Evaluate testing data (multi-class) - Lasso

John Doe

2019-01-30

Contents

Note: The two differences between Lasso and Tree-based methods are:

1. Lasso has its own inherent feature selection process.
2. Lasso's vimp will be based on how many times the feature exist in all runs. Regression coefficients may be presented for binary outcomes

```
## user input
project_home <- "~/EVE/examples"
project_name <- "lasso_multi_outCV_test"
```

0. Load Data

```
## 300 of samples were used
```

```
## 100 of full features
```

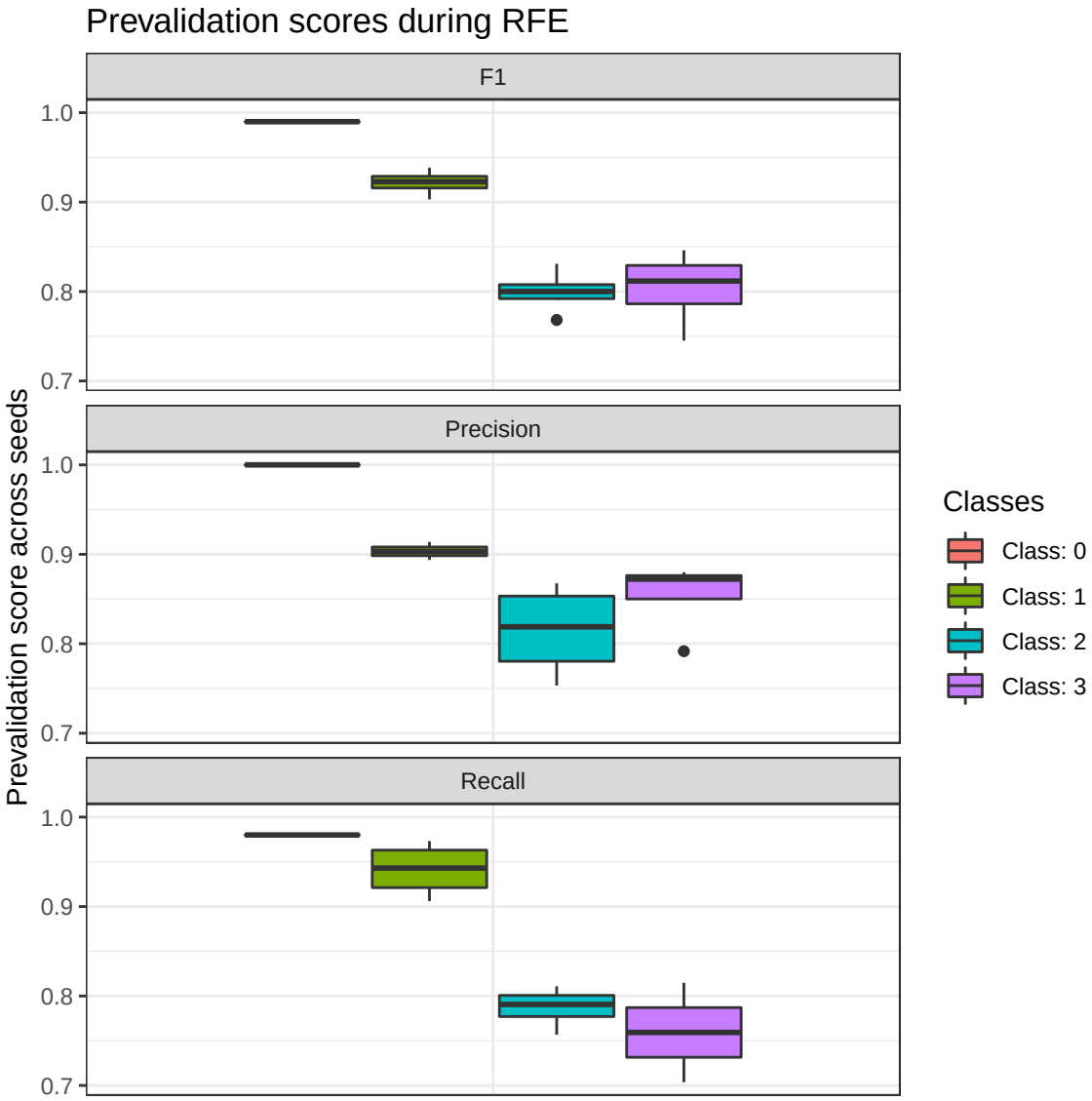
```
## 4 runs, each run contains 3 CVs.
```

```
## Labels:
```

```
##
##   0   1   2   3
##  50 149  74  27
```

1. Scores

1.1 Scores per Class

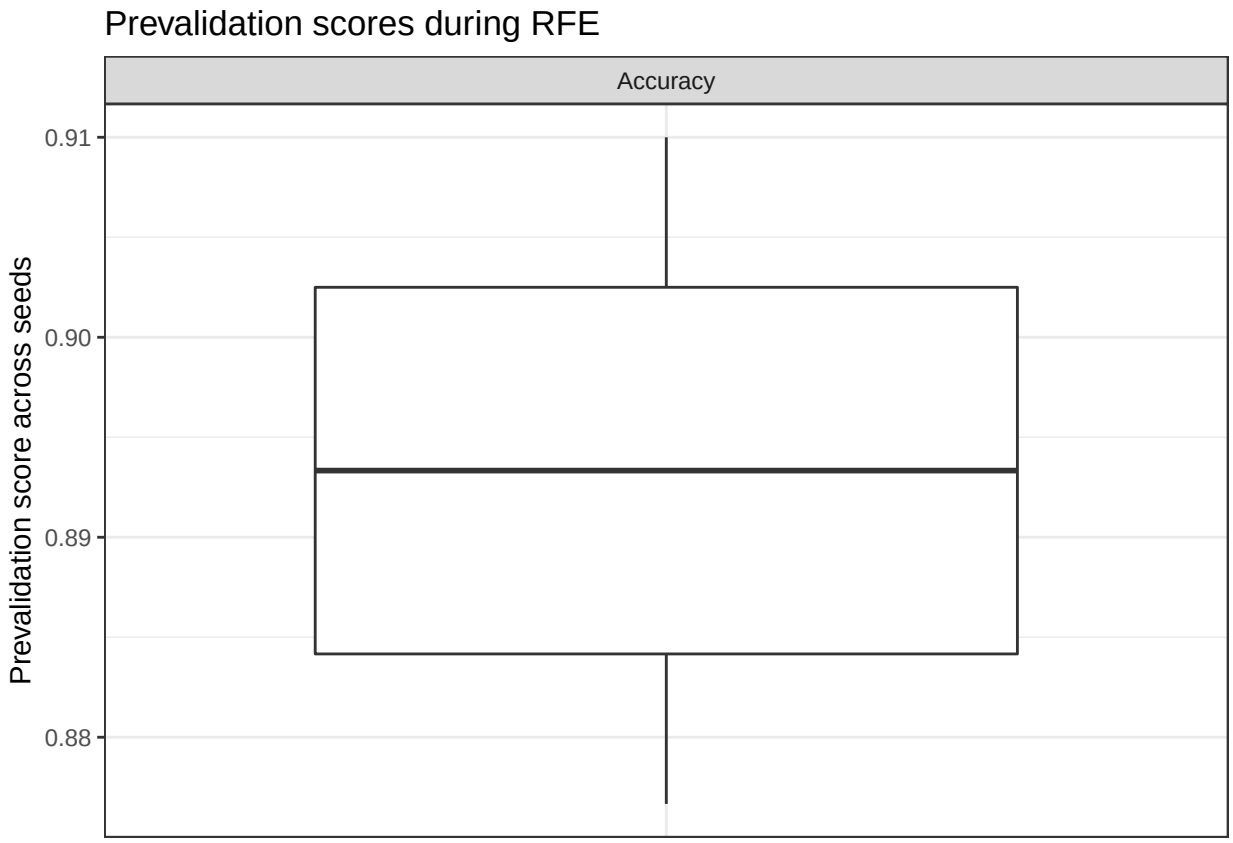


Confusion Matrix

```
## confusion matrix at feature size = 100
## sum across 4 seeds
```

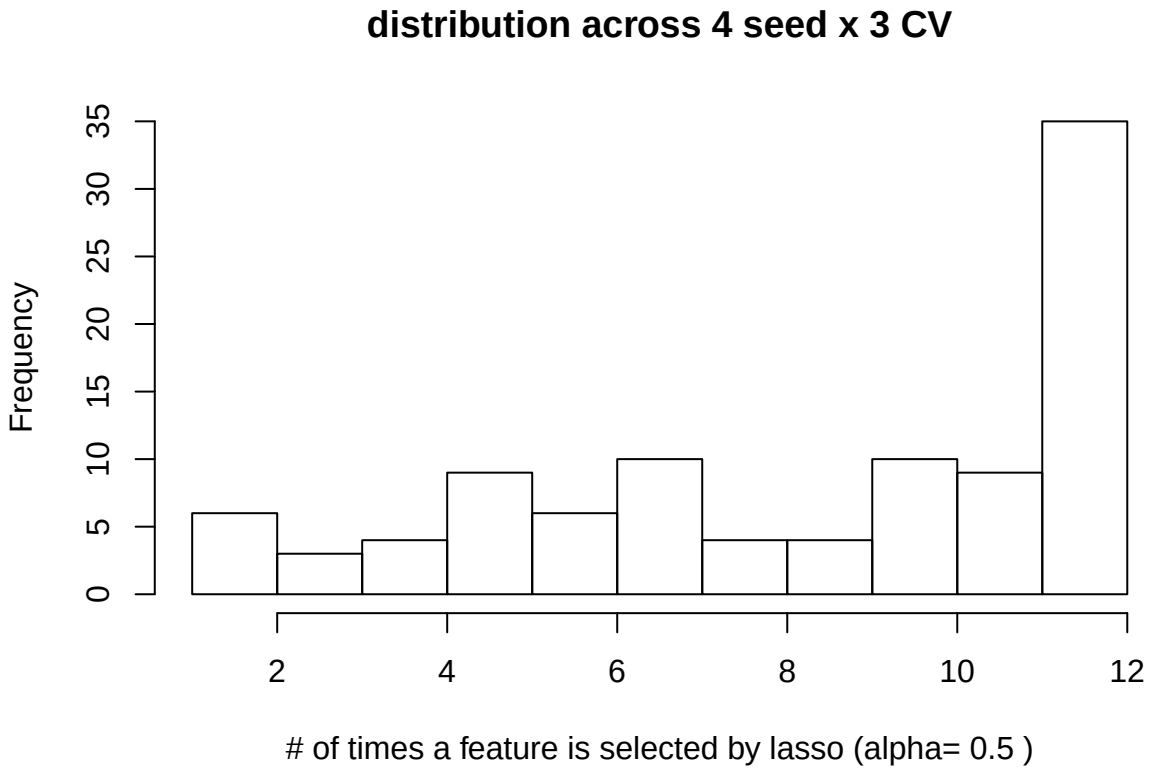
```
##            Reference
## Prediction   0   1   2   3
##          0 196   0   0   0
##          1   0 561  53   7
##          2   0  35 233  19
##          3   4   0  10  82
```

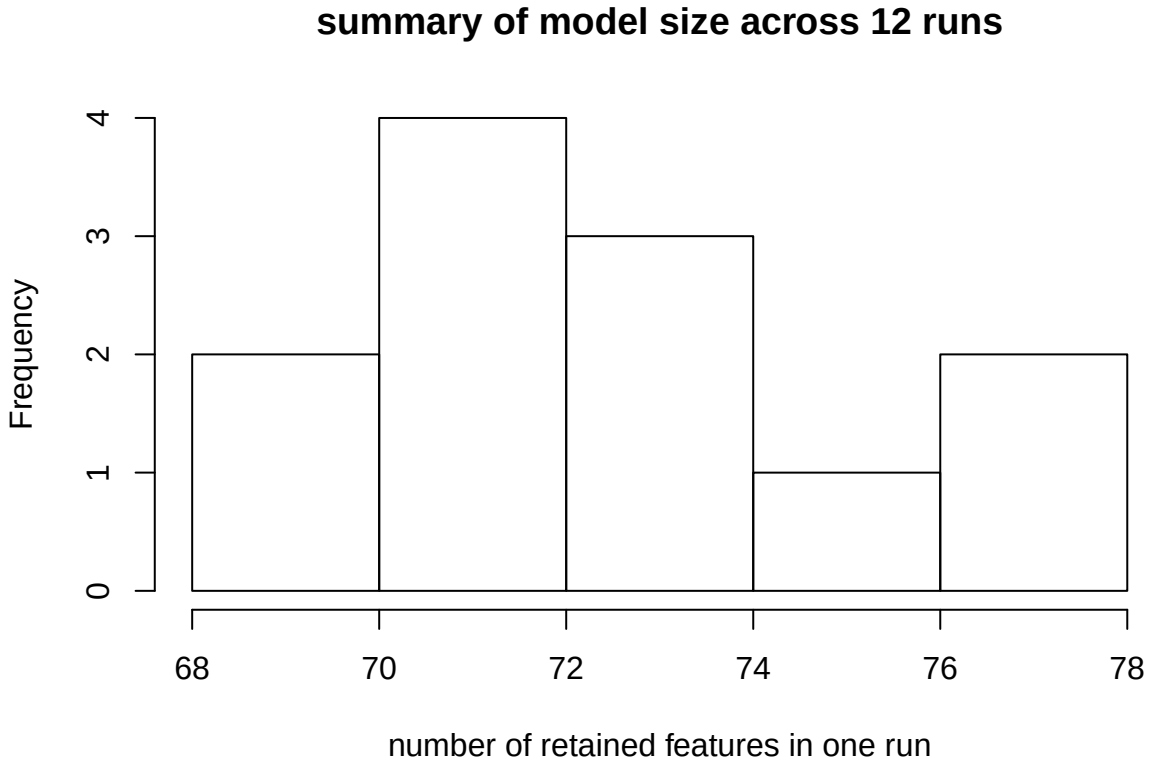
1.2 Average score

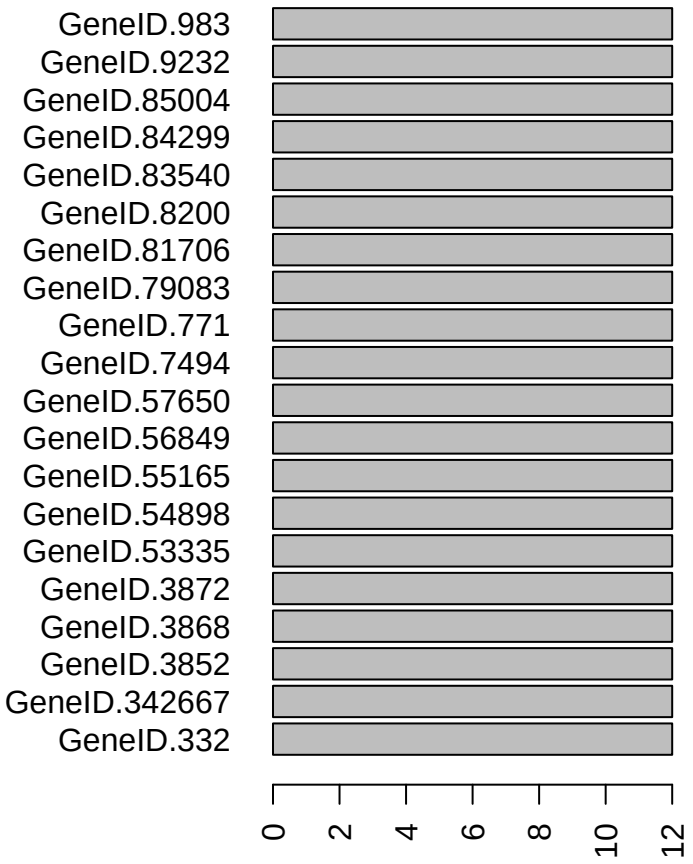


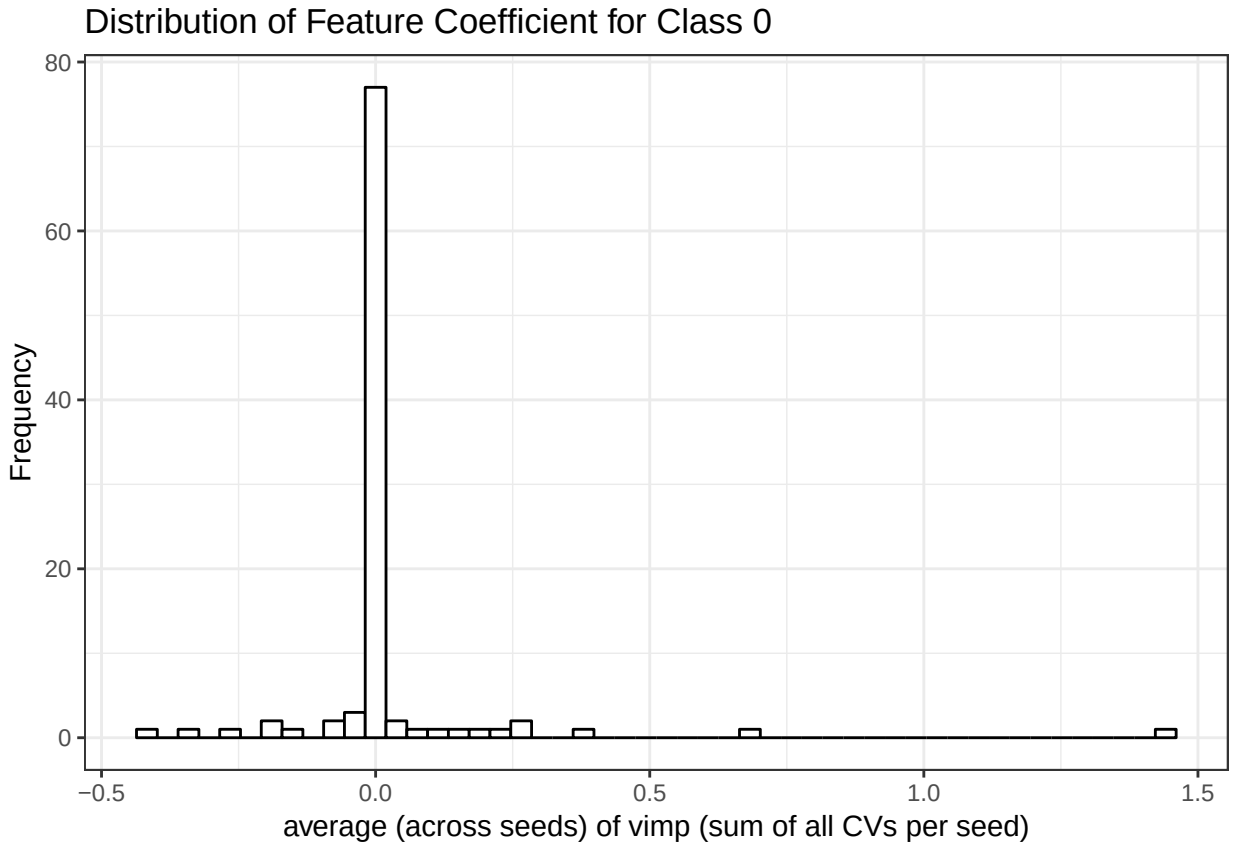
2. Important Features

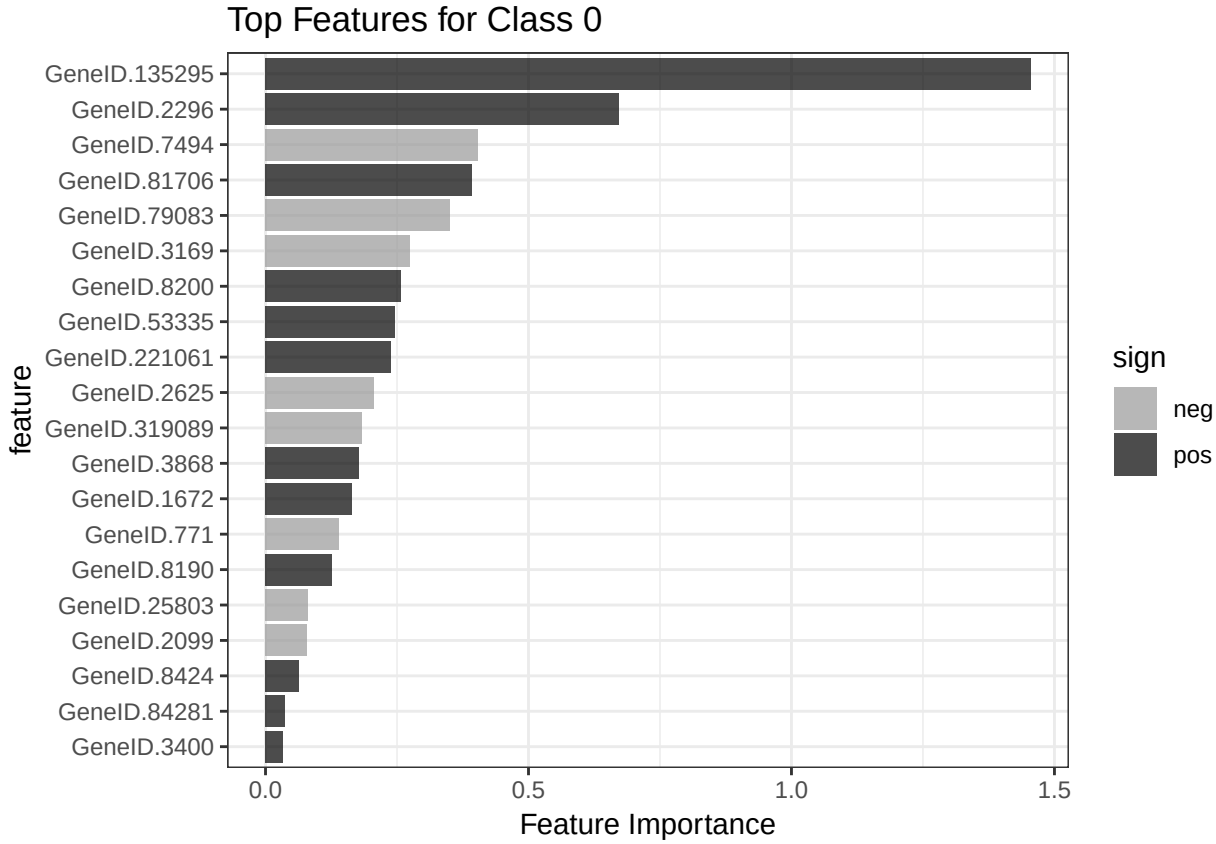
For Lasso, we calculate how many times a given features is being used in all the runs.

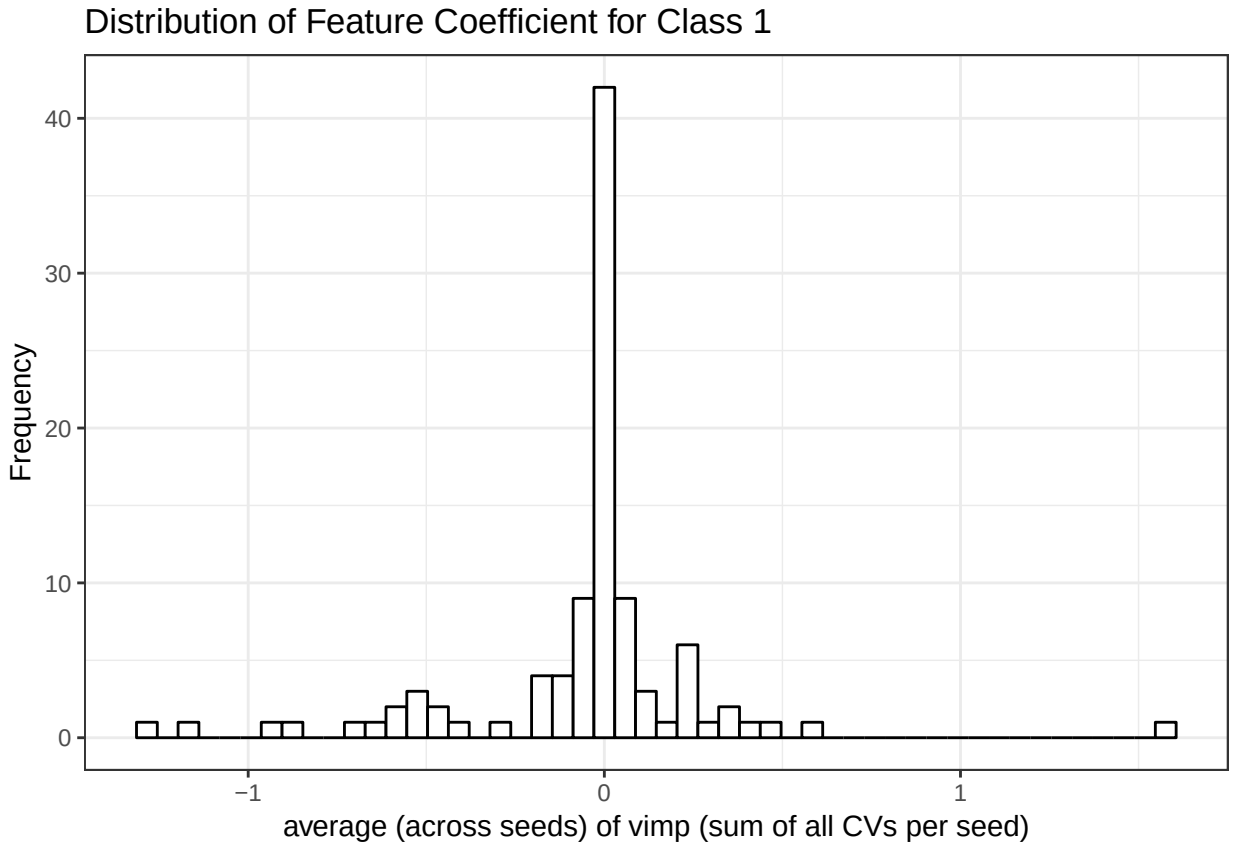

distribution across 4 seed x 3 CV
Frequency
0
5
10
15
20
25
30
35
2
4
6
8
10
12
# of times a feature is selected by lasso (alpha= 0.5)

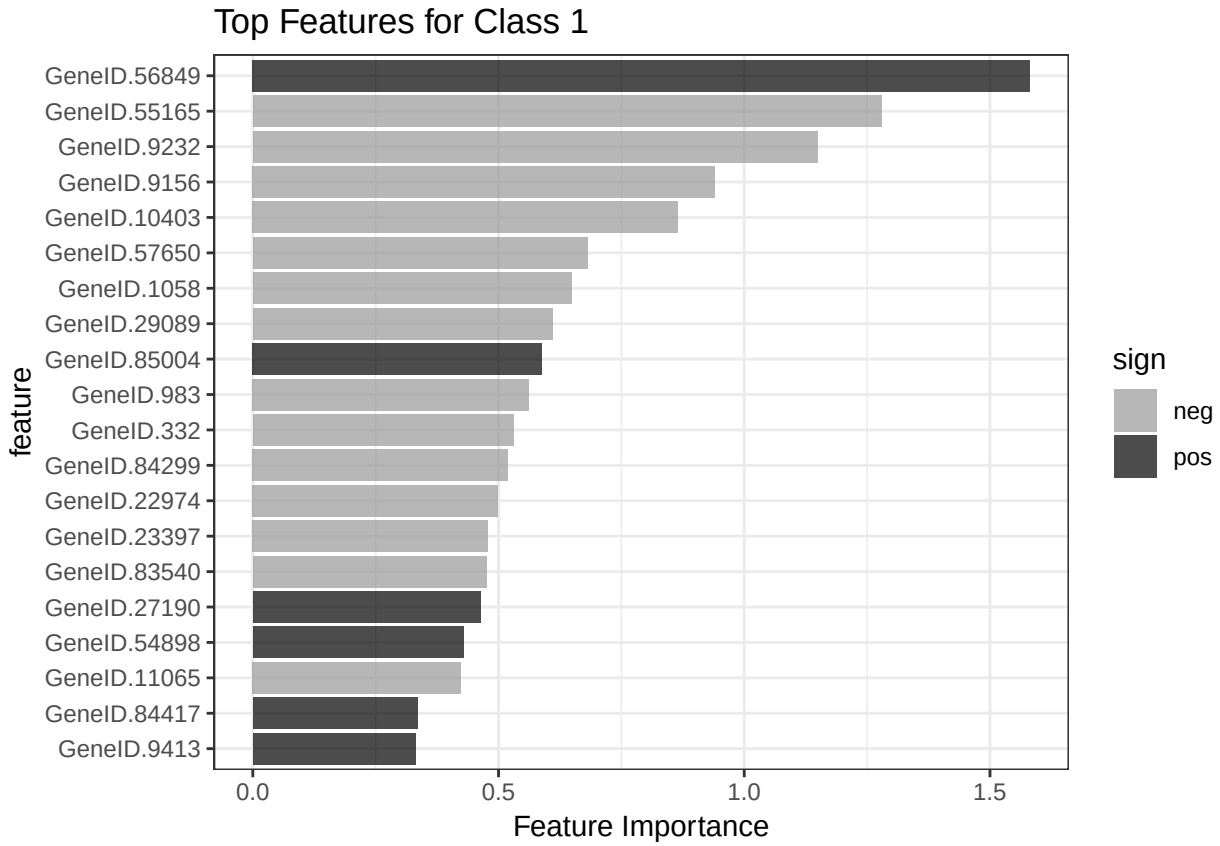

summary of model size across 12 runs
Frequency
0
1
2
3
4
68
70
72
74
76
78
number of retained features in one run

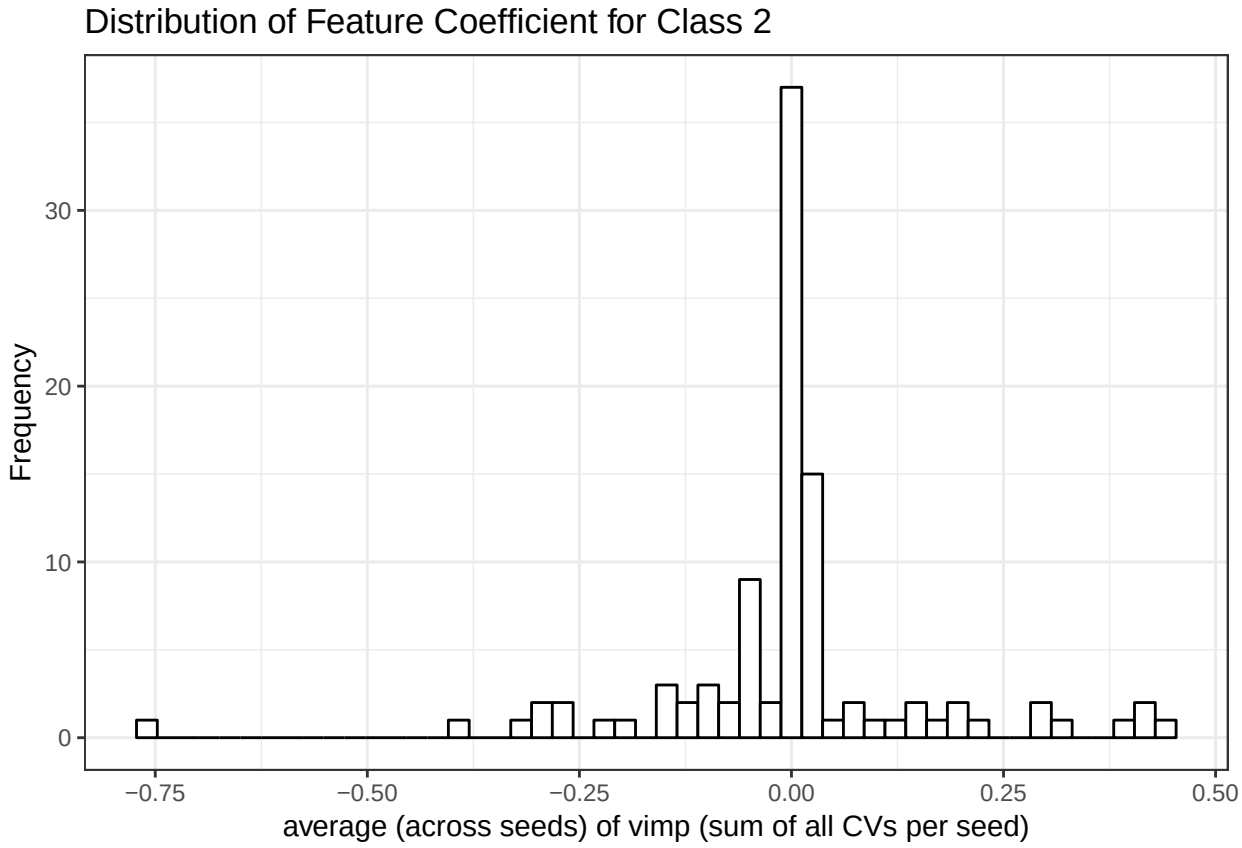

most used features
GeneID.983
GeneID.9232
GeneID.85004
GeneID.84299
GeneID.83540
GeneID.8200
GeneID.81706
GeneID.79083
GeneID.771
GeneID.7494
GeneID.57650
GeneID.56849
GeneID.55165
GeneID.54898
GeneID.53335
GeneID.3872
GeneID.3868
GeneID.3852
GeneID.342667
GeneID.332
0
2
4
6
8
10
12

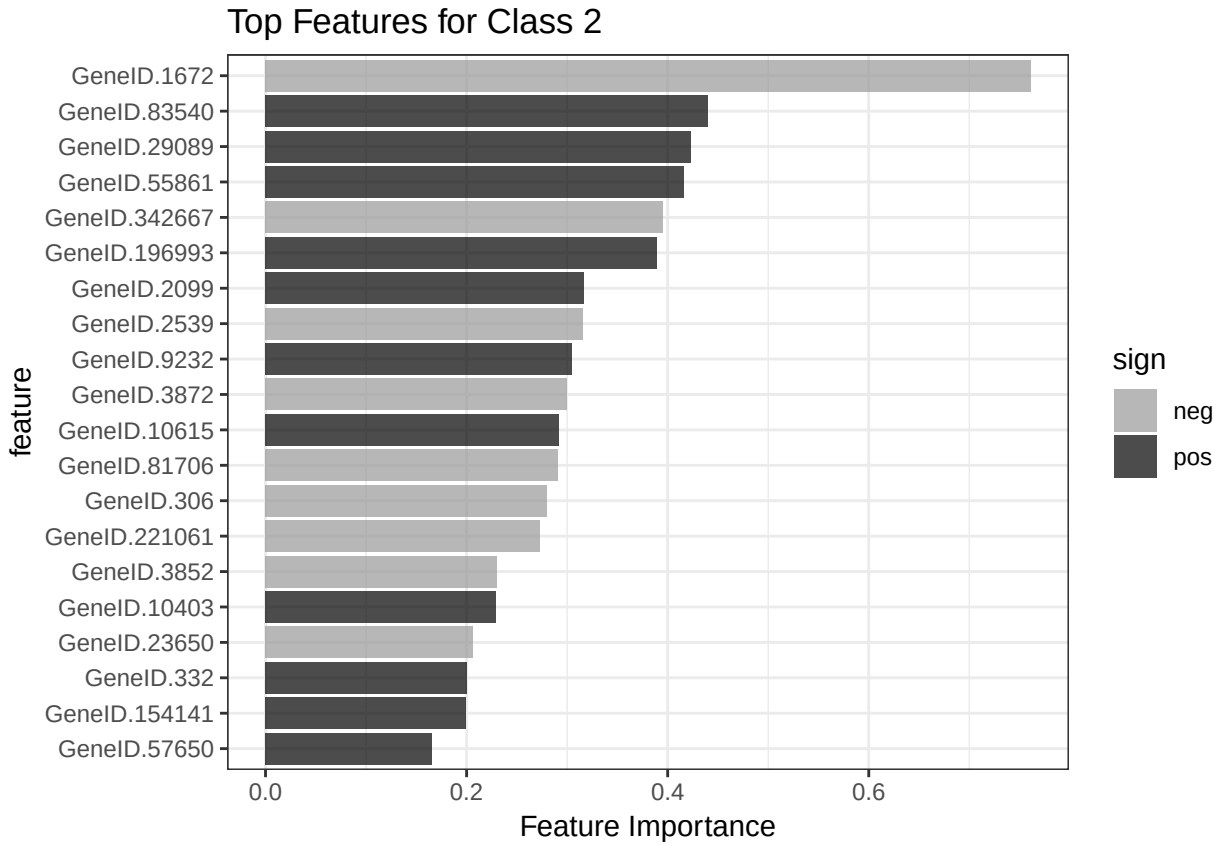

Distribution of Feature Coefficient for Class 0
Frequency
80
60
40
20
0
−0.5
0.0
0.5
1.0
1.5
average (across seeds) of vimp (sum of all CVs per seed)

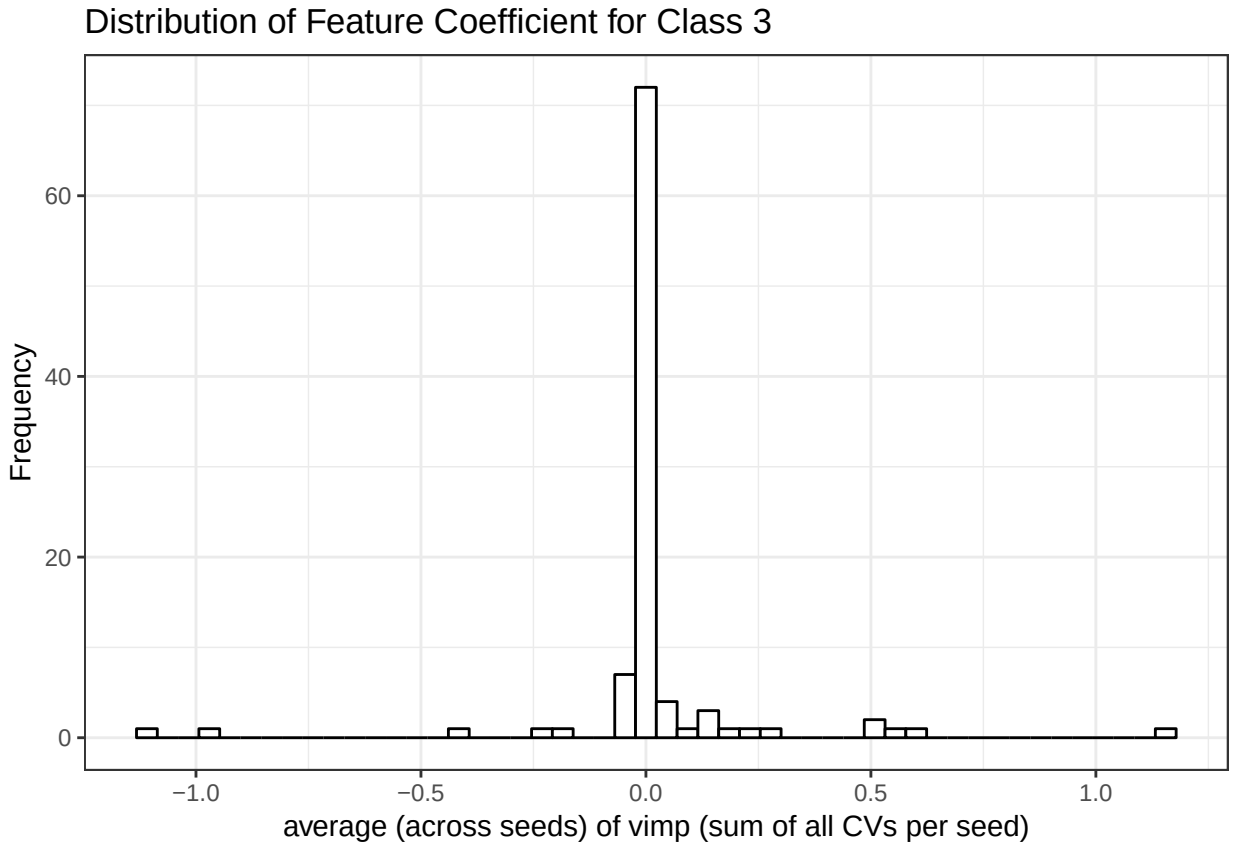

Top Features for Class 0
feature
GeneID.135295
GeneID.2296
GeneID.7494
GeneID.81706
GeneID.79083
GeneID.3169
GeneID.8200
GeneID.53335
GeneID.221061
GeneID.2625
GeneID.319089
GeneID.3868
GeneID.1672
GeneID.771
GeneID.8190
GeneID.25803
GeneID.2099
GeneID.8424
GeneID.84281
GeneID.3400
0.0
0.5
1.0
1.5
Feature Importance
sign
neg
pos

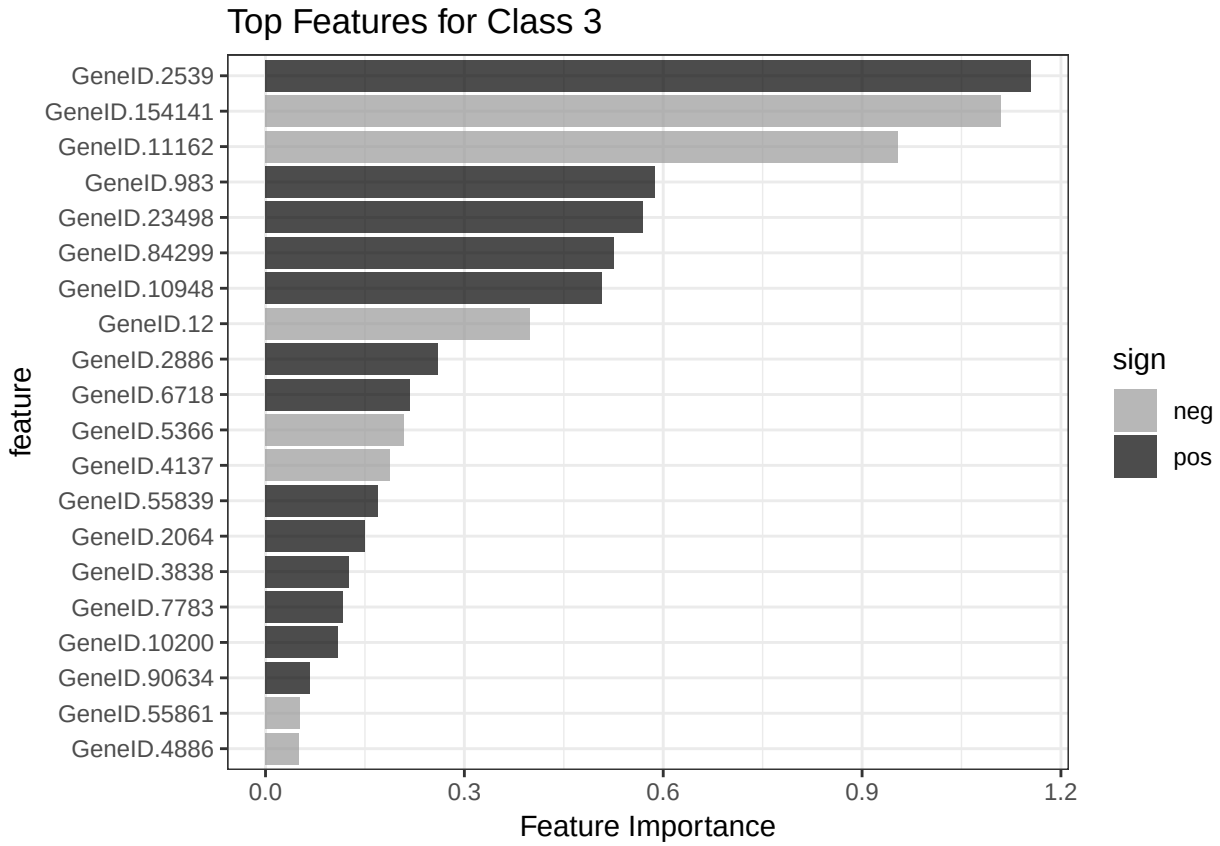

Distribution of Feature Coefficient for Class 1
Frequency
40
30
20
10
0
−1
0
1
average (across seeds) of vimp (sum of all CVs per seed)


Top Features for Class 1
feature
GeneID.56849
GeneID.55165
GeneID.9232
GeneID.9156
GeneID.10403
GeneID.57650
GeneID.1058
GeneID.29089
GeneID.85004
GeneID.983
GeneID.332
GeneID.84299
GeneID.22974
GeneID.23397
GeneID.83540
GeneID.27190
GeneID.54898
GeneID.11065
GeneID.84417
GeneID.9413
0.0
0.5
1.0
1.5
Feature Importance
sign
neg
pos


Distribution of Feature Coefficient for Class 2
Frequency
30
20
10
0
−0.75
−0.50
−0.25
0.00
0.25
0.50
average (across seeds) of vimp (sum of all CVs per seed)


Top Features for Class 2
feature
GeneID.1672
GeneID.83540
GeneID.29089
GeneID.55861
GeneID.342667
GeneID.196993
GeneID.2099
GeneID.2539
GeneID.9232
GeneID.3872
GeneID.10615
GeneID.81706
GeneID.306
GeneID.221061
GeneID.3852
GeneID.10403
GeneID.23650
GeneID.332
GeneID.154141
GeneID.57650
0.0
0.2
0.4
0.6
Feature Importance
sign
neg
pos


Distribution of Feature Coefficient for Class 3
Frequency
60
40
20
0
−1.0
−0.5
0.0
0.5
1.0
average (across seeds) of vimp (sum of all CVs per seed)


Top Features for Class 3
feature
GeneID.2539
GeneID.154141
GeneID.11162
GeneID.983
GeneID.23498
GeneID.84299
GeneID.10948
GeneID.12
GeneID.2886
GeneID.6718
GeneID.5366
GeneID.4137
GeneID.55839
GeneID.2064
GeneID.3838
GeneID.7783
GeneID.10200
GeneID.90634
GeneID.55861
GeneID.4886
0.0
0.3
0.6
0.9
1.2
Feature Importance
sign
neg
pos