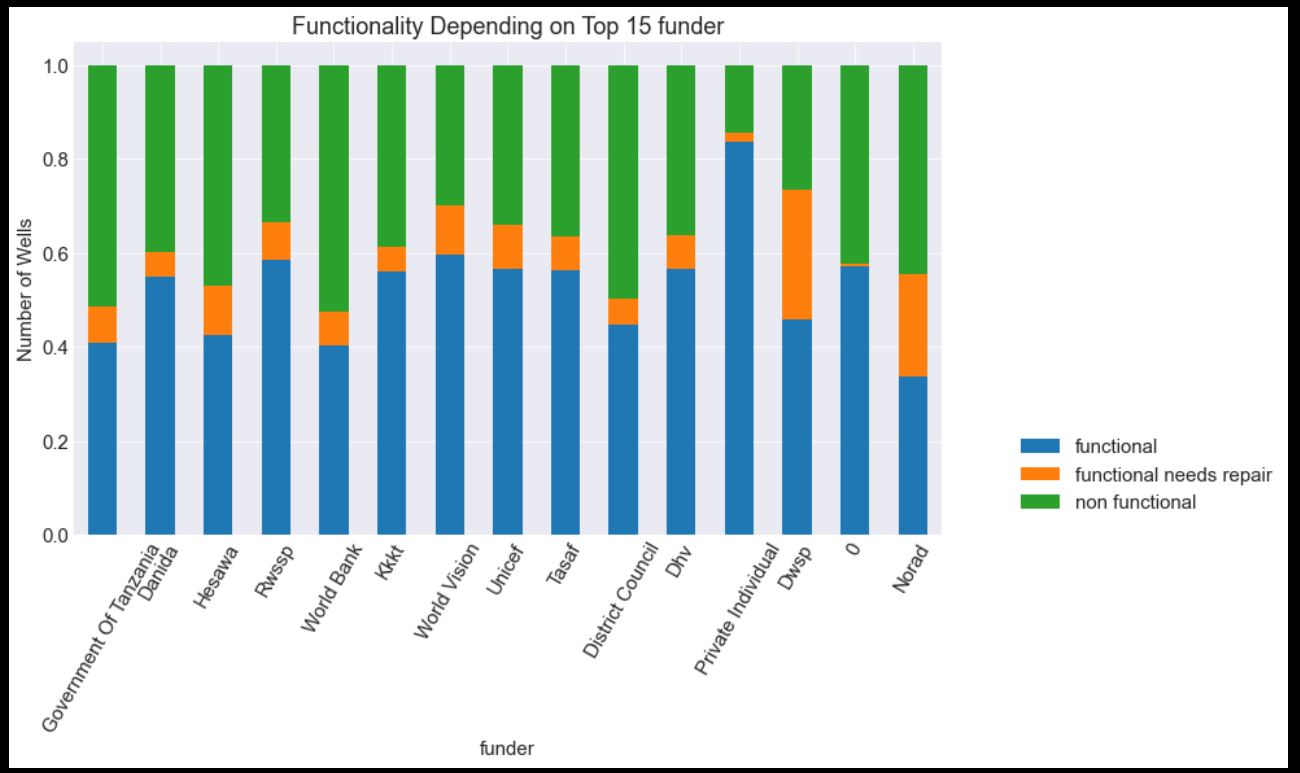


More than 90% non-functional wells for ”dry quantity” wells.

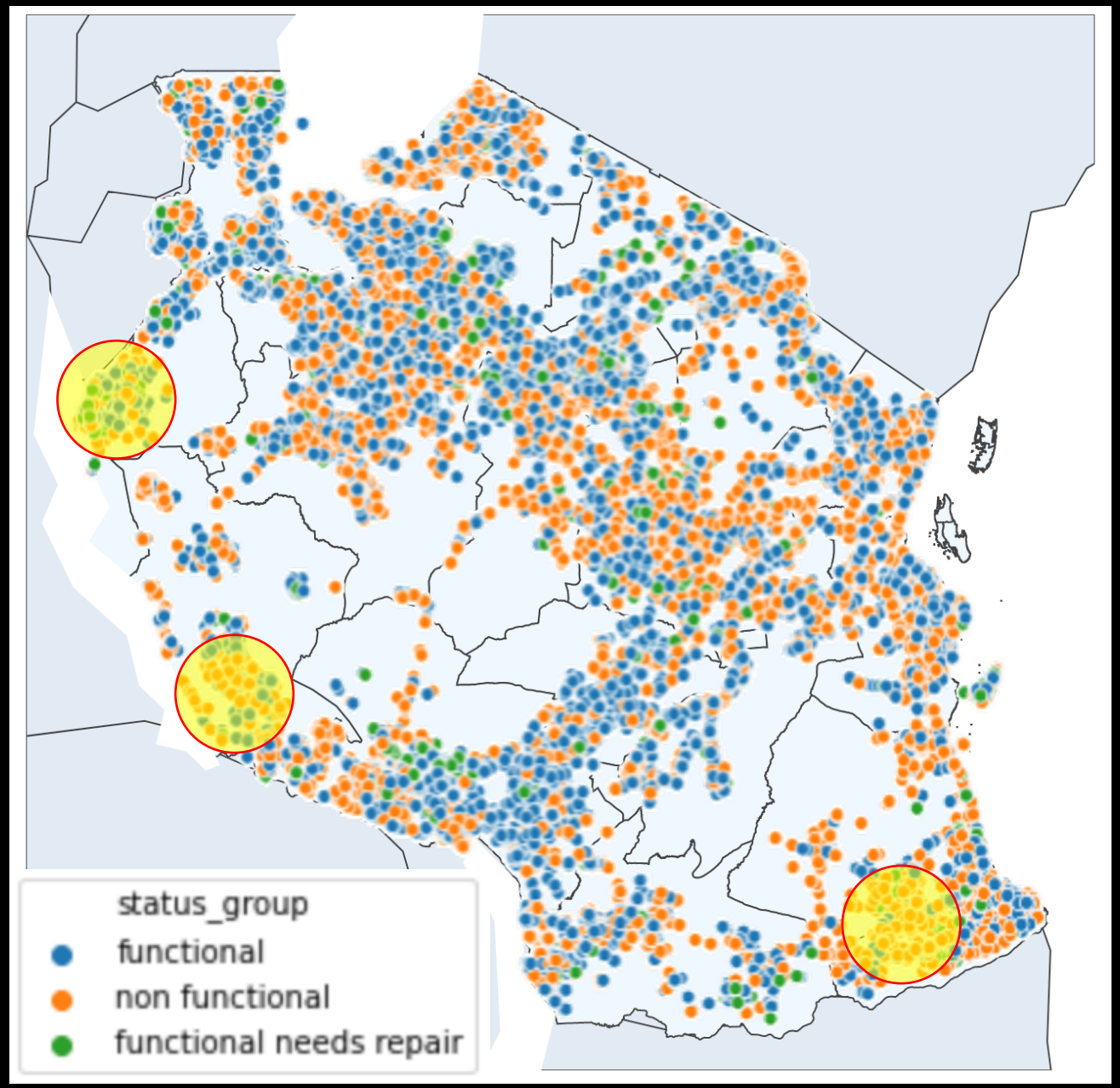
More than 80% non-functional wells for ”unknown quantity” wells.

Action 3: Funders and Installers





- Watch out for government and RWE installed wells
- Watch out for government and Norad funded wells



Action 4: Neighboring

Wells that are near wells that in need of maintenance should raise a flag!

Conclusion

- Our final models can
  - Predict with ~80% recall (Random Forest – 4 1st layers)
- Features correlated with water wells in need of maintenance

  [1] Installer: Government and RWE installers

  [2] Water quantity: Dry and unknown quantities

  [3] Funder: Government and Norad funders

  [4] Extraction type: other than gravity type

  [5] Payment: no payment for usage

  [6] Management: VWC management

  [7] Neighboring: highly dense non-functioning wells area

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
- Cost function for

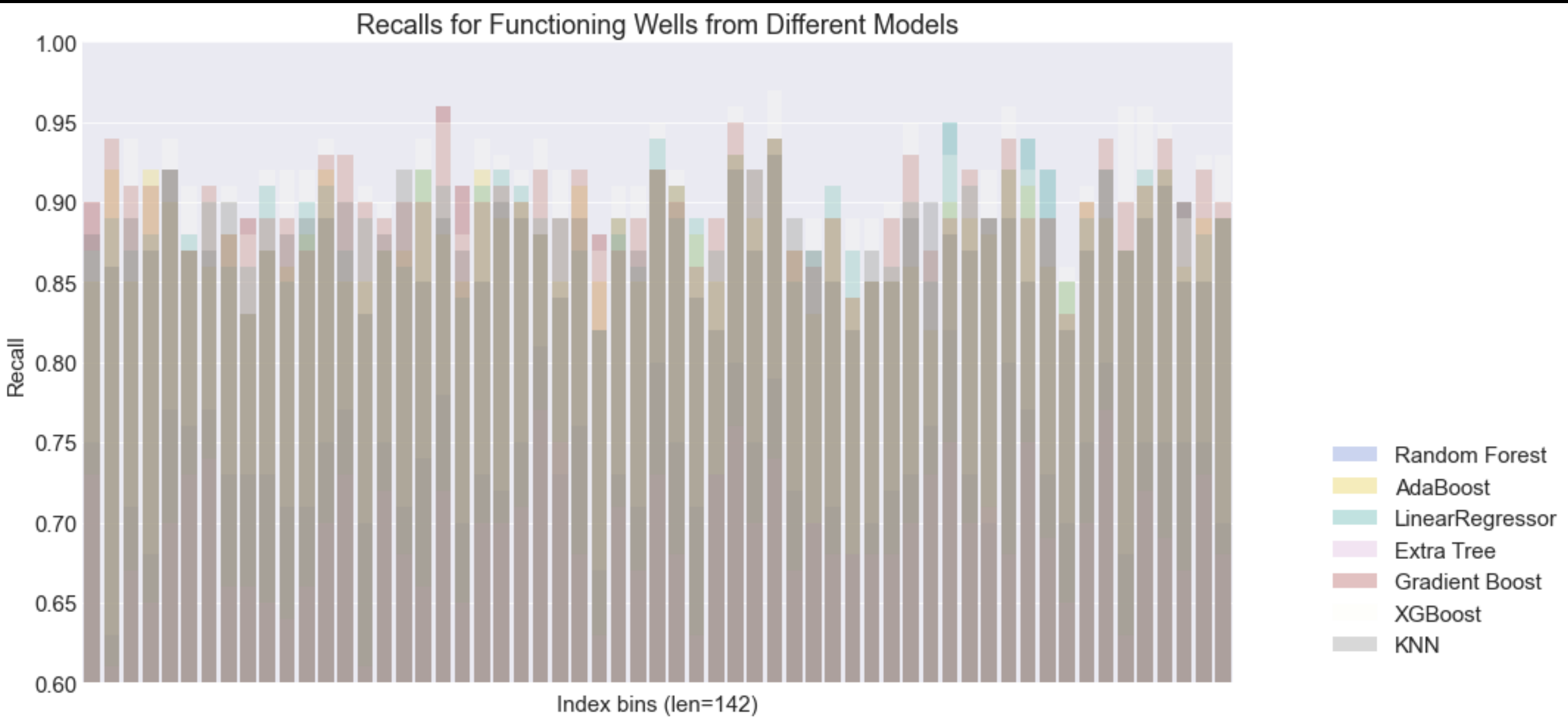
  [1] wrongly identifying functioning wells as need repairing wells and

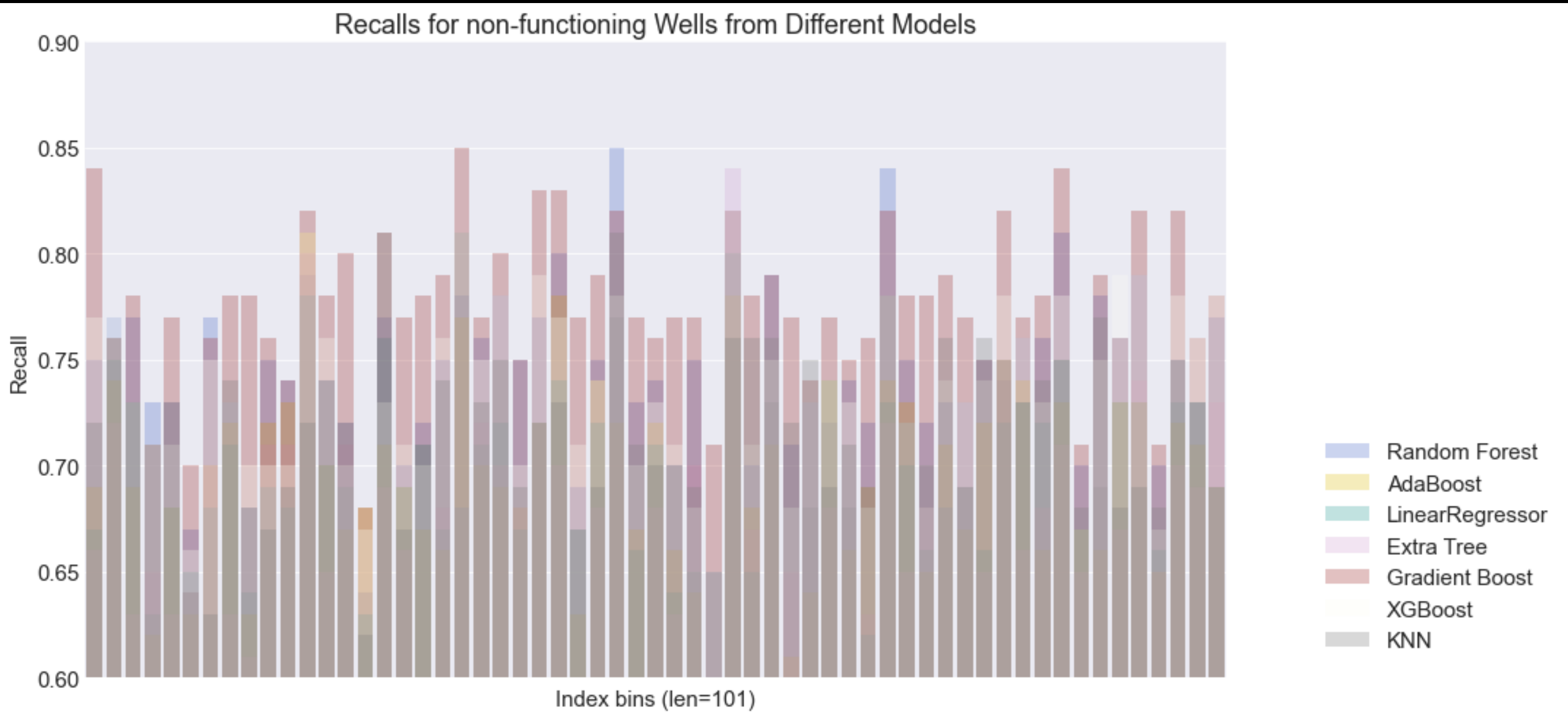
  [2] not being able to identify wells in need of repairing

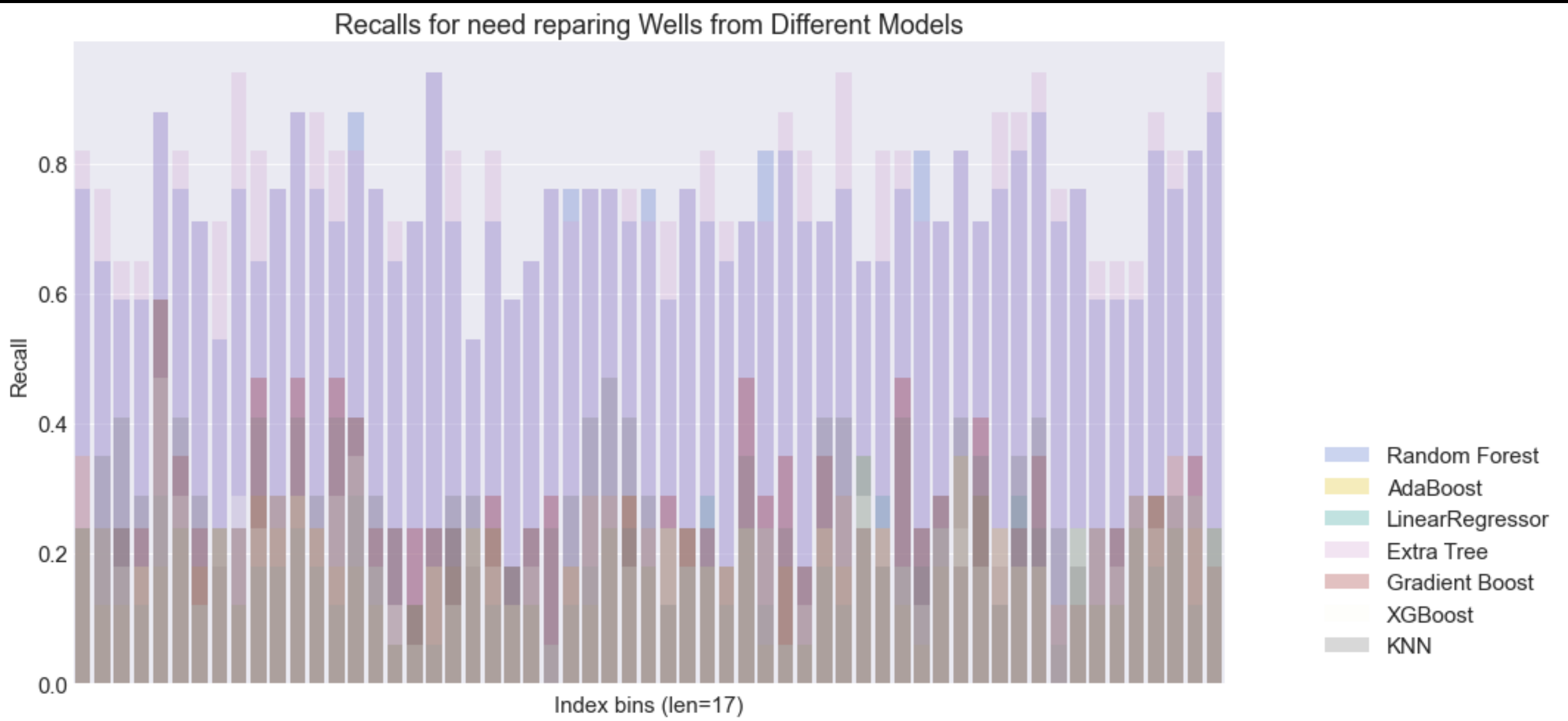
  would improve the model to minimize both economical and sociological cost.

Thank you for listening

Appendix


Recalls for Functioning Wells from Different Models
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
Recall
Index bins (len=142)
Random Forest
AdaBoost
LinearRegressor
Extra Tree
Gradient Boost
XGBoost
KNN


Recalls for non-functioning Wells from Different Models
0.90
0.85
0.80
0.75
0.70
0.65
0.60
Recall
Index bins (len=101)
Random Forest
AdaBoost
LinearRegressor
Extra Tree
Gradient Boost
XGBoost
KNN


Recalls for need reparing Wells from Different Models
Recall
0.0
0.2
0.4
0.6
0.8
Index bins (len=17)
Random Forest
AdaBoost
LinearRegressor
Extra Tree
Gradient Boost
XGBoost
KNN

Best Accuracy Model: Adaboost

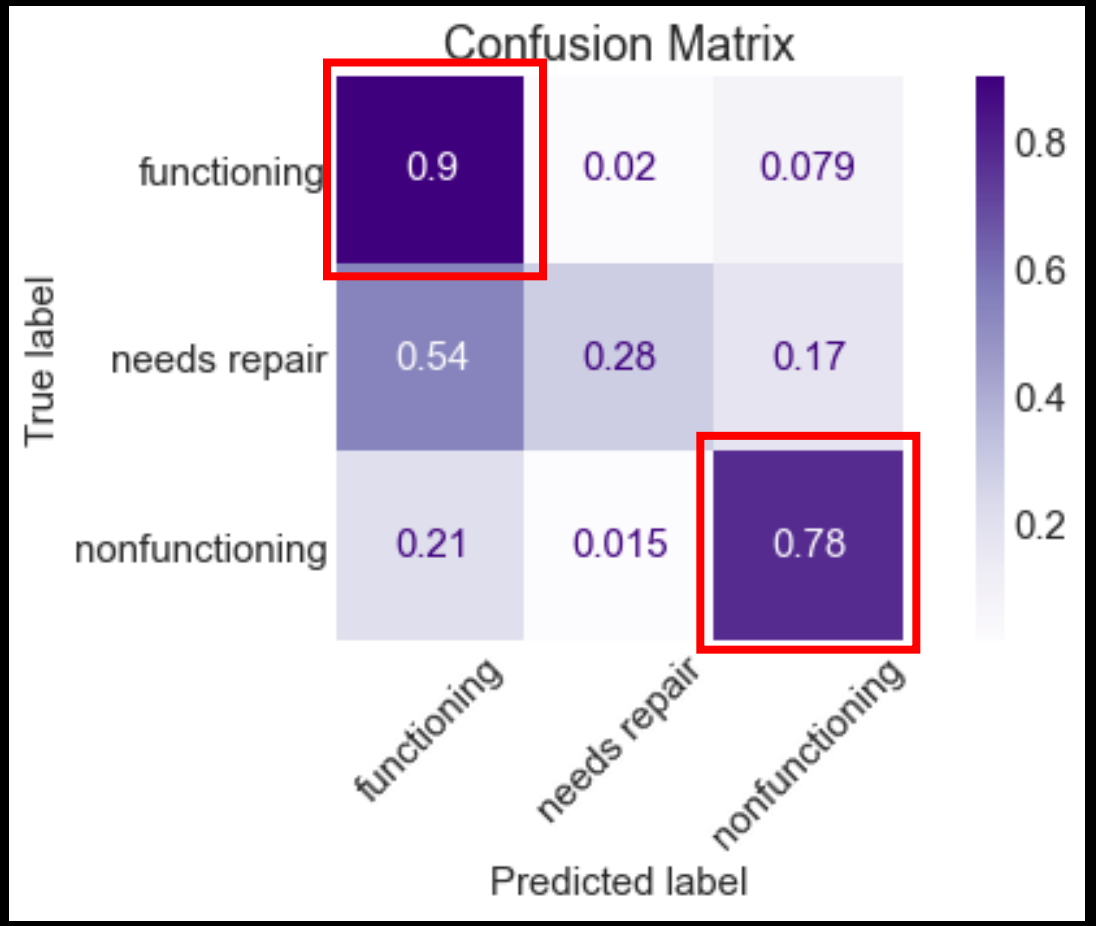
```
----------------------------------------------------------
[i] CLASSIFICATION REPORT
----------------------------------------------------------
Train Accuracy :  0.934
Test Accuracy :  0.8138
Train AUC :  0.9501
Test AUC :  0.8487
CV score (n=3) 0.8971
----------------------------------------------------------

                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

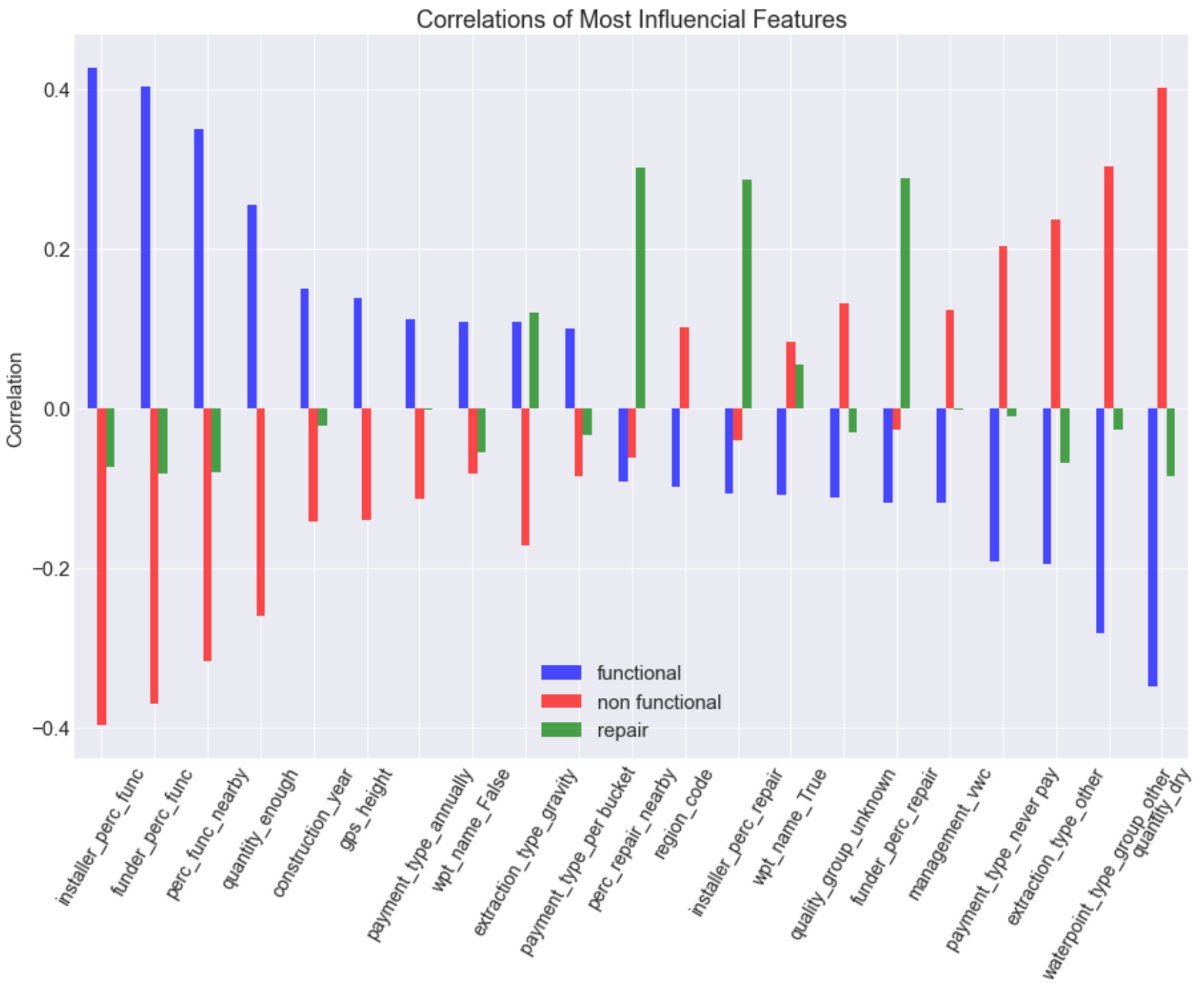
Interpreting the Model

- Expected

Functioning and non-functioning have opposite correlation

- New Finding

Repair is all over the places which makes the prediction extra tough



Region Lists for non-functioning and repair

	region	non_func_perc
9	Lindi	64.23
14	Mtwara	62.43
21	Tabora	54.42
17	Rukwa	53.43
11	Mara	51.96
15	Mwanza	45.68
20	Singida	45.58
1	Dodoma	44.66
25	Dar es Salaam	42.36
22	Tanga	40.75
13	Morogoro	39.62
16	Pwani	39.58
12	Mbeya	39.15
3	Kagera	38.75
18	Ruvuma	37.73
7	Kilimanjaro	32.36
10	Manyara	31.59
19	Shinyanga	31.27
6	Kigoma	30.18
0	Arusha	26.30
2	Iringa	19.46

	region	repair_perc
6	Kigoma	21.41
19	Shinyanga	12.75
12	Mbeya	10.86
1	Dodoma	9.50
3	Kagera	9.17
13	Morogoro	7.49
17	Rukwa	7.47
7	Kilimanjaro	7.35
14	Mtwara	7.28
18	Ruvuma	6.21