# Evaluate testing data (survival) - lasso

EVE W.

2020-05-07

## Contents

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

### 0. Load Data

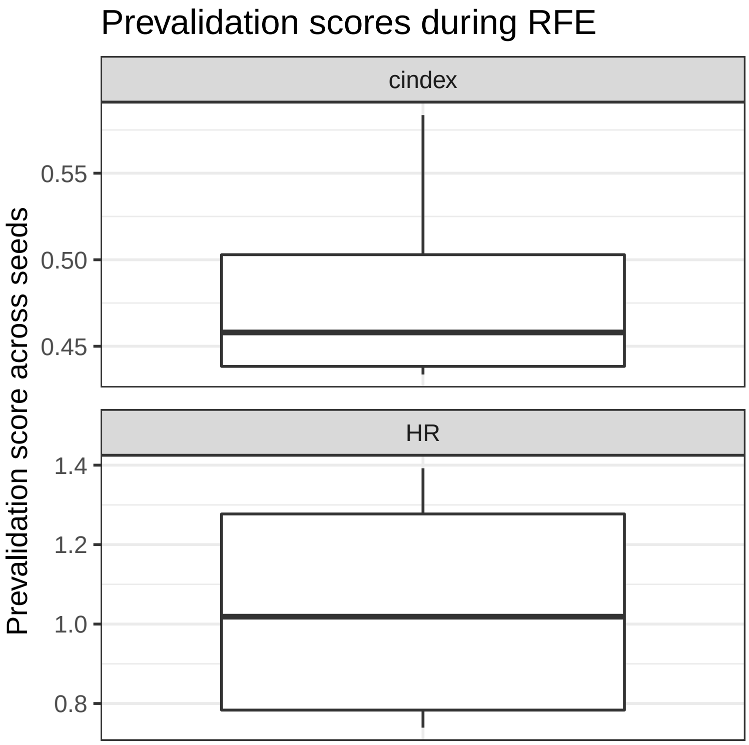
```
300 of samples were used
```

```
100 of full features
```

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

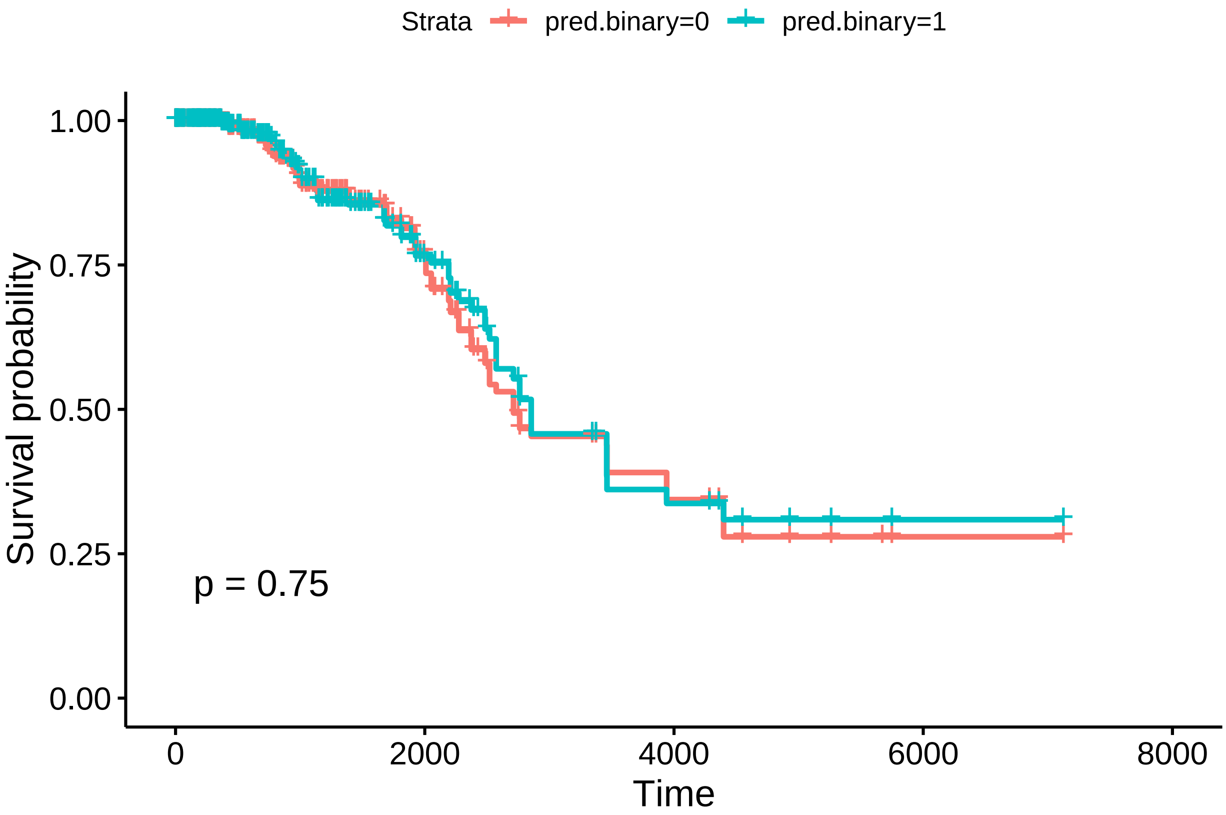
run with lasso.r with alpha = 1.

## 1. Scores

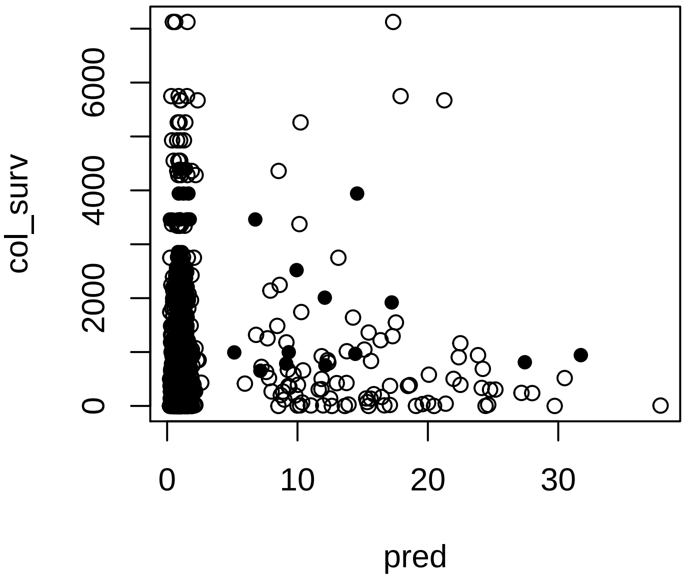


Note for the **HR plot**: A HR value (per seed) is calculated by comparing the survival time between 'long' and 'short' survivors. These two group is defined by splitting samples based on *median* predicted risk score; group_0 is predicted risk scores > median, which can be viewed as 'short survivors'. On the other hand, group_1 can be viewed as 'long survivors'. If the prediction is reasonable, the hazard ratio of group_1/group_0 should be < 1. The actual function used in calculating HR is `coxph(Surv(time, status) ~ group.binary, df)`.

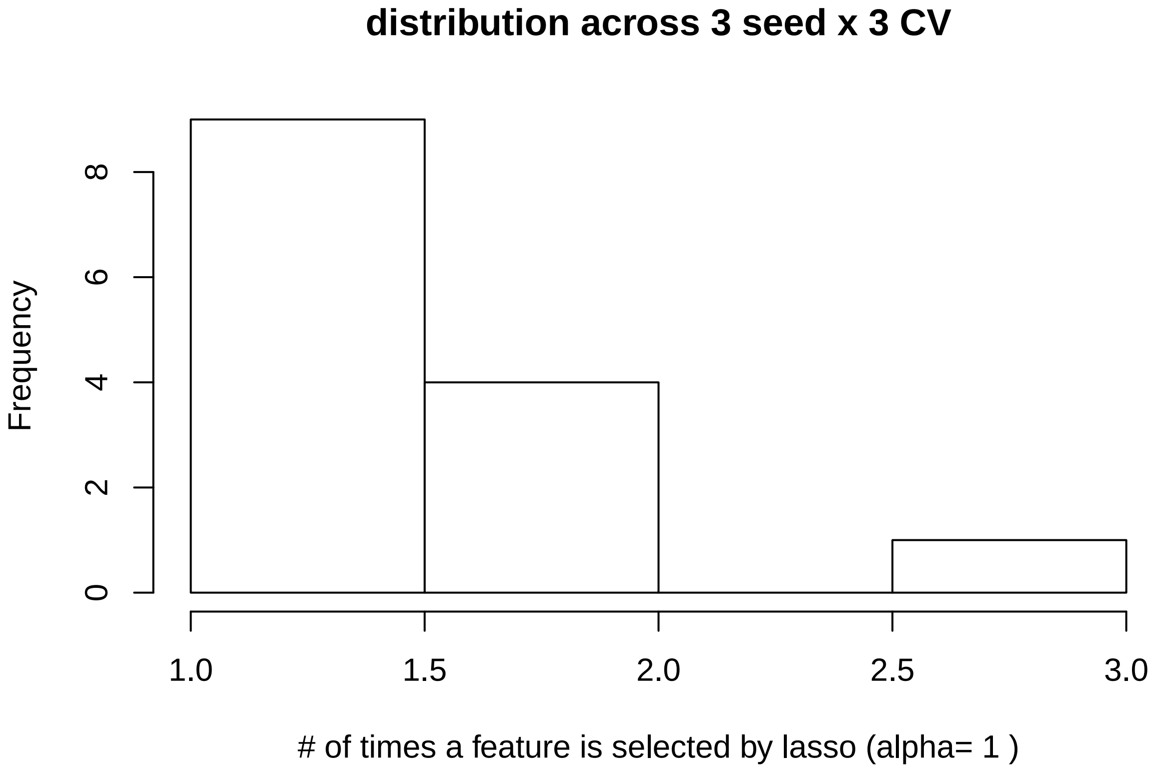
The following plot is to quickly see how well the prediction can separate long and short survivor.



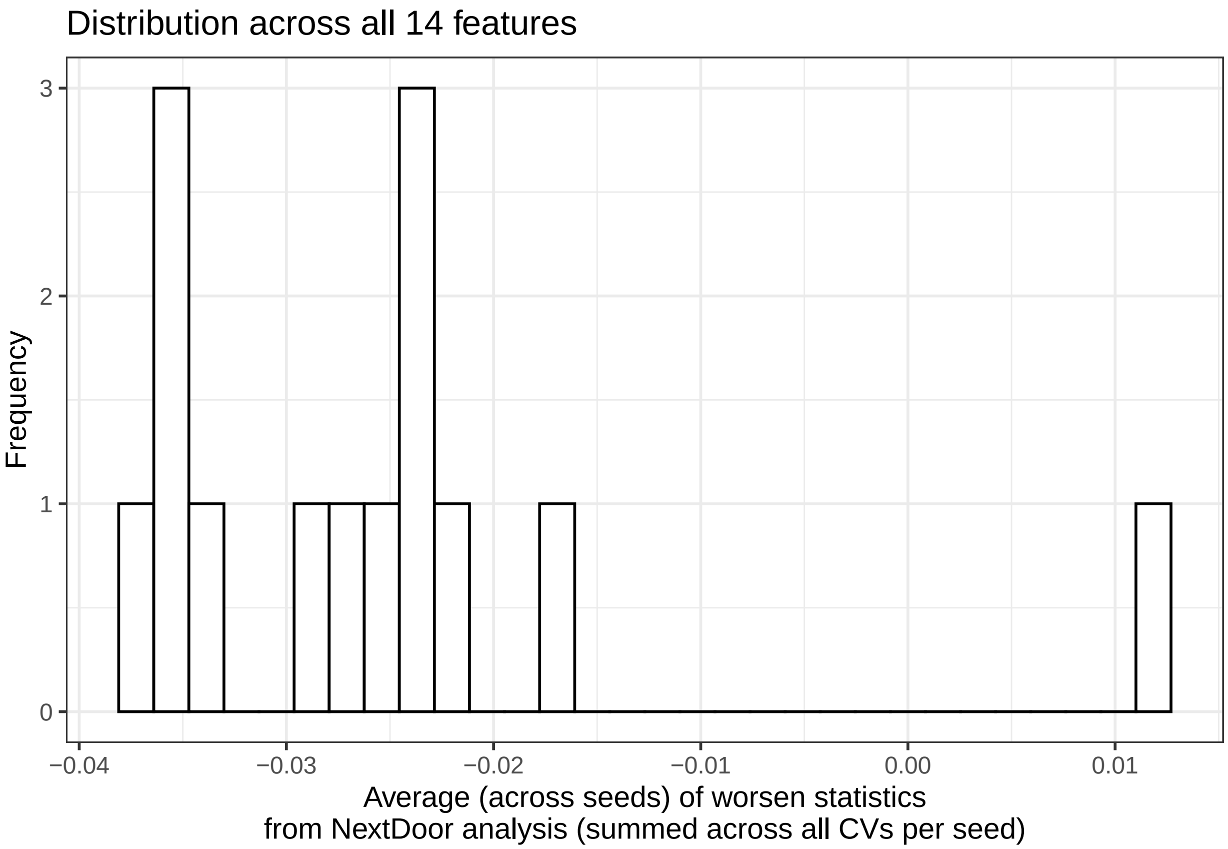
**prediction under seed 1001**

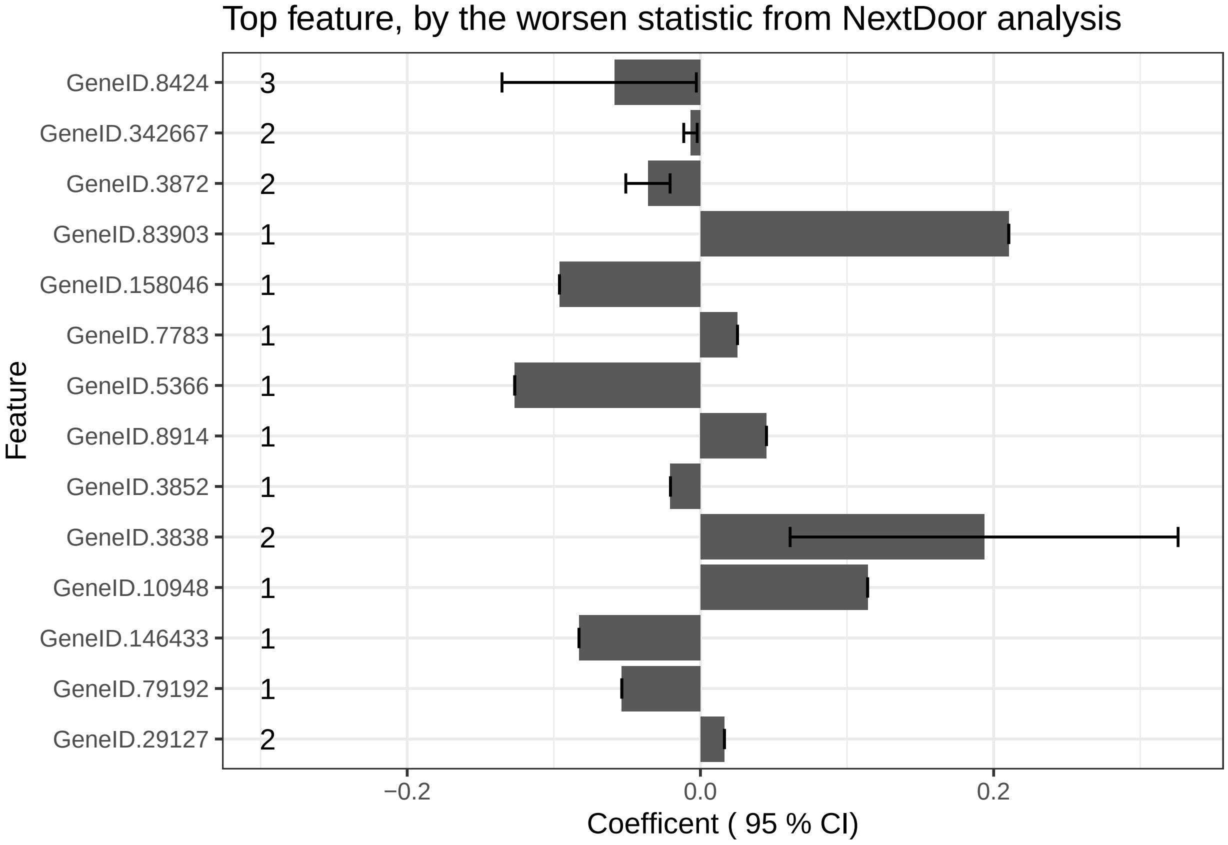


**2. Important Features**



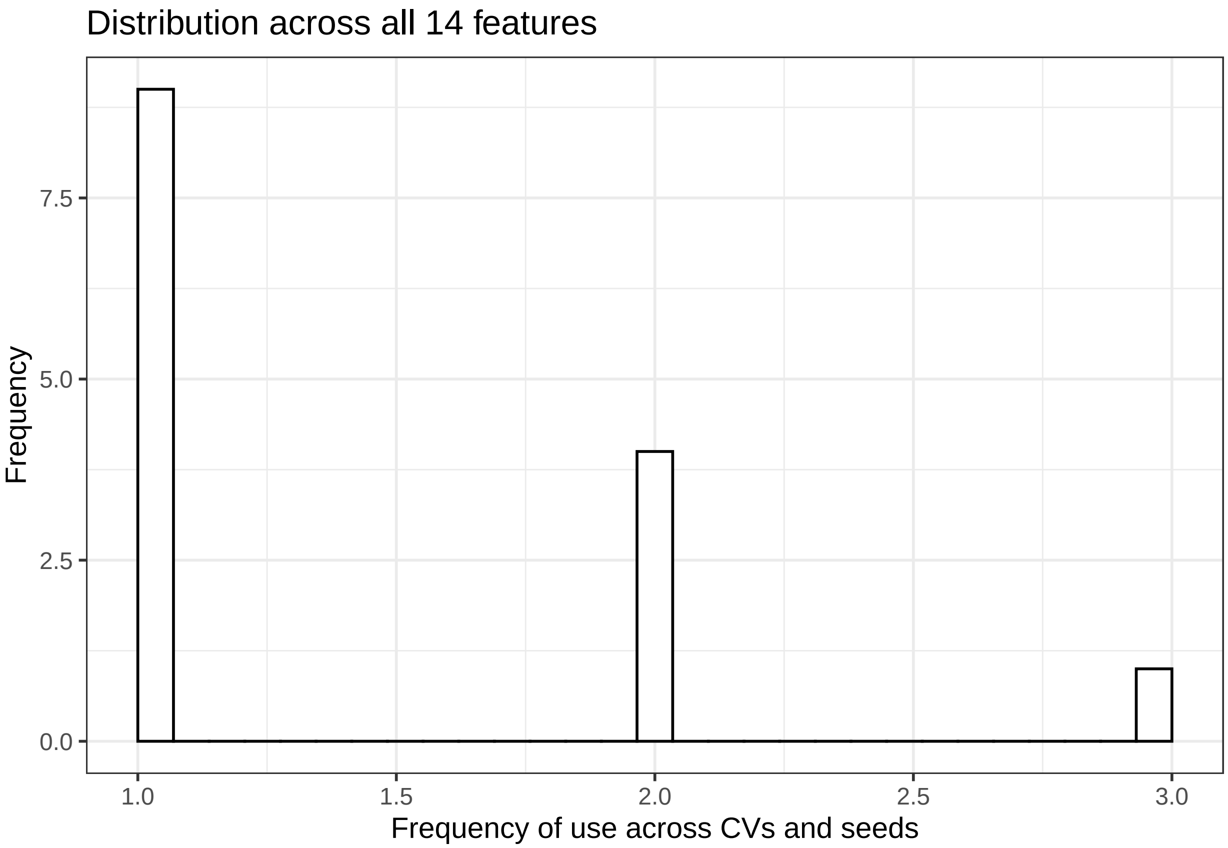
```
## [1] "there are 14 unique features used from the 100 feature set"
## [1] "summary of number of features used in each run under 3 seeds and 3 CVs"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   1.000   2.000   2.222   3.000   6.000

## [1] "there are 6 NA values in vimp before summation within seeds; they are imputed with the smallest
## [1] "there are 23 NA values in vimp after summation within seeds; they are imputed with the smallest

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

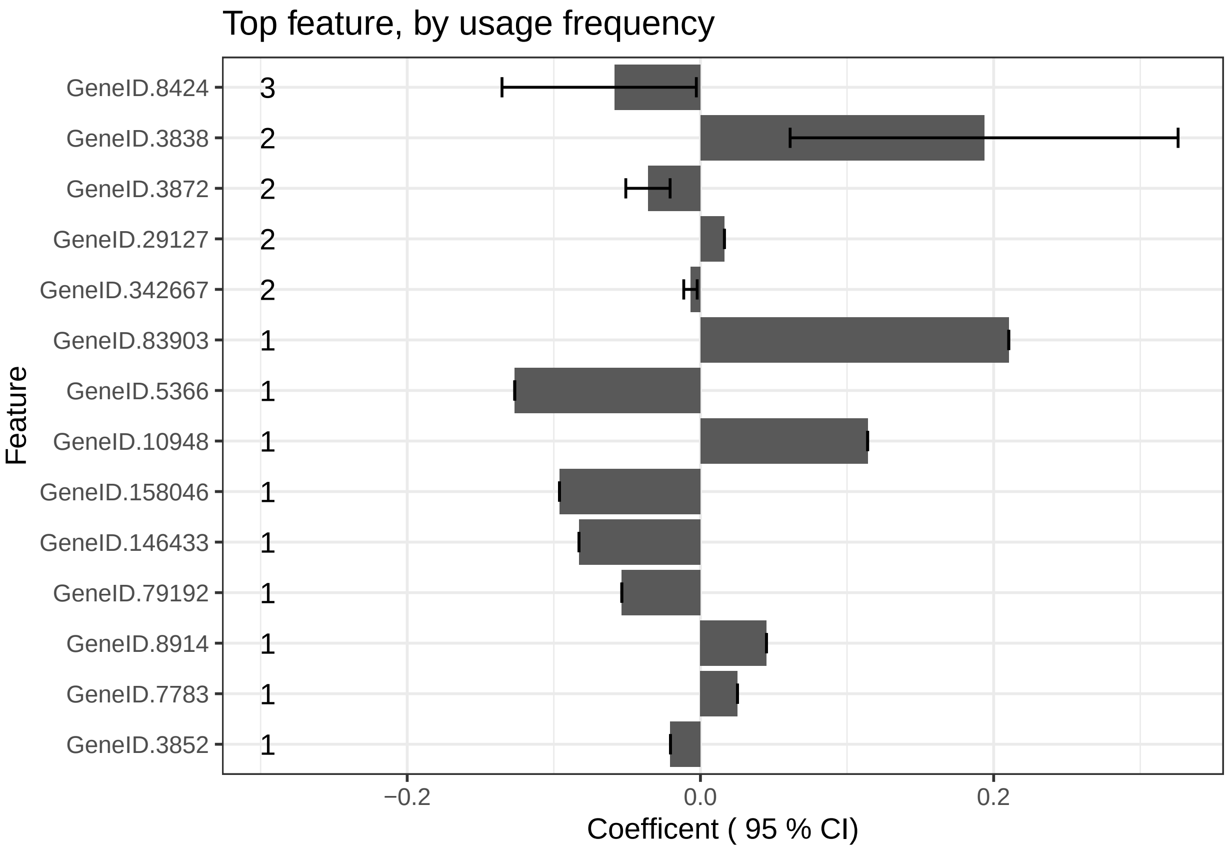

Distribution across all 14 features
Frequency
0
1
2
3
−0.04
−0.03
−0.02
−0.01
0.00
0.01
Average (across seeds) of worsen statistics
from NextDoor analysis (summed across all CVs per seed)

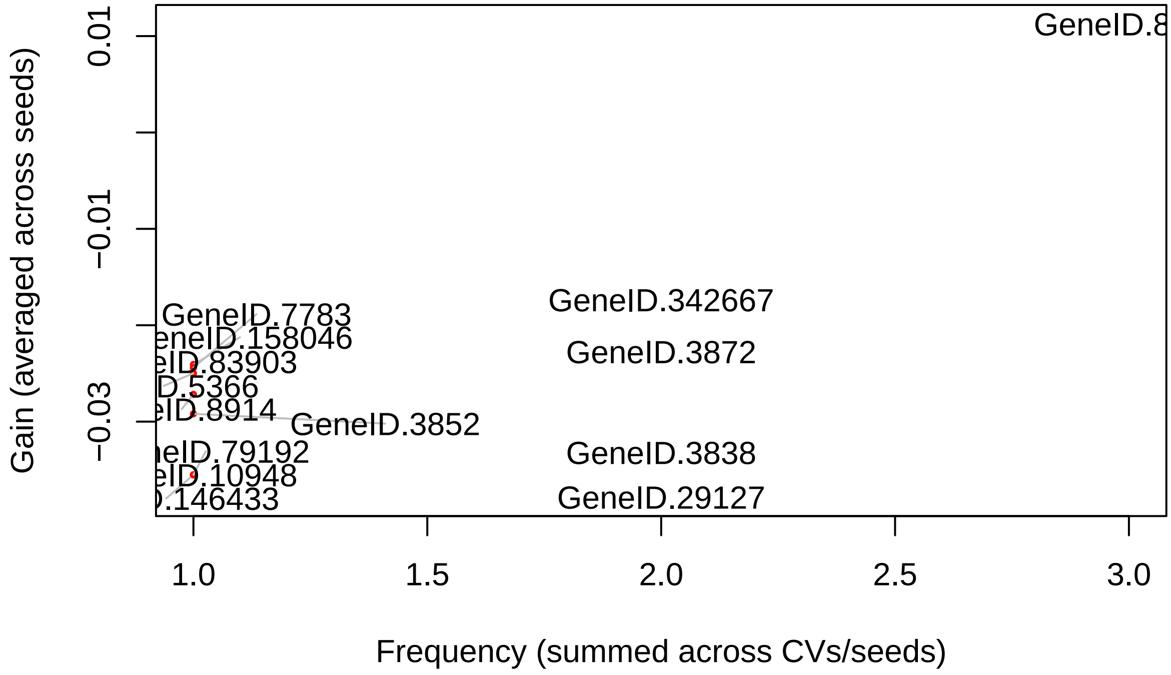

Top feature, by the worsen statistic from NextDoor analysis
Feature
GeneID.8424 3
GeneID.342667 2
GeneID.3872 2
GeneID.83903 1
GeneID.158046 1
GeneID.7783 1
GeneID.5366 1
GeneID.8914 1
GeneID.3852 1
GeneID.3838 2
GeneID.10948 1
GeneID.146433 1
GeneID.79192 1
GeneID.29127 2
−0.2
0.0
0.2
Coefficent ( 95 % CI)

```
## [1] "there are 6 NA values in vimp before summation within seeds; they are imputed with the smallest
## [1] "there are 23 NA values in vimp after summation within seeds; they are imputed with the smallest
```

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




Top feature, by usage frequency
Feature
GeneID.8424 3
GeneID.3838 2
GeneID.3872 2
GeneID.29127 2
GeneID.342667 2
GeneID.83903 1
GeneID.5366 1
GeneID.10948 1
GeneID.158046 1
GeneID.146433 1
GeneID.79192 1
GeneID.8914 1
GeneID.7783 1
GeneID.3852 1
−0.2
0.0
0.2
Coefficent ( 95 % CI)


Gain (averaged across seeds)
0.01
−0.01
−0.03
GeneID.8
GeneID.342667
GeneID.7783
eneID.158046
GeneID.3872
eID.83903
D.5366
eID.8914
GeneID.3852
neID.79192
GeneID.3838
eID.10948
).146433
GeneID.29127
1.0
1.5
2.0
2.5
3.0
Frequency (summed across CVs/seeds)