

Evaluate testing data (binary-class) - Lasso

EVE W.

2020-07-20

Contents

0. Load Data	1
1. Scores	2
1.1 Scores per Class	2
1.2 Average score	3
2. Important Features	4

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

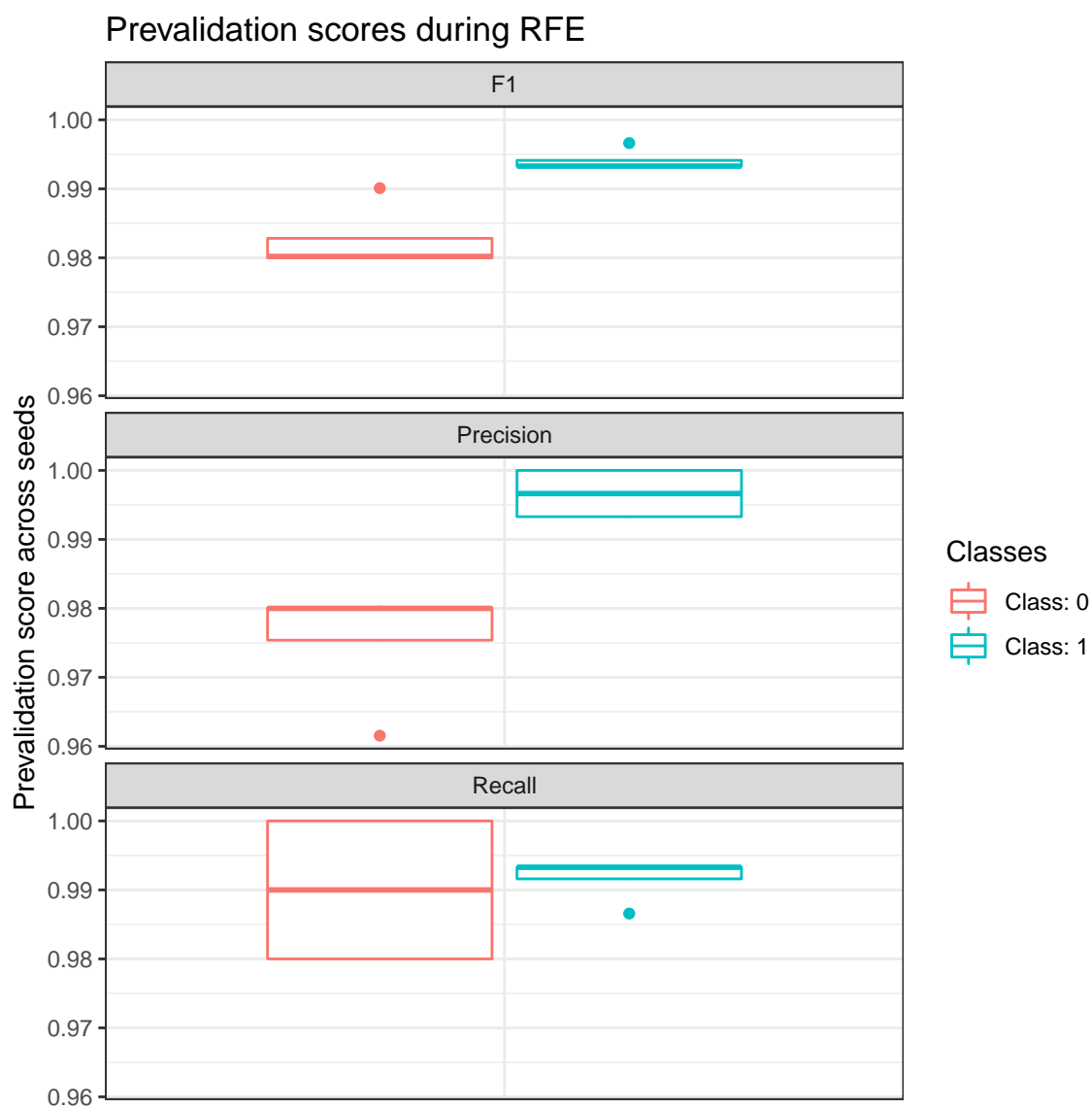
```
## user input
project_home <- "~/EVE/examples"
project_name <- "lasso_binary_outCV_test"
```

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

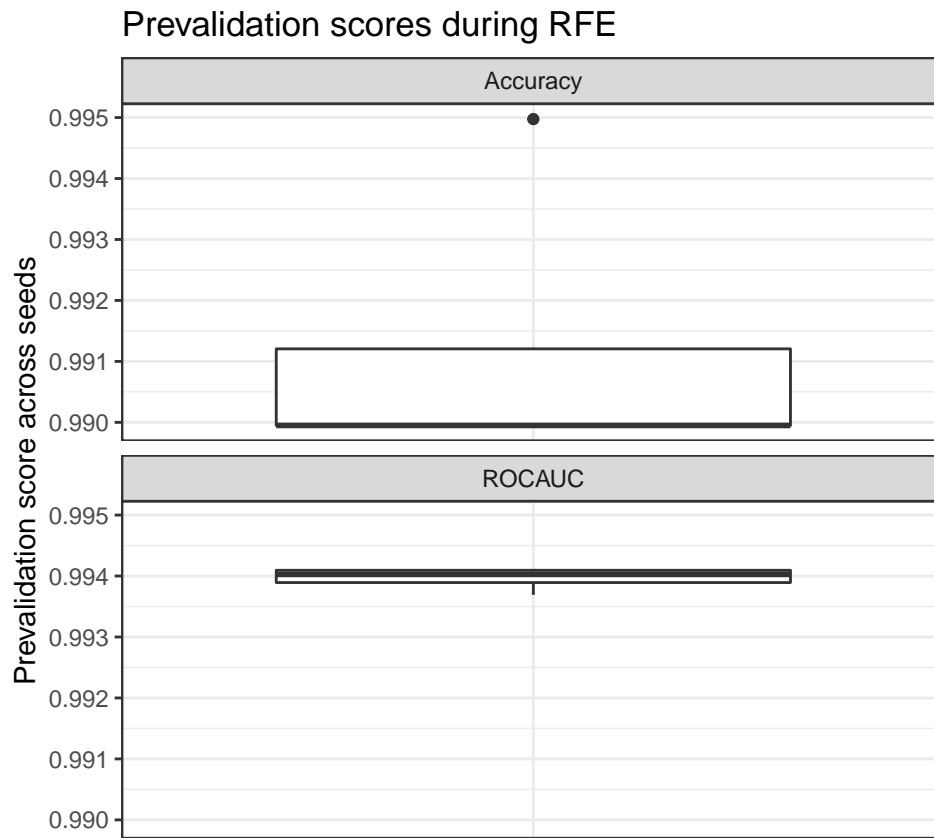


Confusion Matrix

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

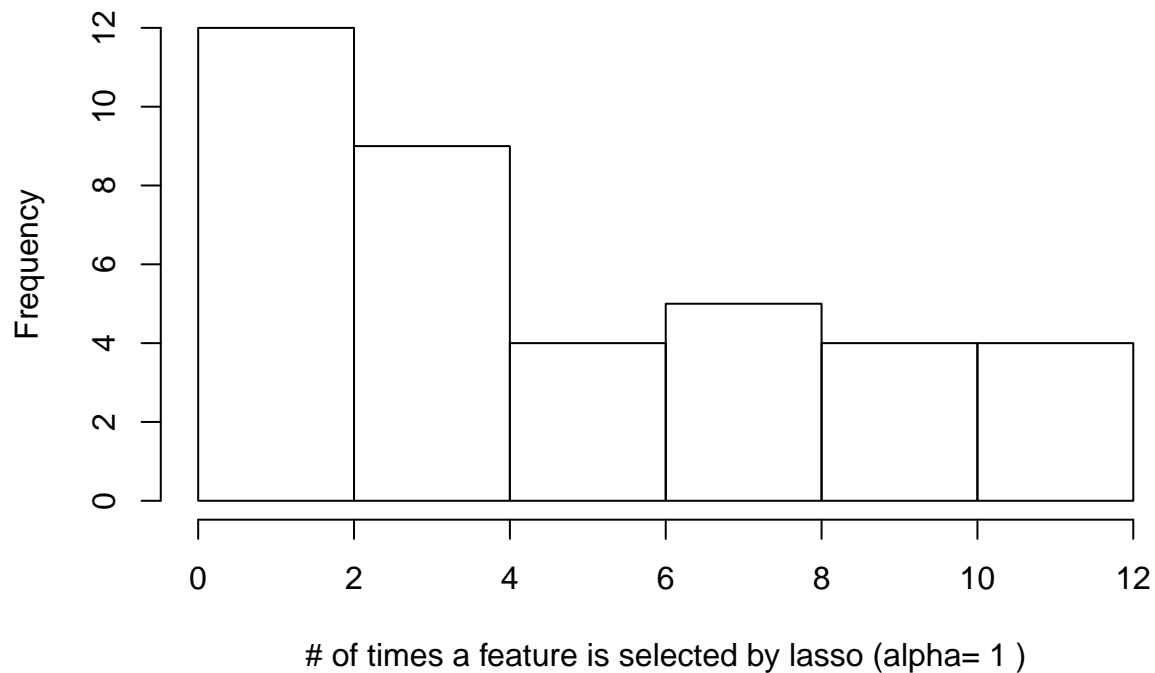
##      Reference
## Prediction  0   1
##      0 198   5
##      1   2 591
```

1.2 Average score



2. Important Features

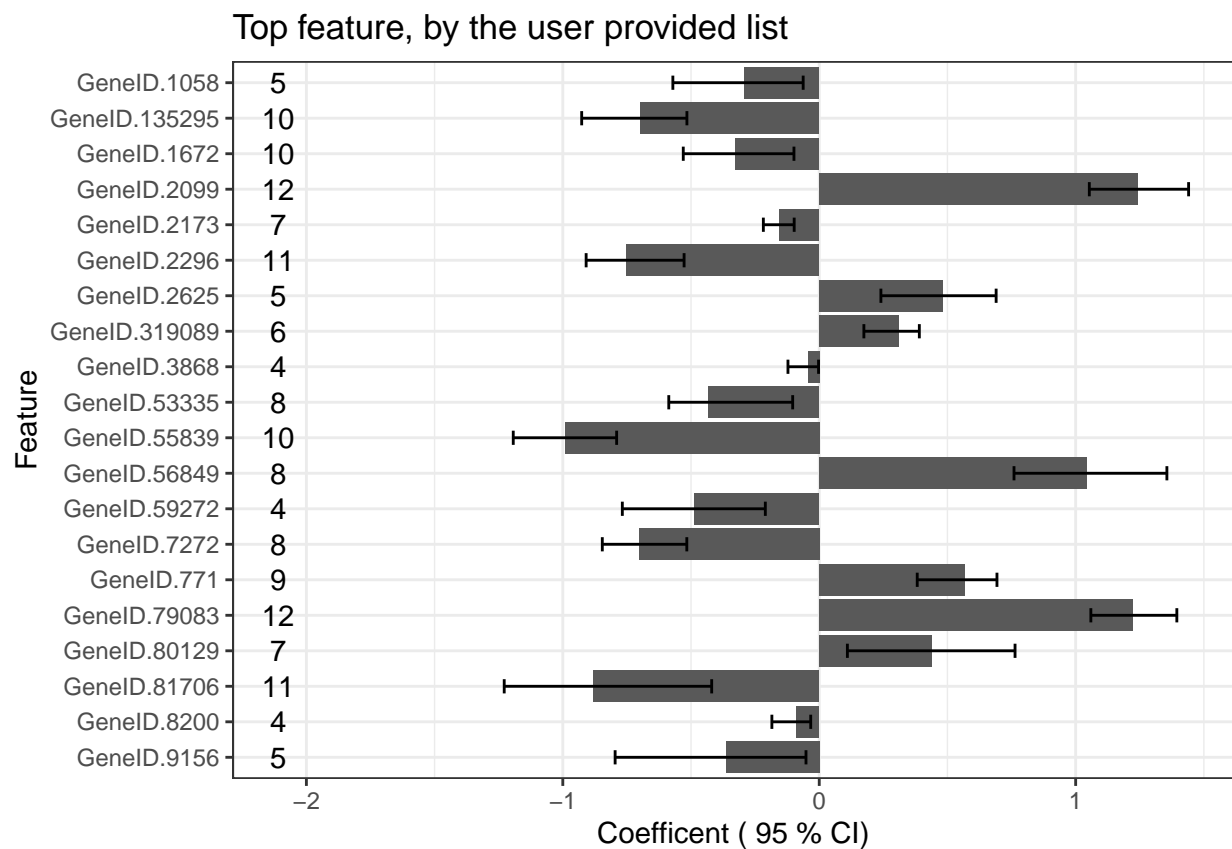
distribution across 4 seed x 3 CV



```
## [1] "there are 38 unique features used from the 100 feature set"
## summary of number of features used in 4 seeds and 3 CVs

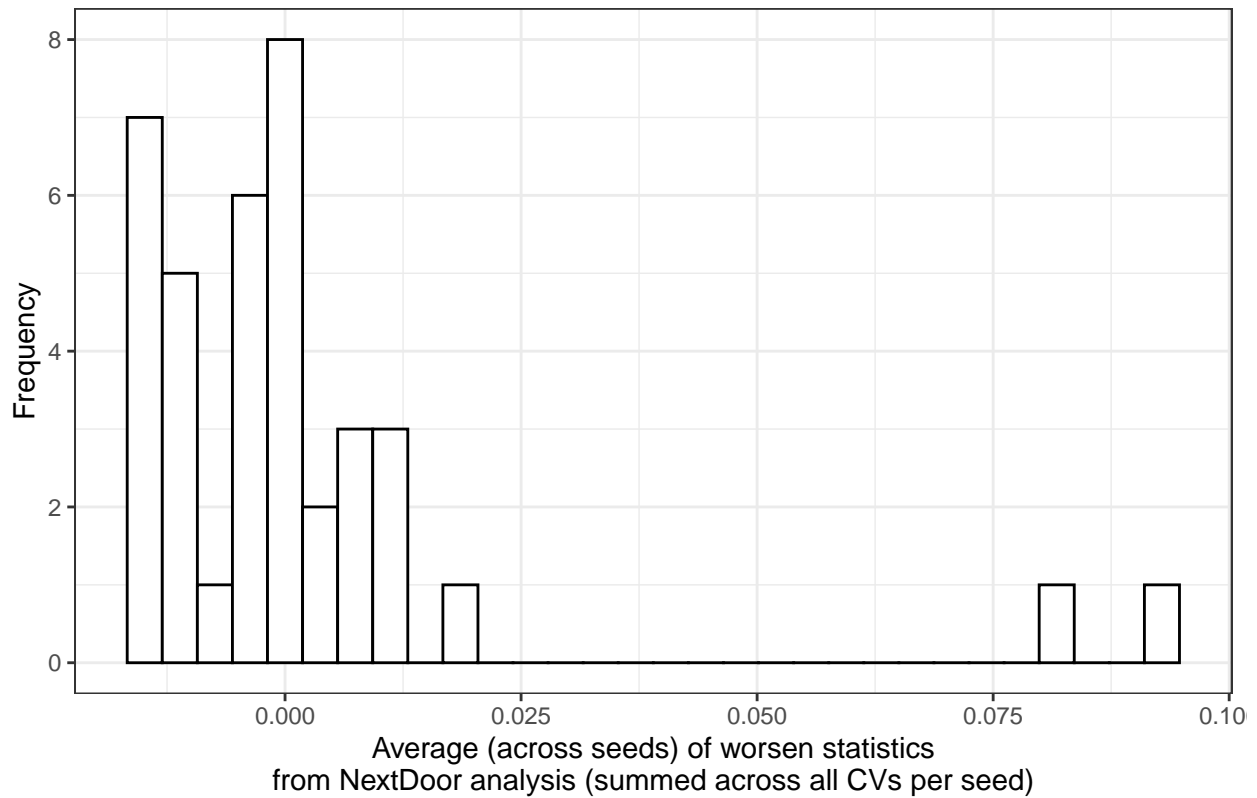
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  13.00  14.75   16.50   16.08  17.25   18.00

## [1] "there are 37 NA values in vimp after summation within seeds; they are imputed with the smallest
```

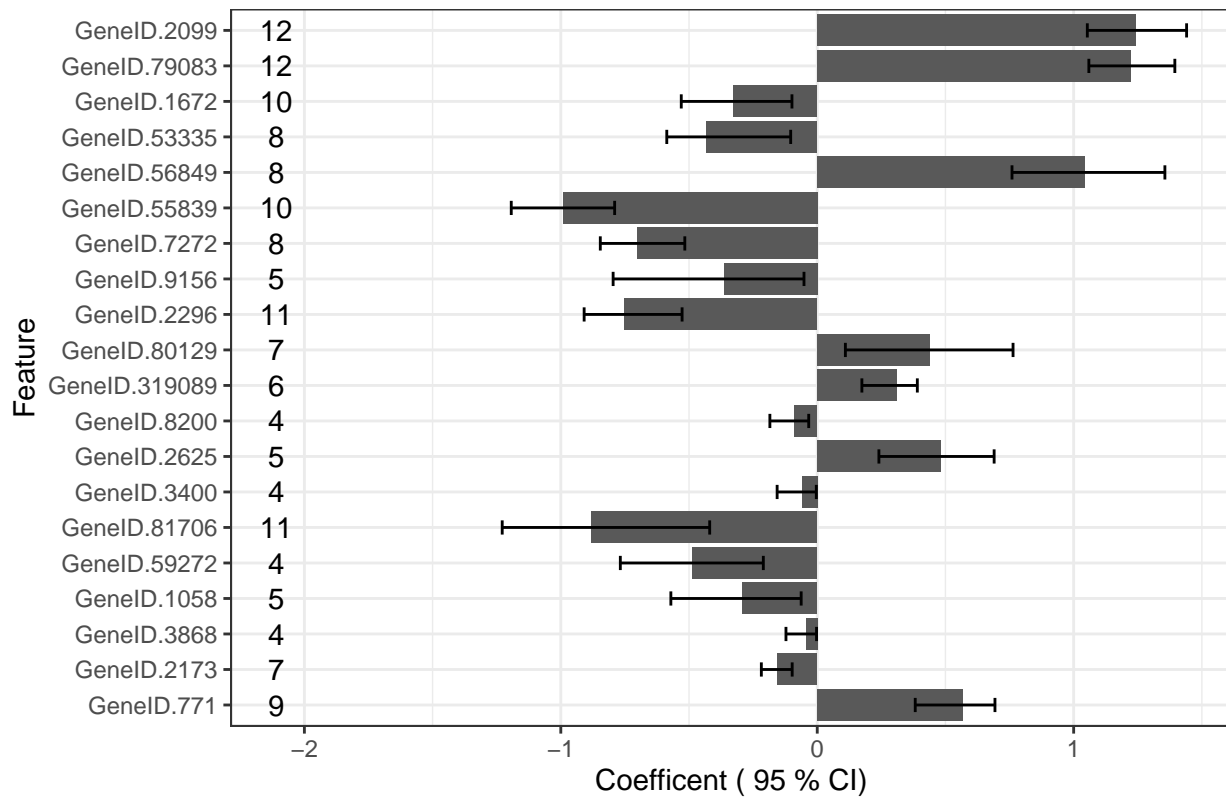


```
## [1] "there are 37 NA values in vimp after summation within seeds; they are imputed with the smallest
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`."
```

Distribution across all 38 features

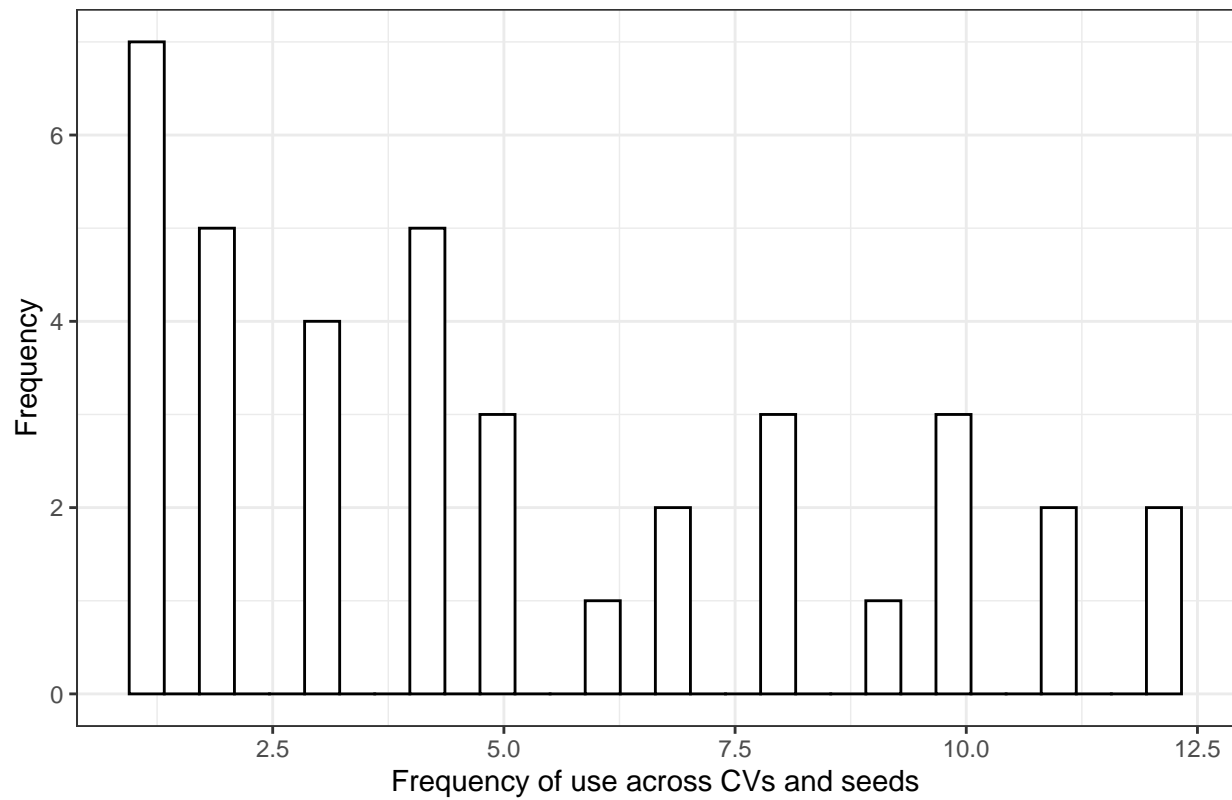


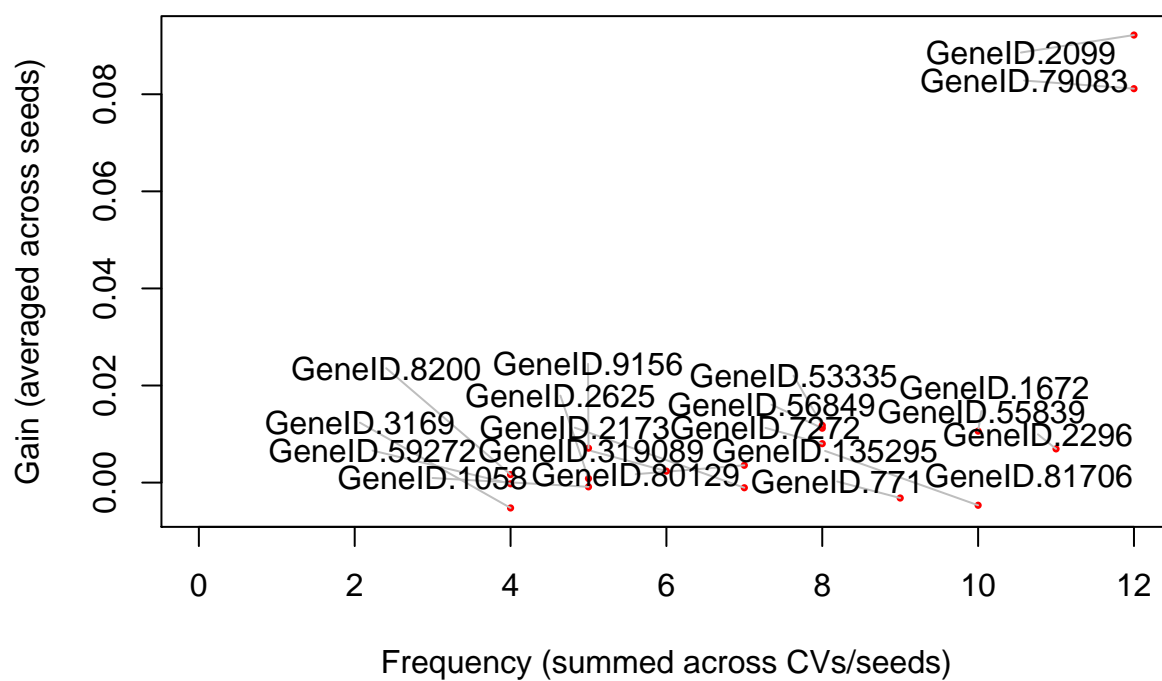
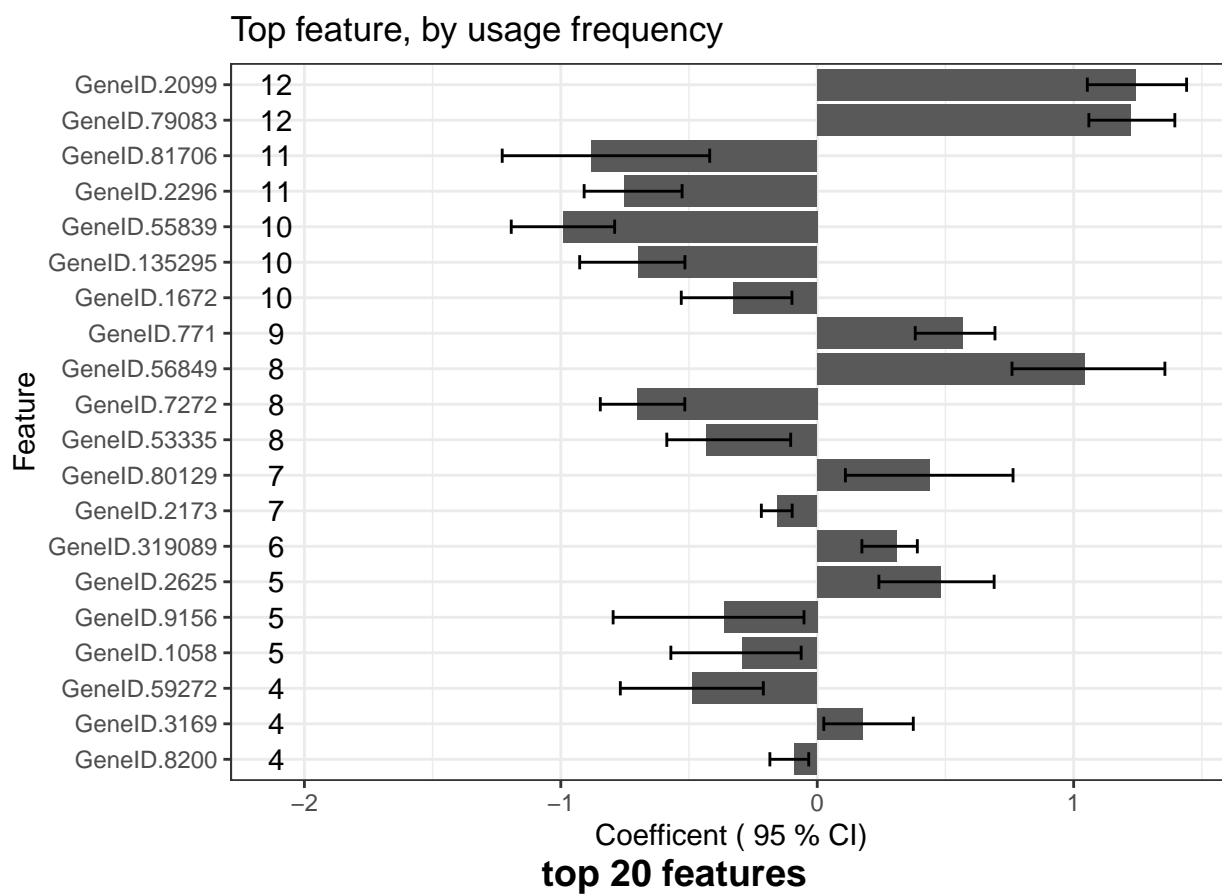
Top feature, by the worsen statistic from NextDoor analysis



```
## [1] "there are 37 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 38 features





Heatmap of top 20 important features

