UPDATES

June 24, 2024

Notes

- detection limit issues fixed
- focusing on boosted gradient descent model

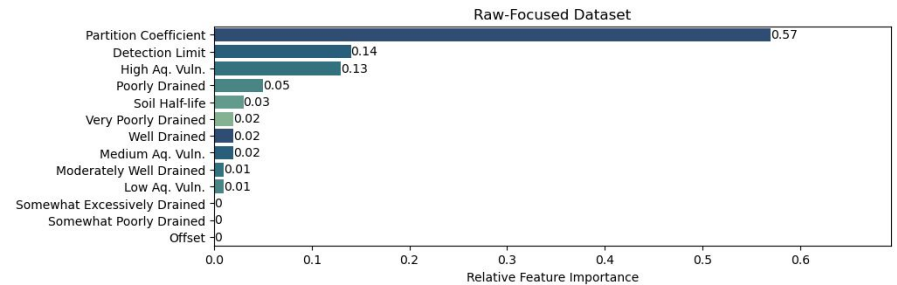# Old dataset, with detection limit, simple TGUS equation (3.4 assumption)

Results - classification accuracy (detected or nondetected)

	Avg. Train %	Avg. Validation %	Avg. Test %	Best Sc. Train %	Best Sc. Validation %	Best Sc. Test %
All	98.2	98.1	94.9	97	97.1	100
GUS Focus	98.1	98.2	95.3	98	95.8	99.1
TGUS Focus	98.2	98.1	95.1	98	97.1	99.1
Raw Focus	98.2	98.2	95	98	98.6	99.1

- no significant differences in t-tests

Feature Importances

- detection limit still high
- redundancy
- GUS/TGUS perform similarly
- drainage class helpful in every scenario
- aquifer vulnerability helpful in every scenario

GUS-Focused Dataset

Feature	Relative Feature Importance
GUS	0.66
Detection Limit	0.12
High Aq. Vuln.	0.1
Poorly Drained	0.03
Somewhat Excessively Drained	0.03
Well Drained	0.02
Very Poorly Drained	0.01
Low Aq. Vuln.	0.01
Medium Aq. Vuln.	0.01
Moderately Well Drained	0
Somewhat Poorly Drained	0
Offset	0

TGUS-Focused Dataset

Feature	Relative Feature Importance
TGUS	0.59
Detection Limit	0.15
High Aq. Vuln.	0.14
Low Aq. Vuln.	0.04
Poorly Drained	0.02
Very Poorly Drained	0.02
Medium Aq. Vuln.	0.02
Moderately Well Drained	0.01
Somewhat Excessively Drained	0.01
Somewhat Poorly Drained	0.01
Well Drained	0
Offset	0

All-Value Dataset

Feature	Relative Feature Importance
GUS	0.45
Partition Coefficient	0.11
High Aq. Vuln.	0.1
TGUS	0.06
Detection Limit	0.06
Somewhat Excessively Drained	0.04
Very Poorly Drained	0.04
Medium Aq. Vuln.	0.03
Soil Half-life	0.03
Somewhat Poorly Drained	0.02
Well Drained	0.02
Low Aq. Vuln.	0.02
Moderately Well Drained	0.01
Poorly Drained	0.01
Offset	0

Raw-Focused Dataset

Feature	Relative Feature Importance
Partition Coefficient	0.57
Detection Limit	0.14
High Aq. Vuln.	0.13
Poorly Drained	0.05
Soil Half-life	0.03
Very Poorly Drained	0.02
Well Drained	0.02
Medium Aq. Vuln.	0.02
Moderately Well Drained	0.01
Low Aq. Vuln.	0.01
Somewhat Excessively Drained	0
Somewhat Poorly Drained	0
Offset	0

Old dataset, without detection limit, simple TGUS equation (3.4 assumption)
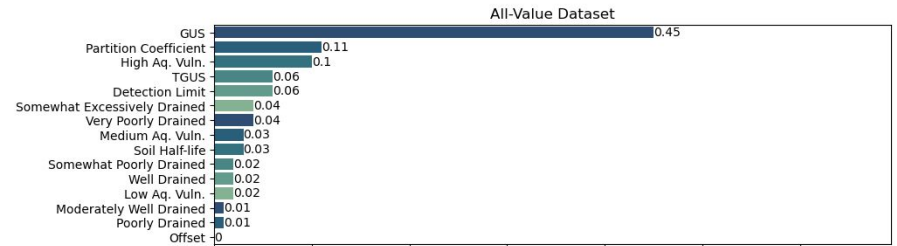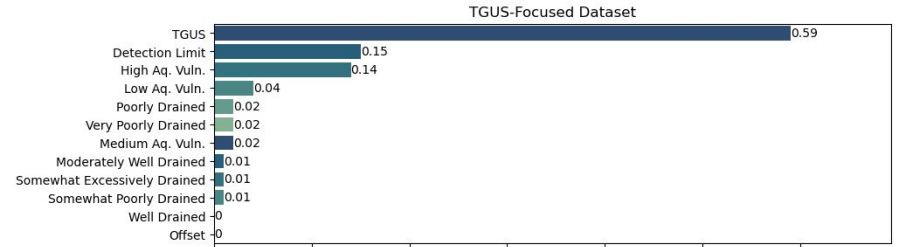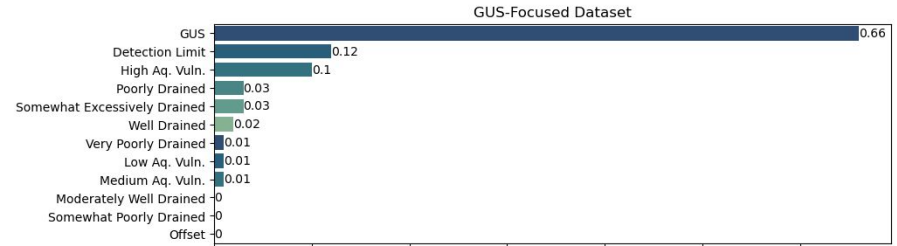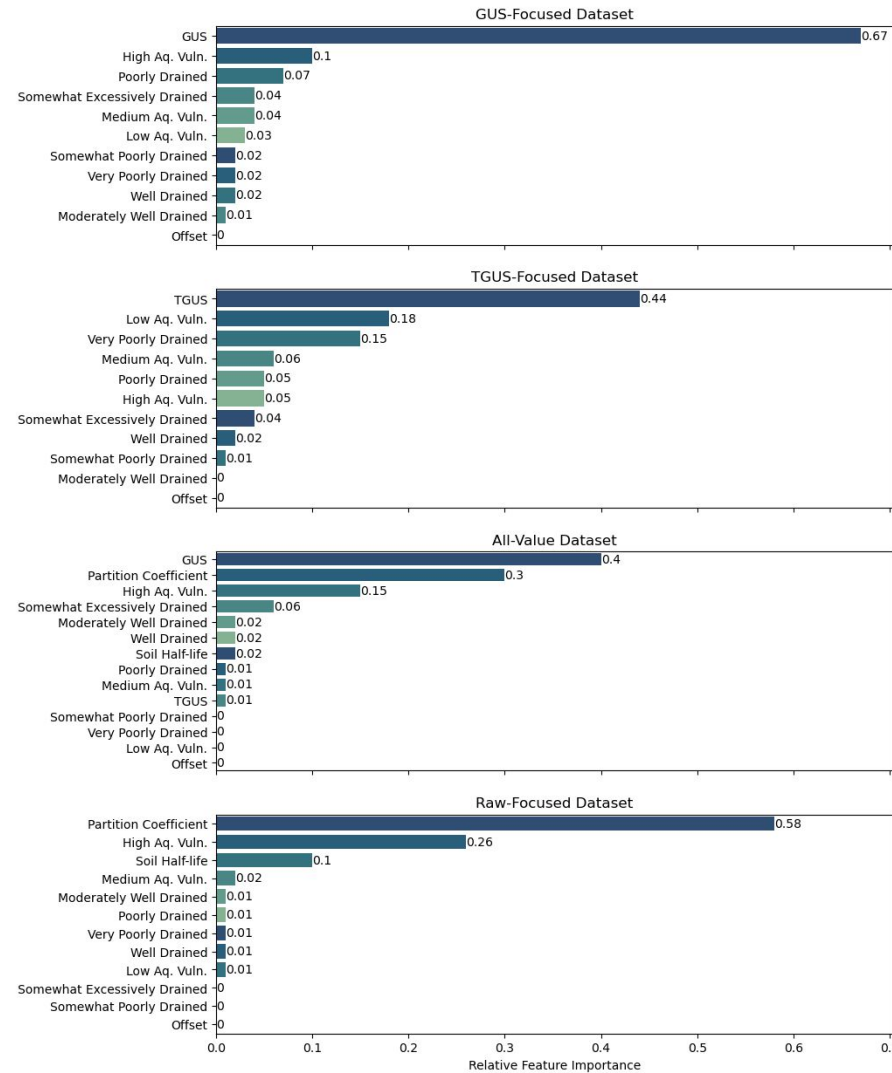
Results - classification accuracy (detected or nondetected)

	Avg. Train %	Avg. Validation %	Avg. Test %	Best Sc. Train %	Best Sc. Validation %	Best Sc. Test %
All	96.3	96.8	93	95	95.7	98.1
GUS Focus	96.4	97	93.4	95	98.6	99.1
TGUS Focus	96.4	96.7	93	95	94.4	98.1
Raw Focus	96.5	97.1	92.9	95	95.8	99.1

- one significant t-test diff between TGUS and Raw in the validation portion…not super important

Feature Importances

- different than first report...unsure why
- TGUS and GUS both highest in their sets
- categorical data still important

GUS-Focused Dataset

Feature	Value
GUS	0.67
High Aq. Vuln.	0.1
Poorly Drained	0.07
Somewhat Excessively Drained	0.04
Medium Aq. Vuln.	0.04
Low Aq. Vuln.	0.03
Somewhat Poorly Drained	0.02
Very Poorly Drained	0.02
Well Drained	0.02
Moderately Well Drained	0.01
Offset	0

TGUS-Focused Dataset

Feature	Value
TGUS	0.44
Low Aq. Vuln.	0.18
Very Poorly Drained	0.15
Medium Aq. Vuln.	0.06
Poorly Drained	0.05
High Aq. Vuln.	0.05
Somewhat Excessively Drained	0.04
Well Drained	0.02
Somewhat Poorly Drained	0.01
Moderately Well Drained	0
Offset	0

All-Value Dataset

Feature	Value
GUS	0.4
Partition Coefficient	0.3
High Aq. Vuln.	0.15
Somewhat Excessively Drained	0.06
Moderately Well Drained	0.02
Well Drained	0.02
Soil Half-life	0.02
Poorly Drained	0.01
Medium Aq. Vuln.	0.01
TGUS	0.01
Somewhat Poorly Drained	0
Very Poorly Drained	0
Low Aq. Vuln.	0
Offset	0

Raw-Focused Dataset

Feature	Value
Partition Coefficient	0.58
High Aq. Vuln.	0.26
Soil Half-life	0.1
Medium Aq. Vuln.	0.02
Moderately Well Drained	0.01
Poorly Drained	0.01
Very Poorly Drained	0.01
Well Drained	0.01
Low Aq. Vuln.	0.01
Somewhat Excessively Drained	0
Somewhat Poorly Drained	0
Offset	0

Relative Feature Importance

Moving Forward

- feature importance research
- fix TGUS equation
- new data