Evaluate testing data (regression) - Lasso

EVE W.

2020-07-20
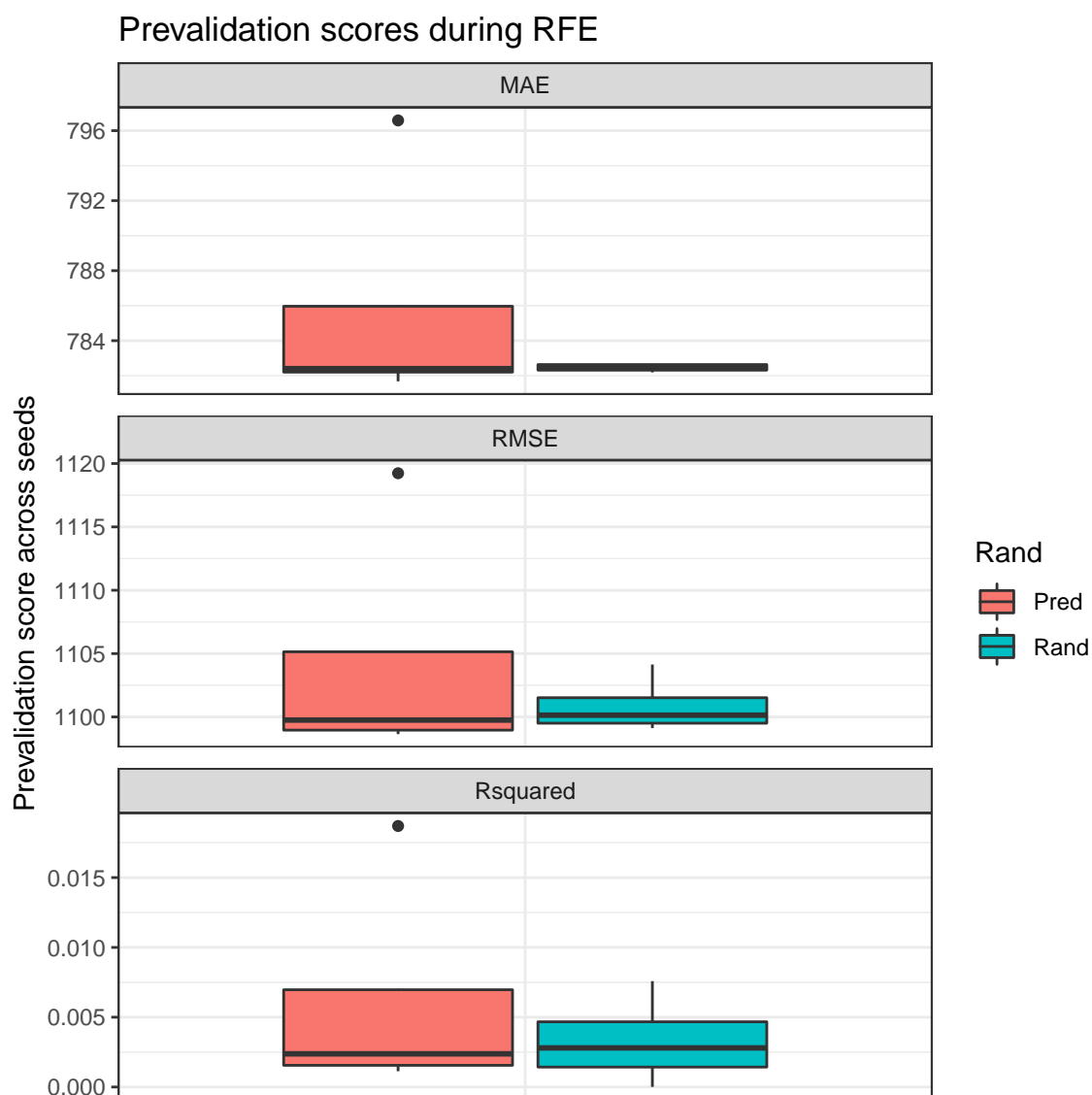
Contents

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

0. Load Data

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

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

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

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

```
## os_time :
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   182.8   480.0   889.4  1221.2  7125.0
```

run with lasso.r with alpha = 0.01.

1. Scores

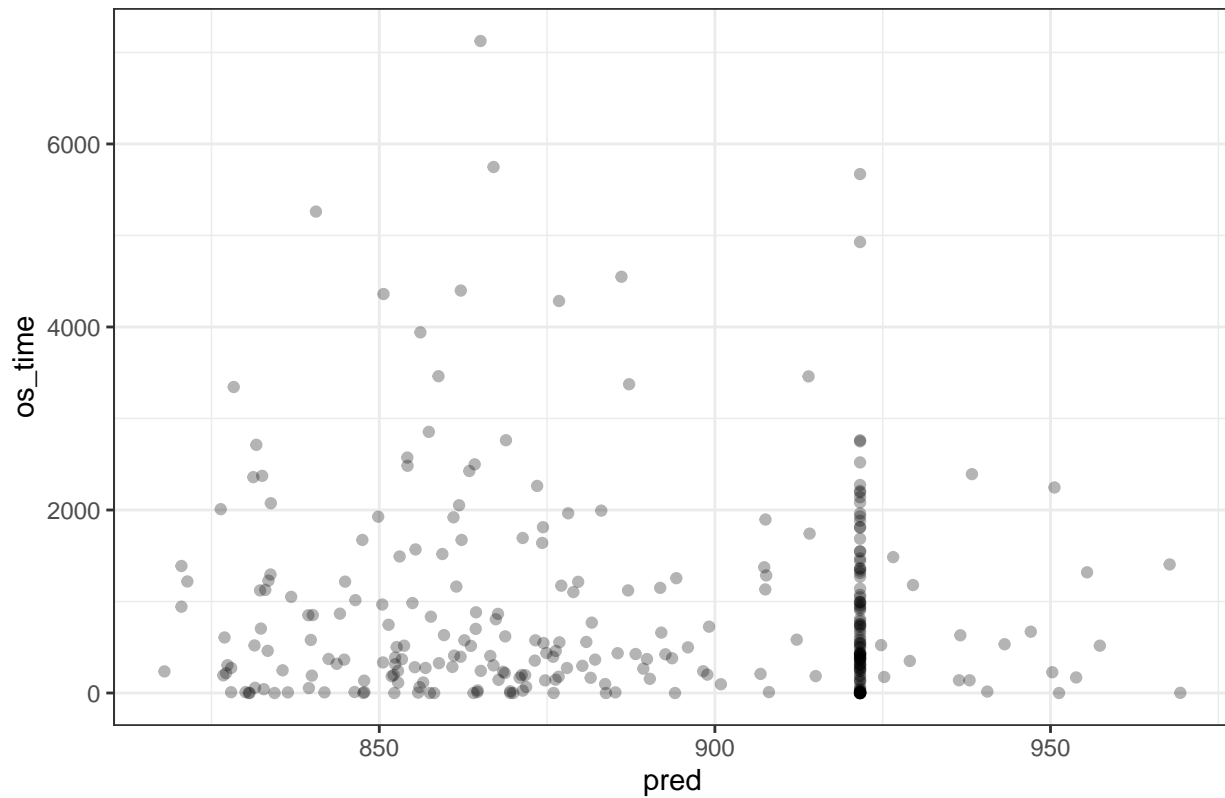## Prevalidation scores during RFE

'Pred' compares the actual CV prediction with observed value. 'Rand' compares permuted CV prediction with observed to mimic random prediction.
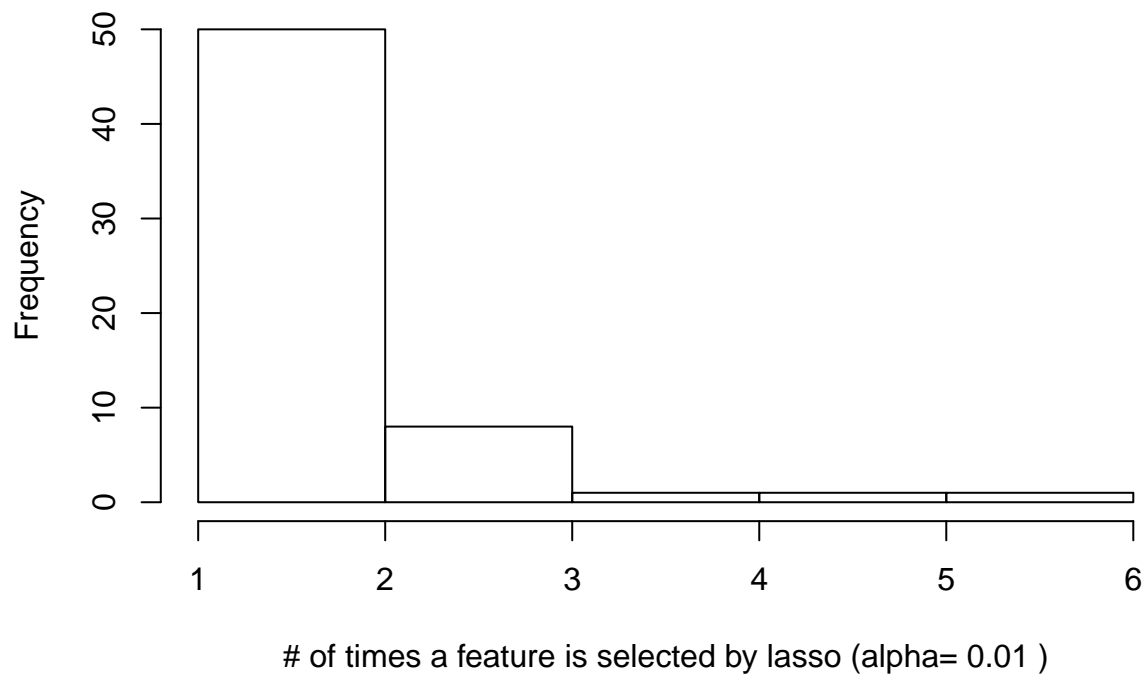
correlation

```
##
##
## Table: Averaged pearson correlation across seeds
##
##     cor.avg      cor.sdt
## -----------   ----------
##  -0.0667016    0.0475382
```

Correlation at seed = 1003 using 100 feature set input

2. Important Features

distribution across 4 seed x 3 CV

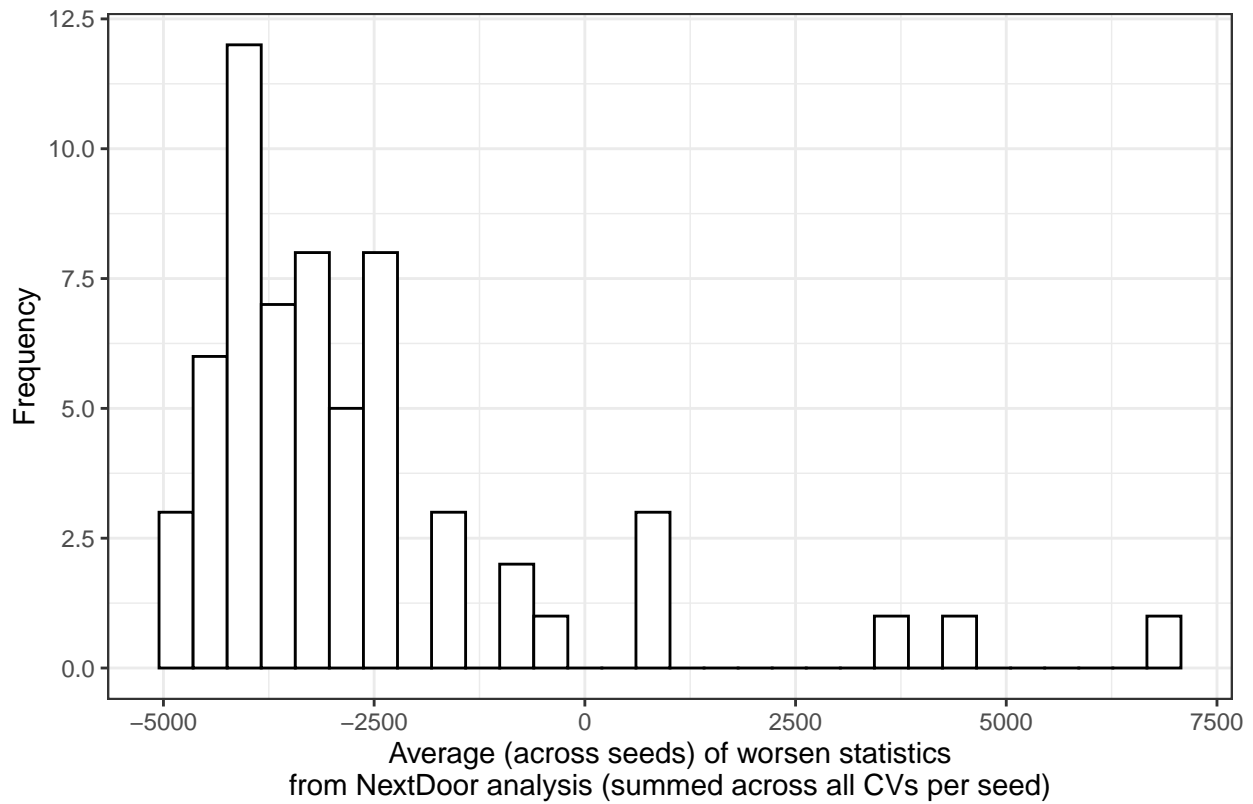# of times a feature is selected by lasso (alpha= 0.01)

```
## [1] "there are 61 unique features used from the 100 feature set"
## [1] "summary of number of features used in each run under 4 seeds and 3 CVs"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    8.25   12.00   17.17   18.75   50.00       6

## [1] "there are 151 NA values in vimp after summation within seeds; they are imputed with the smalles

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
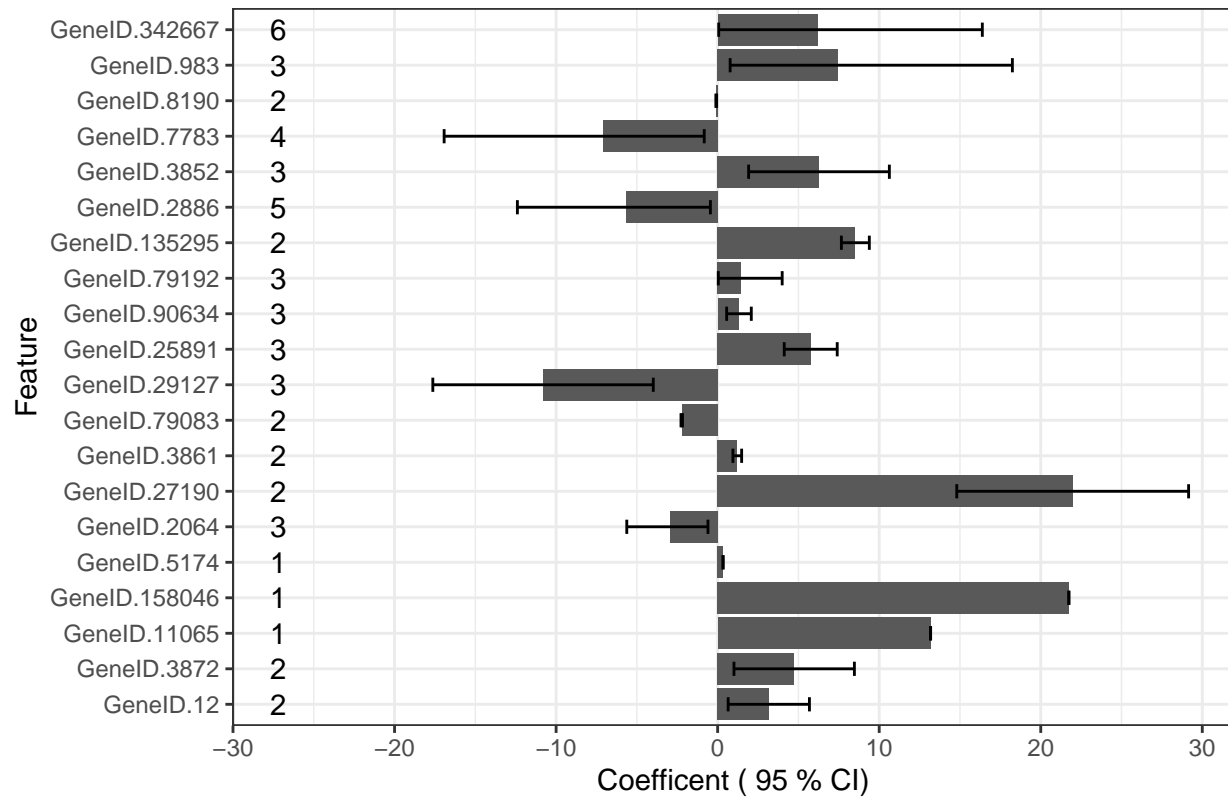
Distribution across all 61 features

Average (across seeds) of worsen statistics
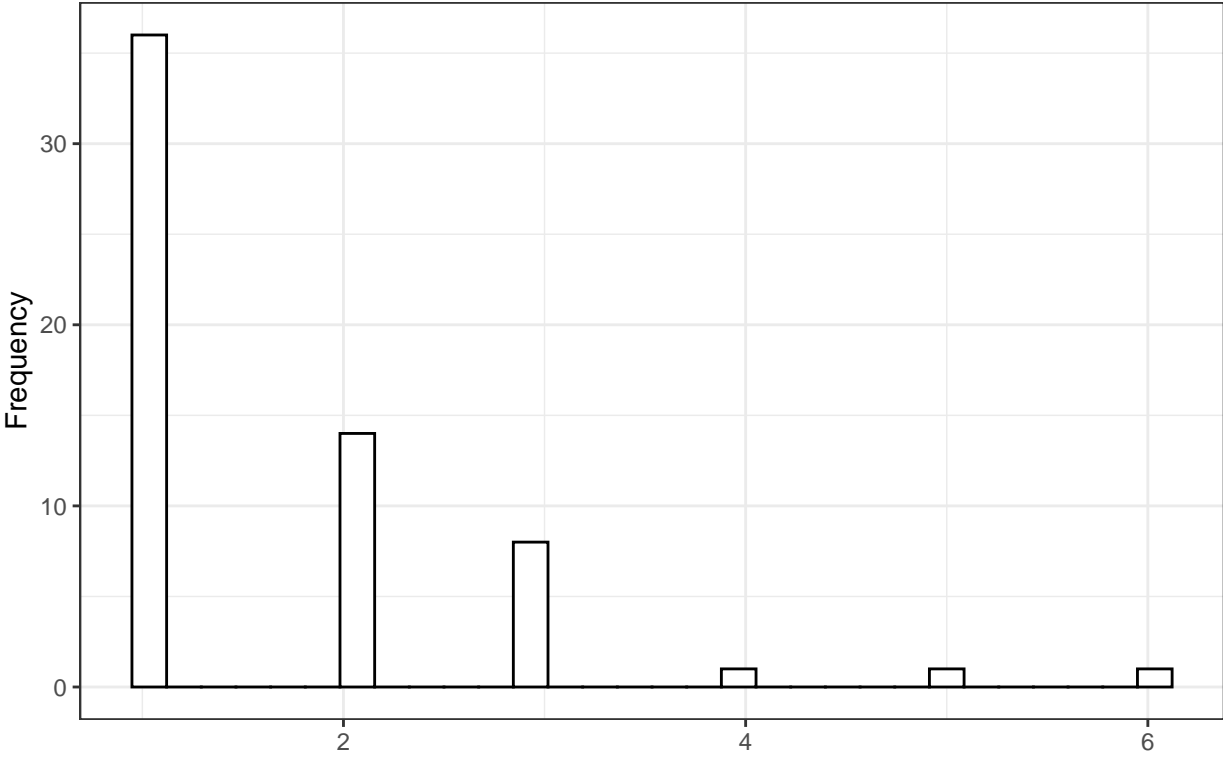from NextDoor analysis (summed across all CVs per seed)

Top feature, by the worsen statistic from NextDoor analysis

[1] "there are 151 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 61 features

Top feature, by usage frequency

Gain (averaged across seeds)

GeneID.34
GeneID.983
GeneID.3852   GeneID.7783   GeneID.2886
ID.135295   GeneID.90634
GeneID.79192
3 GeneID.12 GeneID.25891
eneID.3872   GeneID.29127
eID.27190   GeneID.2064
ID.149461
eID.8914
.8424   GeneID.11162

Frequency (summed across CVs/seeds)

Heatmap of top 20 important features

os_time

GeneID.90634
GeneID.12
GeneID.79083
GeneID.5174
GeneID.158046
GeneID.135295
GeneID.8190
GeneID.79192
GeneID.29127
GeneID.983
GeneID.11065
GeneID.25891
GeneID.27190
GeneID.342667
GeneID.3861
GeneID.3852
GeneID.3872
GeneID.7783
GeneID.2886
GeneID.2064

os_time
7000
0

vimp
6000
−2000