Evaluate testing data (binary-class) - Lasso

EVE W.

2020-04-11

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_binary_outCV_test"
```

0. Load Data

```
## Error : $ operator is invalid for atomic vectors
```

```
## 199 of samples were used
```

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

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

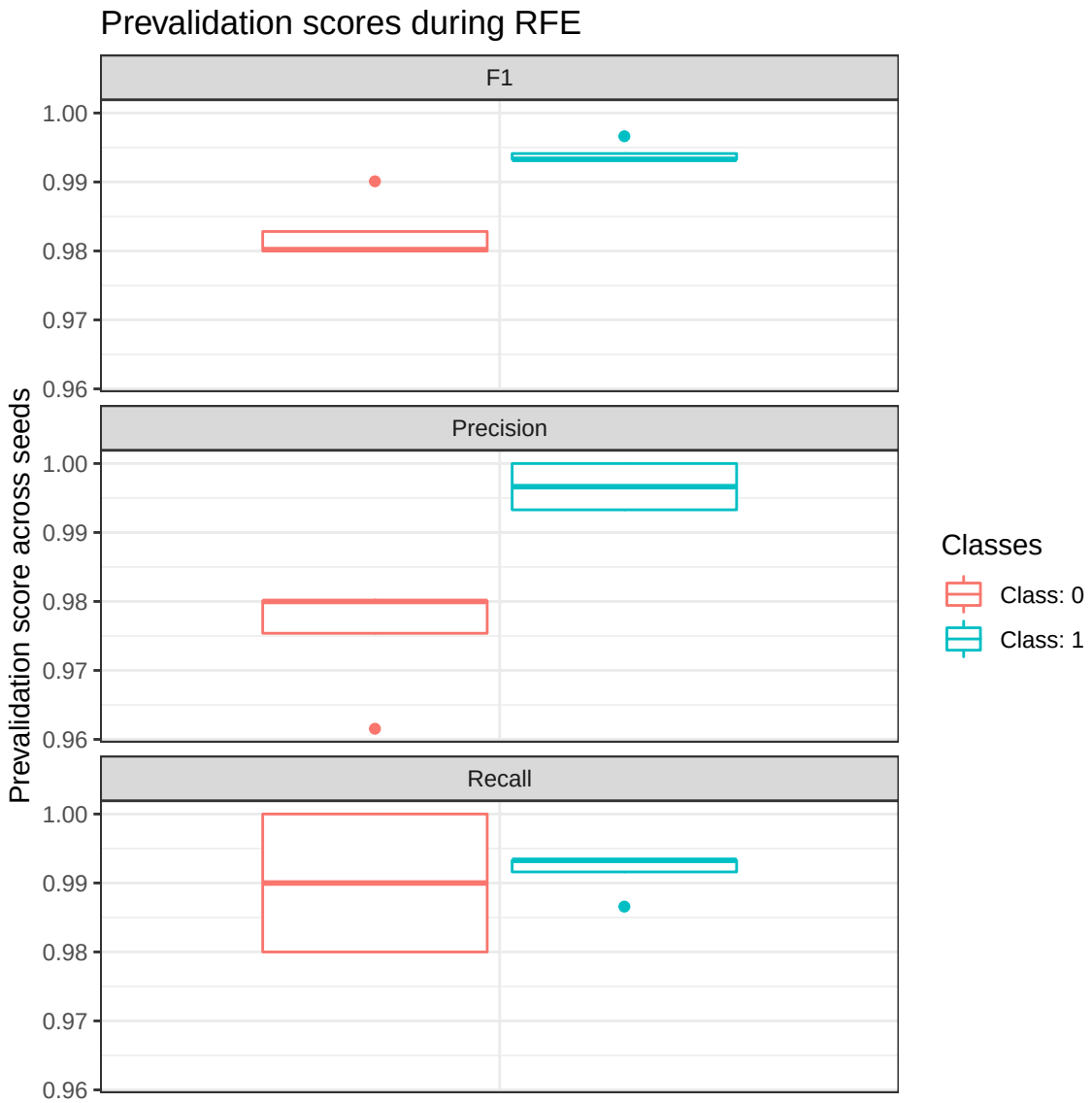
```
## Labels:
```

```
##
##   0   1
##  50 149
```

run with lasso.r.

1. Scores

1.1 Scores per Class

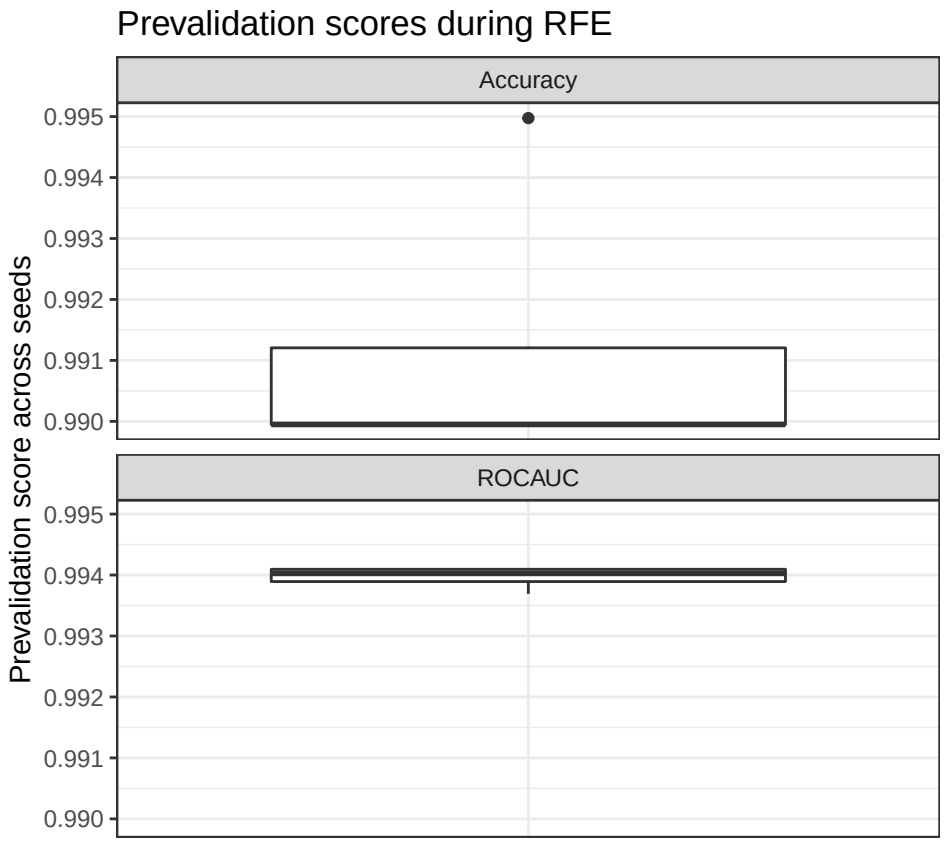


Confusion Matrix

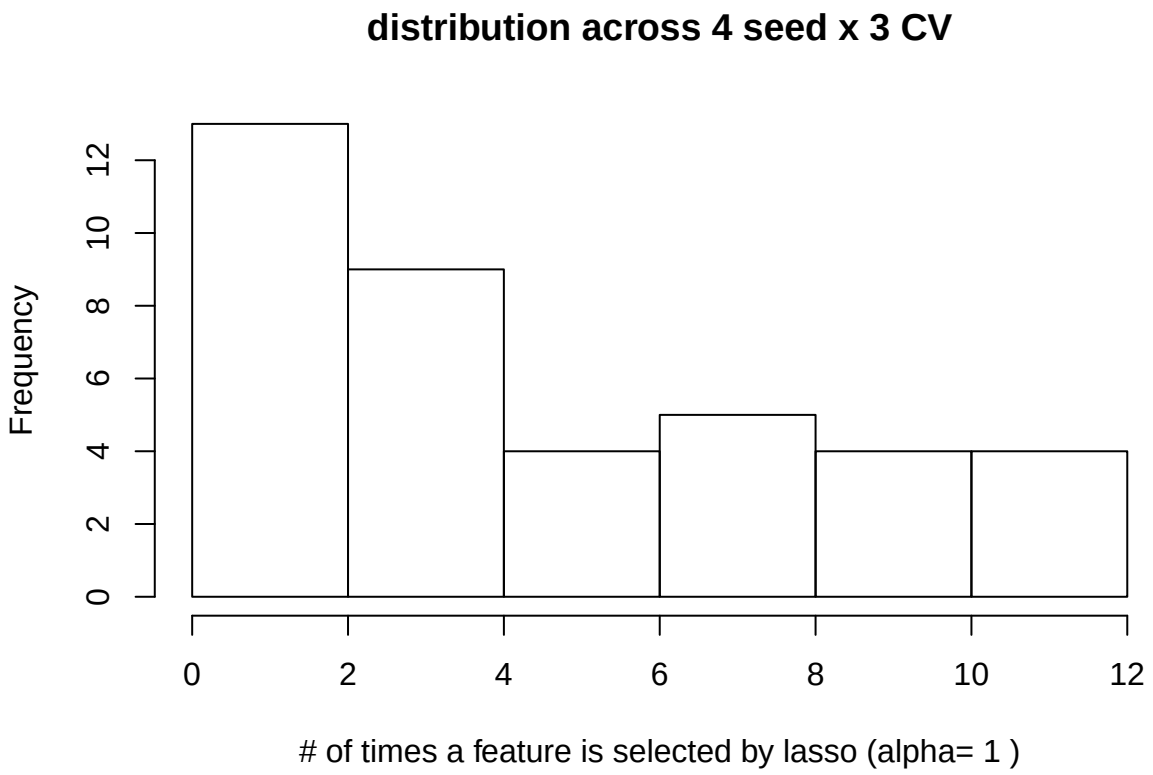
```
## confusion matrix at feature size = 100
## sum across 4 seeds

##            Reference
## Prediction   0   1
##          0 198   5
##          1   2 591
```

1.2 Average score

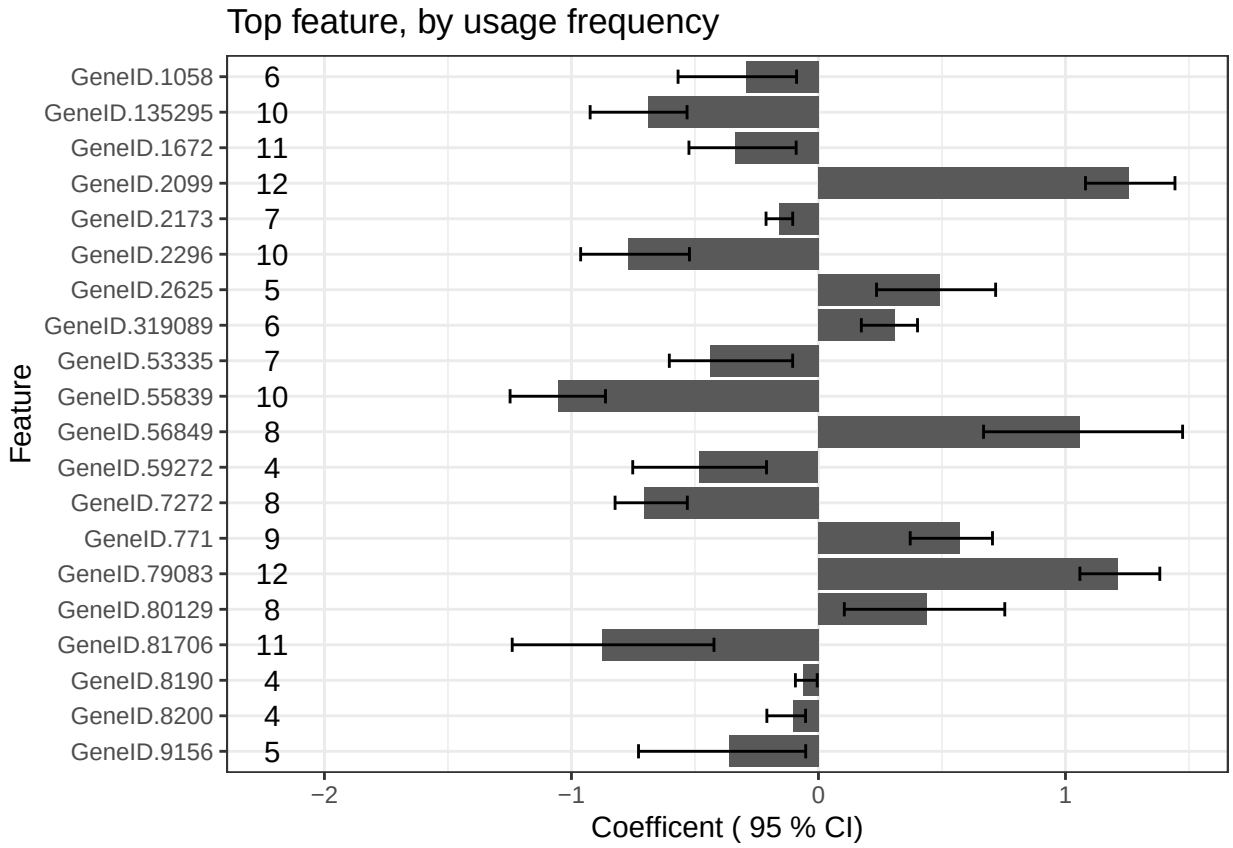


2. Important Features

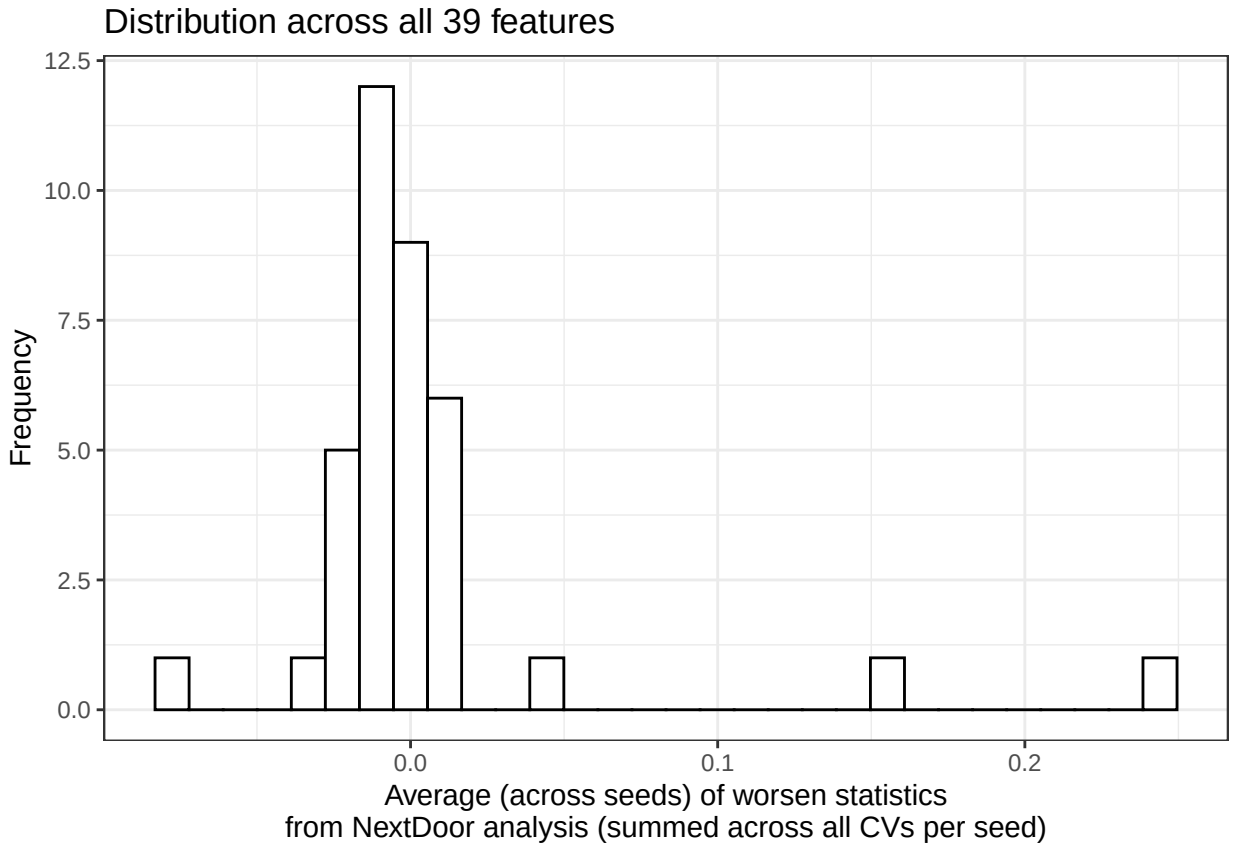


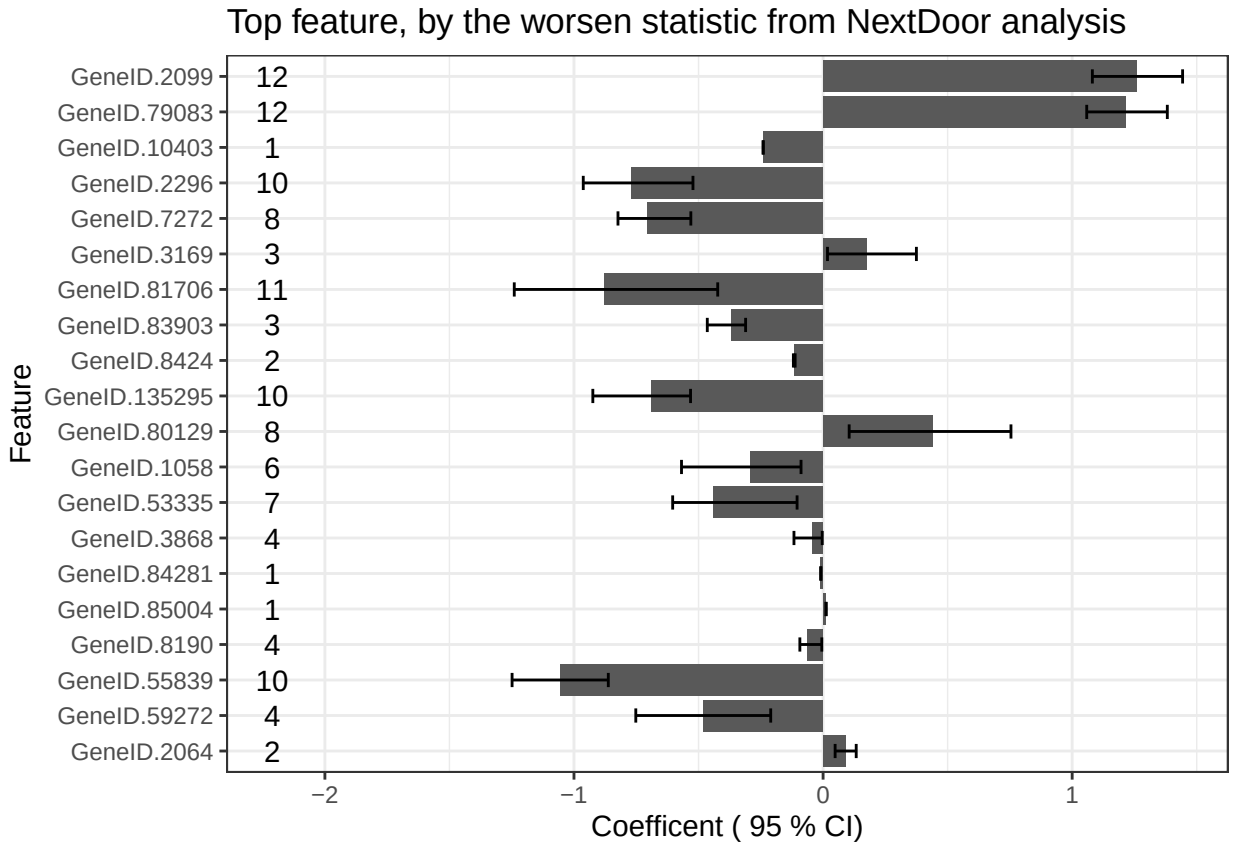

```
## [1] "there are 39 unique features used from the 100 feature set"
## summary of numer of features used in 4 seeds and 3 CVs
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   14.75   16.00   16.25   18.00   19.00
```

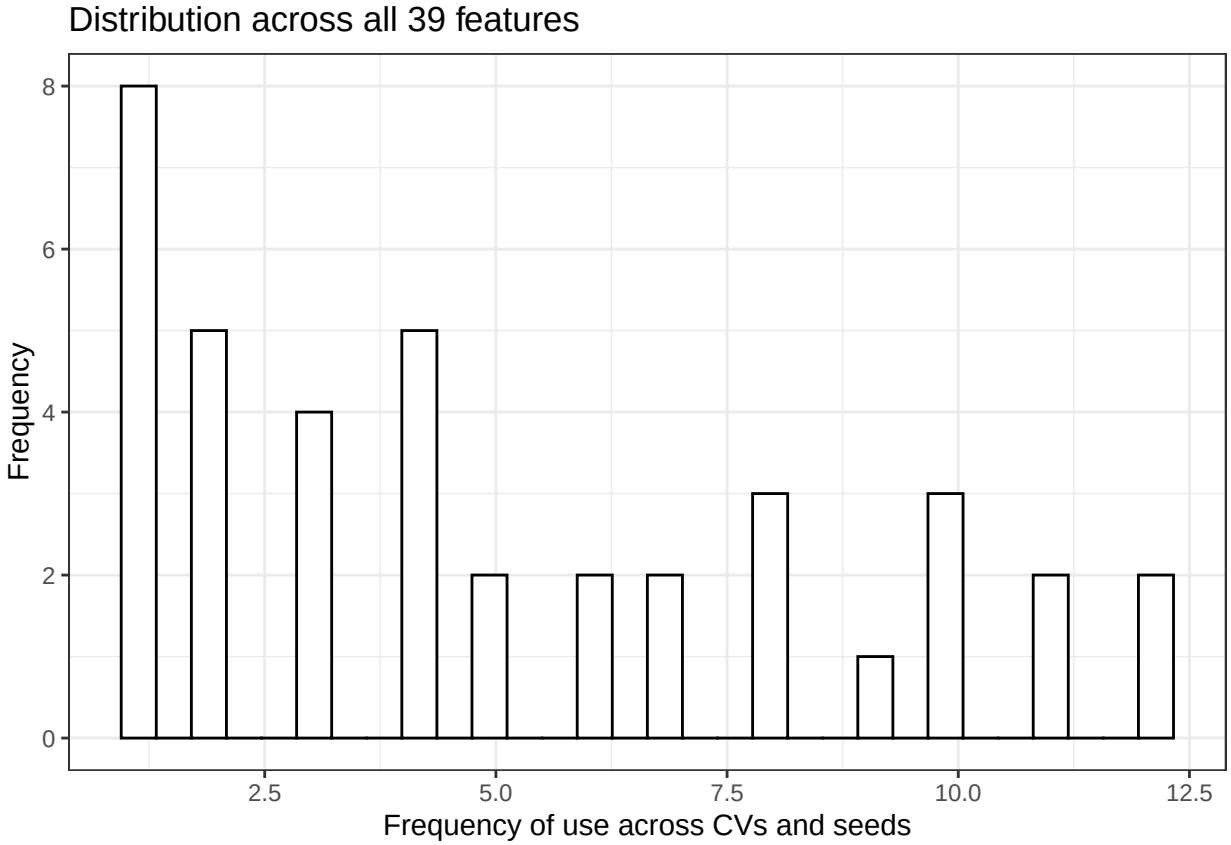


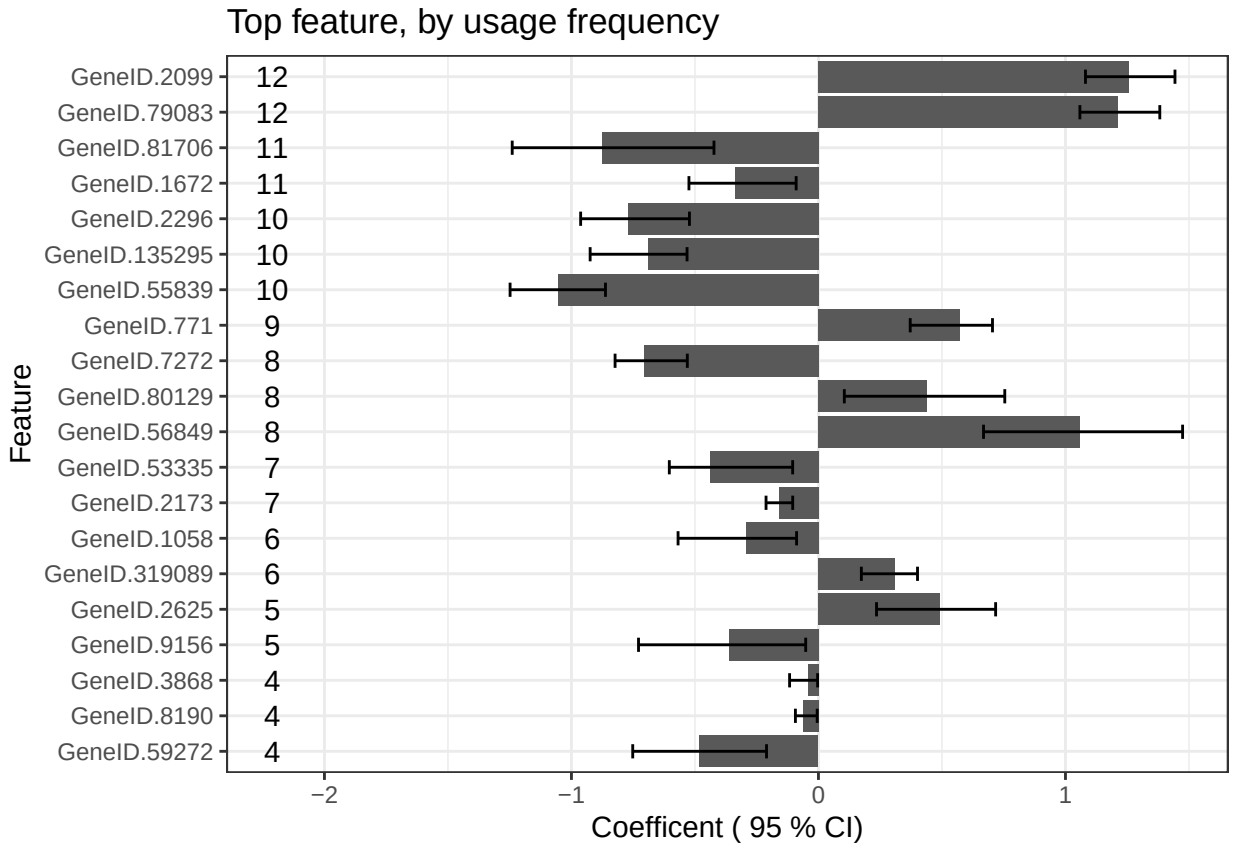

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

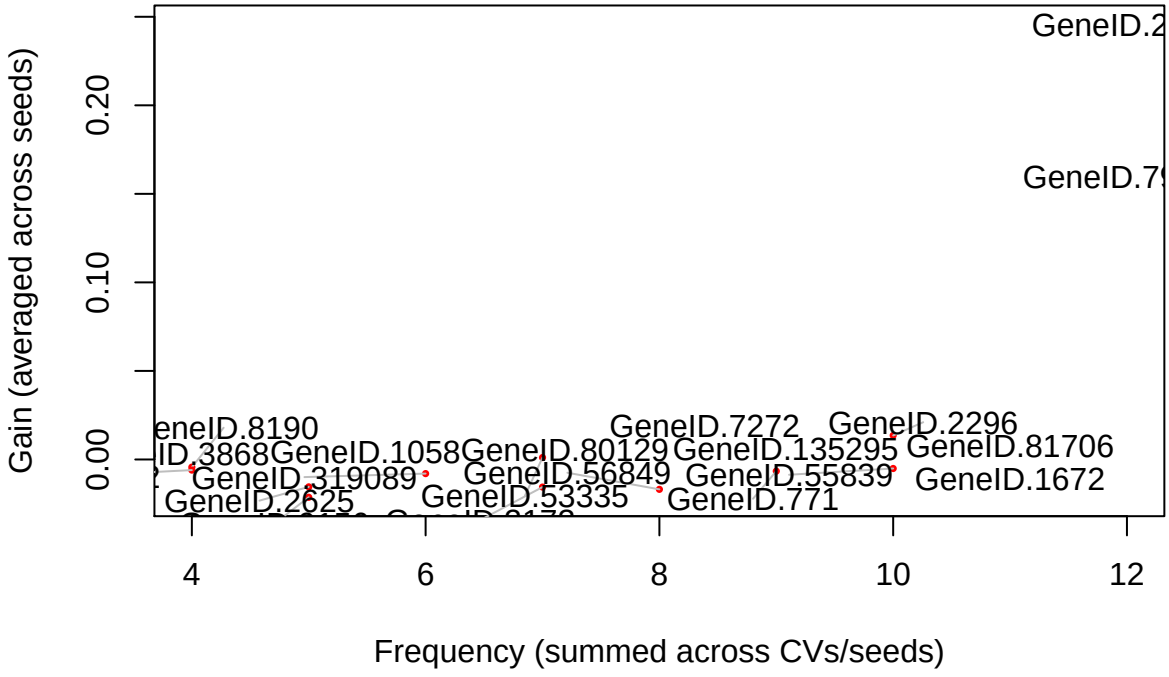

Distribution across all 39 features
Frequency
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.1
0.2
Average (across seeds) of worsen statistics
from NextDoor analysis (summed across all CVs per seed)

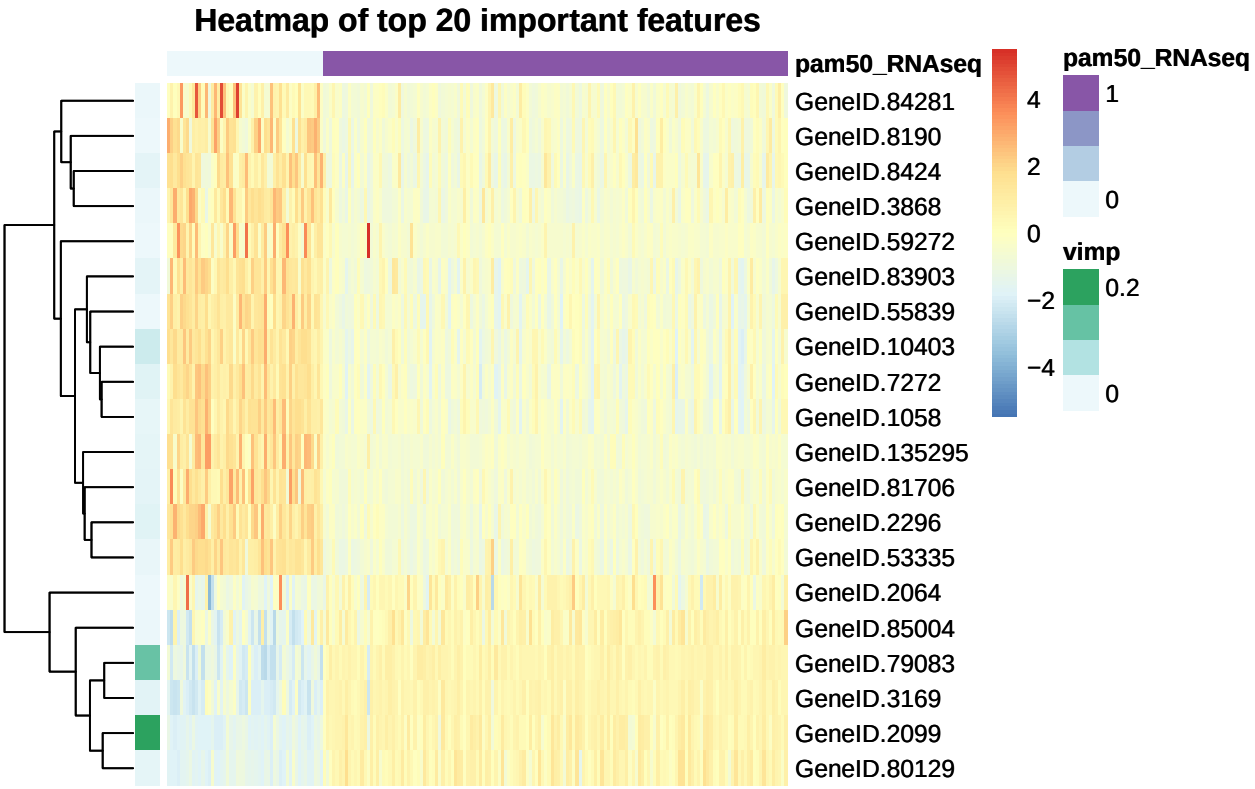

Top feature, by the worsen statistic from NextDoor analysis
Feature
GeneID.2099 12
GeneID.79083 12
GeneID.10403 1
GeneID.2296 10
GeneID.7272 8
GeneID.3169 3
GeneID.81706 11
GeneID.83903 3
GeneID.8424 2
GeneID.135295 10
GeneID.80129 8
GeneID.1058 6
GeneID.53335 7
GeneID.3868 4
GeneID.84281 1
GeneID.85004 1
GeneID.8190 4
GeneID.55839 10
GeneID.59272 4
GeneID.2064 2
−2
−1
0
1
Coefficent (95 % CI)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




Top feature, by usage frequency
GeneID.2099 12
GeneID.79083 12
GeneID.81706 11
GeneID.1672 11
GeneID.2296 10
GeneID.135295 10
GeneID.55839 10
GeneID.771 9
GeneID.7272 8
GeneID.80129 8
GeneID.56849 8
GeneID.53335 7
GeneID.2173 7
GeneID.1058 6
GeneID.319089 6
GeneID.2625 5
GeneID.9156 5
GeneID.3868 4
GeneID.8190 4
GeneID.59272 4
Feature
−2
−1
0
1
Coefficent (95 % CI)


Gain (averaged across seeds)
0.00
0.10
0.20
GeneID.2
GeneID.7
eneID.8190
ID.3868
GeneID.1058
GeneID.80129
GeneID.135295
GeneID.81706
GeneID.319089
GeneID.56849
GeneID.55839
GeneID.1672
GeneID.2625
GeneID.53335
GeneID.771
GeneID.7272
GeneID.2296
4
6
8
10
12
Frequency (summed across CVs/seeds)


Heatmap of top 20 important features
pam50_RNAseq
GeneID.84281
GeneID.8190
GeneID.8424
GeneID.3868
GeneID.59272
GeneID.83903
GeneID.55839
GeneID.10403
GeneID.7272
GeneID.1058
GeneID.135295
GeneID.81706
GeneID.2296
GeneID.53335
GeneID.2064
GeneID.85004
GeneID.79083
GeneID.3169
GeneID.2099
GeneID.80129
4
2
0
−2
−4
pam50_RNAseq
1
0
vimp
0.2
0