UPDATES

July 29, 2024

Notes

- dataset imbalanced, lacking detections
- new type of feature importance

Current dataset - imbalanced

- 391 tests
  - 27 parameters -> 74 TGUS values
  - 11 sites
- 46 detects
  - 11 parameters -> 28 TGUS values
  - 11 sites
- 345 nondetects
  - 25 parameters -> 54 TGUS values
  - 8 sites
- (345 nondetects) / (391 total) = 88%
- model can be poor and still highly accurate

Feature importances can be misleading

- built on the training set
- overfitting & high cardinality inflate FIs
- cardinality order
    - TGUS: 76
    - GUS: 27
    - detection limit: 9
    - drainage class: 5
    - aquifer vulnerability: 3

Current model

Results - binary classification accuracy (detected or nondetected)

- **All**: GUS, TGUS, drainage class, aquifer vulnerability, detection limit
- **GUS Focus**: GUS, drainage class, aquifer vulnerability, detection limit
- **TGUS Focus**: TGUS, drainage class, aquifer vulnerability, detection limit
- **Indices**: GUS, TGUS

	Avg. Train %	Avg. Validation %	Avg. Test %	Best Test %	Worst Test &
All	97.2	96.2	93.9	98.7	87.3
GUS Focus	95.8	95.4	92.1	98.7	86.1
TGUS Focus	96.8	96	93.7	100	88.6
Indices	97	95.9	93.1	98.7	84.8

- each set set statistically similar
- potentially some slight overfitting

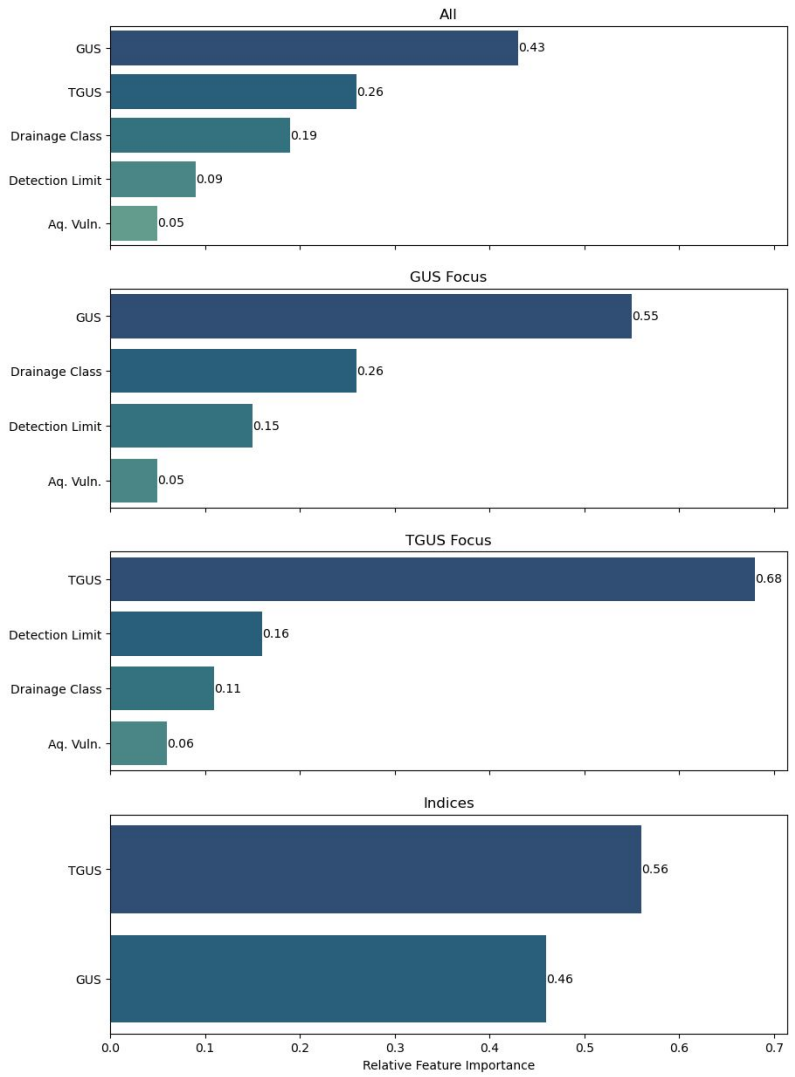
Model struggling with more with identifying detections

Model Average - True/False Positive/Negatives

	Test %	True Positive	True Negative	False Positive	False Negative
All	93.9	5.47	70.22	0.96	3.96
GUS Focus	92.1	5.55	68.69	1.84	4.53
TGUS Focus	93.7	4.78	70.8	1.06	3.98
Indices	93.1	5.14	69.92	1.37	4.18

Feature Importances (FI)

- possible overfitting/inflation of values
- permutation importance can help
- still shows correlation regardless



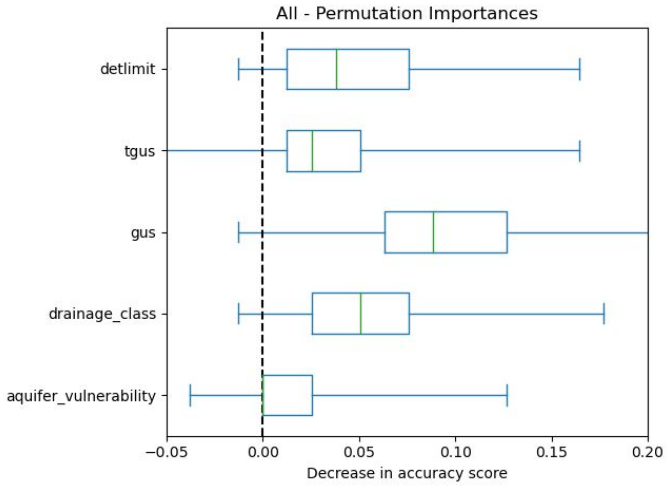
Permutation importance (PI)

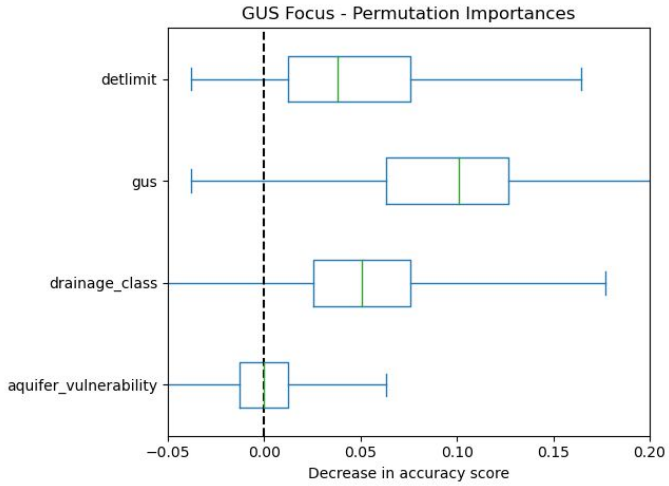
1. score model on held out test set
2. shuffle feature column and retest model
3. record drop in accuracy
4. repeat many times

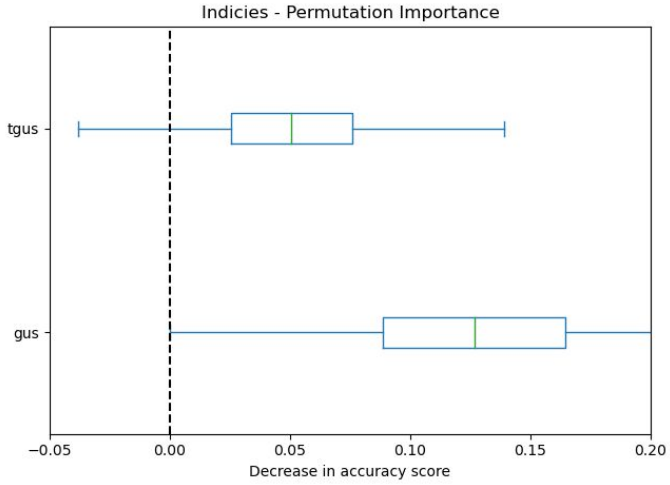
“Permutation importance does not reflect to the intrinsic predictive value of a feature by itself but **how important this feature is for a particular model**.”

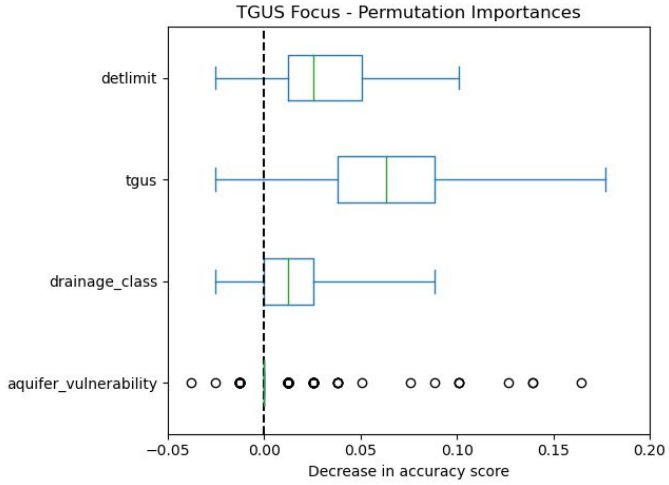
https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-py

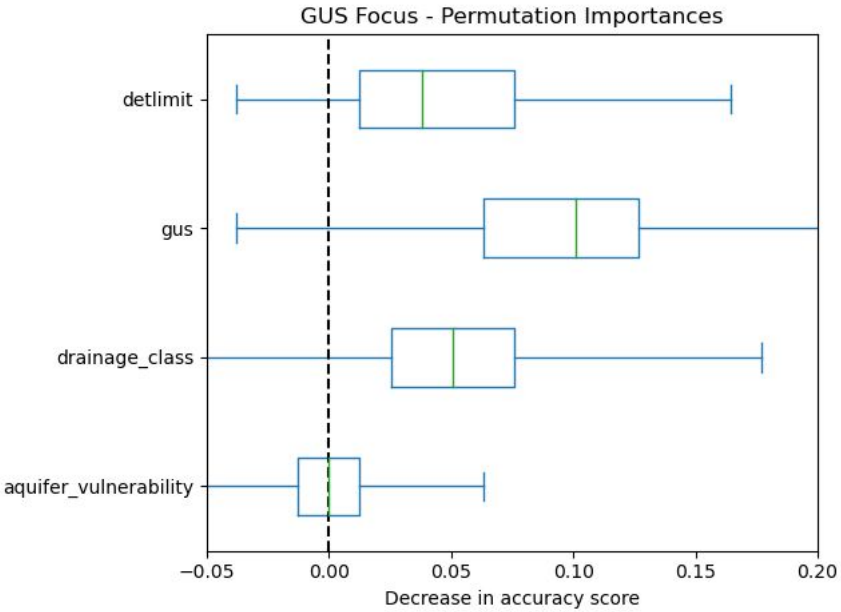
https://scikit-learn.org/stable/modules/permutation_importance.html

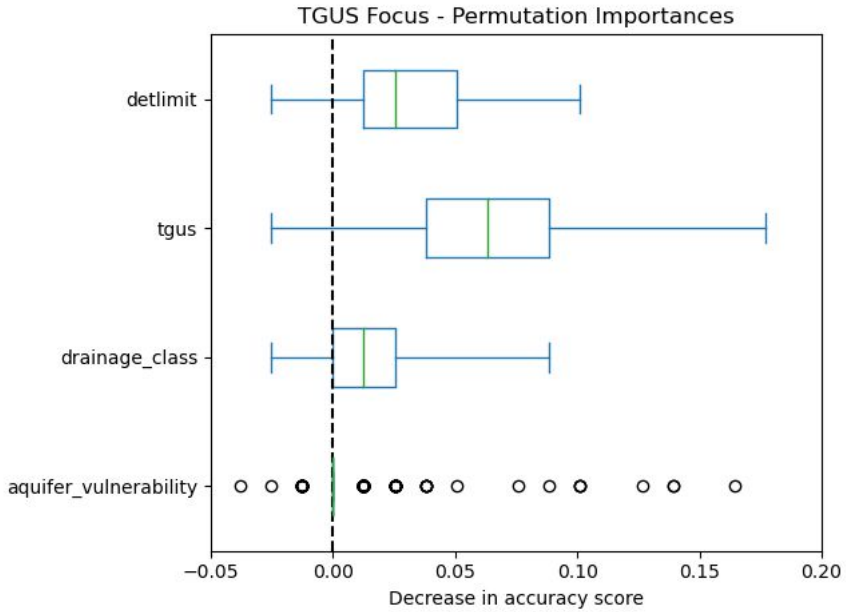

All - Permutation Importances
detlimit
tgus
gus
drainage_class
aquifer_vulnerability
−0.05
0.00
0.05
0.10
0.15
0.20
Decrease in accuracy score


GUS Focus - Permutation Importances
detlimit
gus
drainage_class
aquifer_vulnerability
−0.05
0.00
0.05
0.10
0.15
0.20
Decrease in accuracy score


Indicies - Permutation Importance
tgus
gus
−0.05
0.00
0.05
0.10
0.15
0.20
Decrease in accuracy score


TGUS Focus - Permutation Importances
detlimit
tgus
drainage_class
aquifer_vulnerability
−0.05
0.00
0.05
0.10
0.15
0.20
Decrease in accuracy score





- detection limit and drainage class drop
- aquifer vulnerability provides no value
- consider feature correlation

Moving Forward

- graphs for DEC report?
- balance/add data if possible
- mineral size with permutation importance