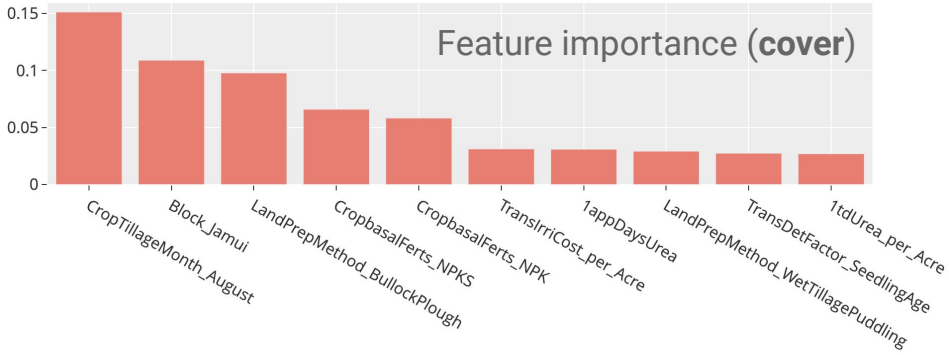
Classification

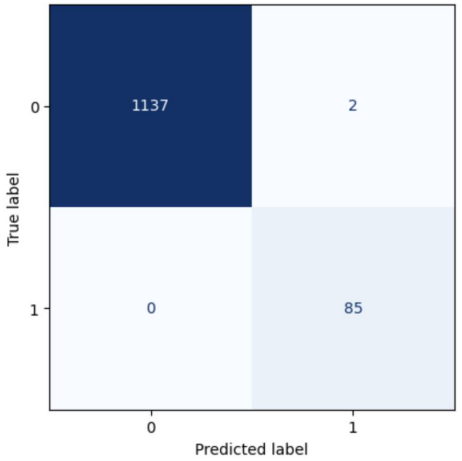
Outliers vs. non-outliers on yield per acre

Goal: predicting which rows are outliers (1) and which rows are not (0) in the test set

- **New binary variable**: “1” for the 21 outliers, “0” for others (determined by Yield_per_Acre > 5000)
- **Class imbalance**: duplicated outlier rows * 10
  - → 231:3849 outlier : non-outlier ratio
- **Train-test split**: only used the train set, and did a 70-30% split
  - number of class 1 in train set: 170
  - number of class 1 in test set: 61
- **Columns dropped**: ID, Set, Yield, Yield_per_Acre, Group_Outlier
  - are there any others I should have removed?
- **Model**: XGBoost Classifier
  - parameters (almost default, did not do much to optimize it): random_state=0, scale_pos_weight=1, min_child_weight=1, max_depth=7, learning_rate=0.1, gamma=0.1, colsample_bytree=0.9, subsample=0.7, eval_metric="error"

Modeling summary





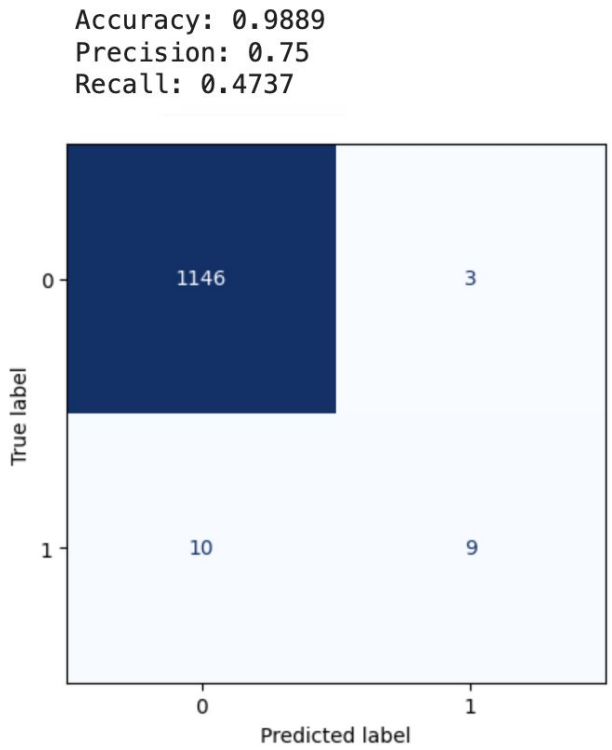
Accuracy: 0.9984
Precision: 0.977
Recall: 1.0

Model 1

duplicated outlier rows *** 10**; training set size = 2856 (**146 class 1**), train set size = 1224 (**85 class 1**)

Number of class 1 predictions in the actual test set: **6** (out of 1290) (all Jamui; all LandPrepMethod BullockPlough & WetTillagePuddling = True)

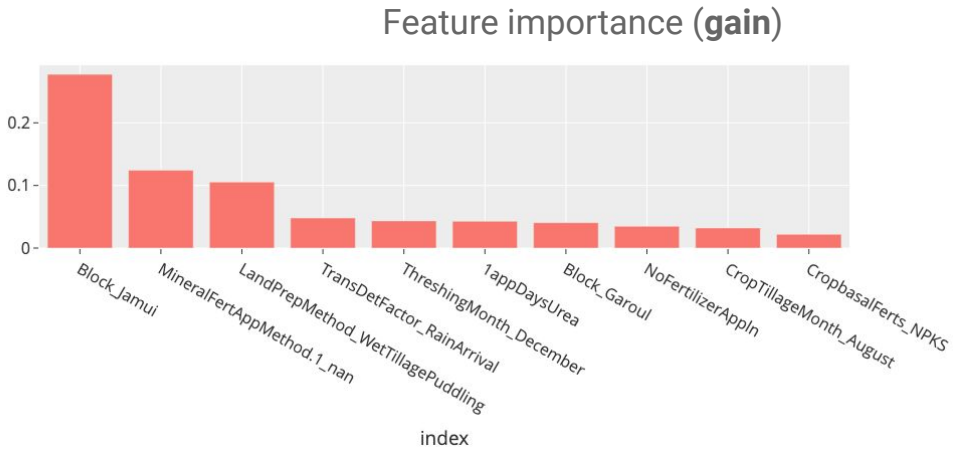
Number of class 1 predictions in the actual test set: 4
(3 of which match with those identified by Model 1)



Model 2

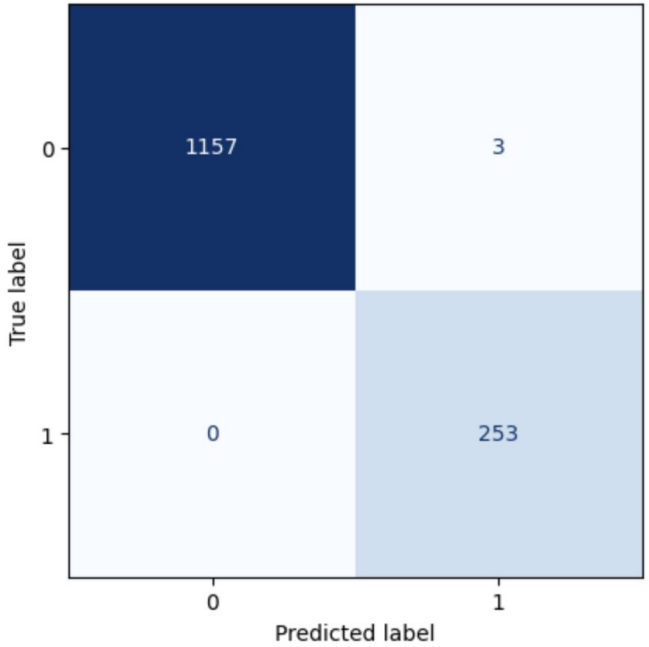
duplicated outlier rows **once**; training set n=2723 (23 class 1); testing set n=1168 (19 class 1)

Number of class 1 predictions in the actual test set: **4** (out of 1290)



Number of class 1 predictions in the actual test set: 11
(which include those identified by Models 1 and 2)

```
Accuracy: 0.9979
Precision: 0.9883
Recall: 1.0
```



Model 3

duplicated outlier rows * 40; training set n=3297 (608 class 1); testing set n=1413 (253 class 1)

Number of class 1 predictions in the actual test set: **11** (out of 1290)

Ran the model to predict outlier/not-outlier on the test set, exported df as “preprocessed_with_outlier_classif.csv”

Modeling summary