

# Evaluate testing data (survival) - lasso

EVE W.

2020-05-07

## Contents

0. Load Data . . . . .	1
1. Scores . . . . .	2
2. Important Features . . . . .	4

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

## 0. Load Data

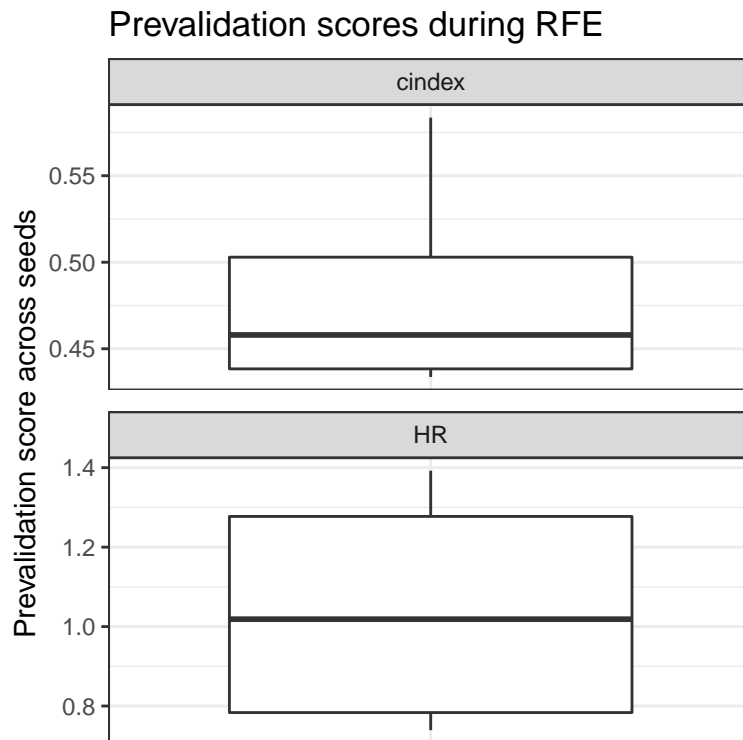
300 of samples were used

100 of full features

4 runs, each run contains 3 CVs.

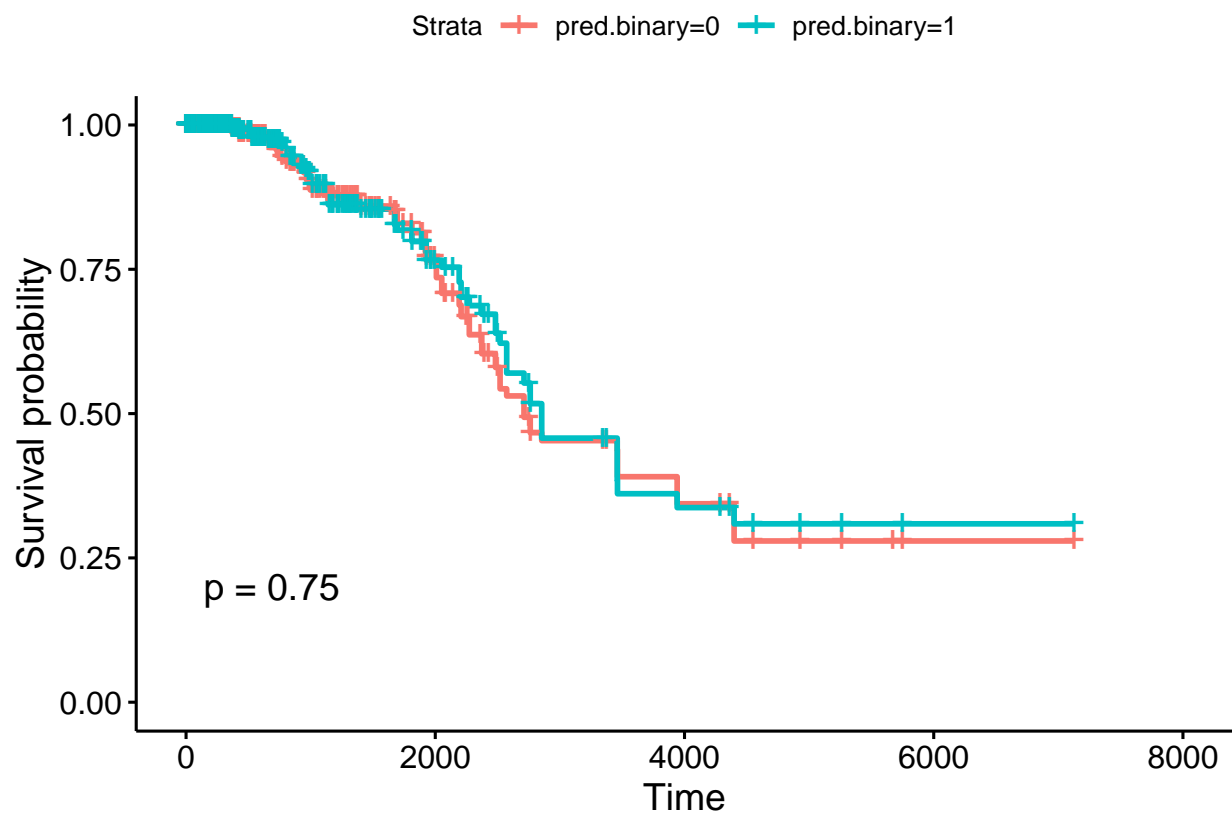
run with lasso.r with  $\alpha = 1$ .

## 1. Scores

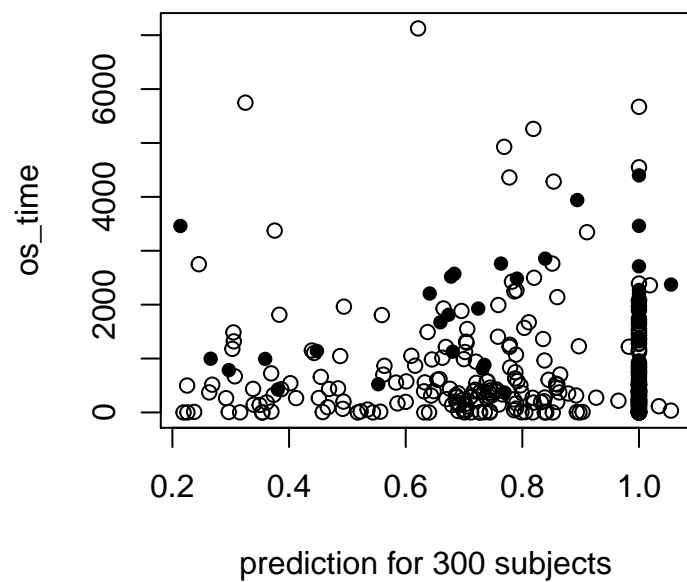


Note for the **HR plot**: A HR value (per seed) is calculated by comparing the survival time between 'long' and 'short' survivors. These two groups are defined by splitting samples based on *median* predicted risk score; group\_0 is predicted risk scores > median, which can be viewed as 'short survivors'. 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.

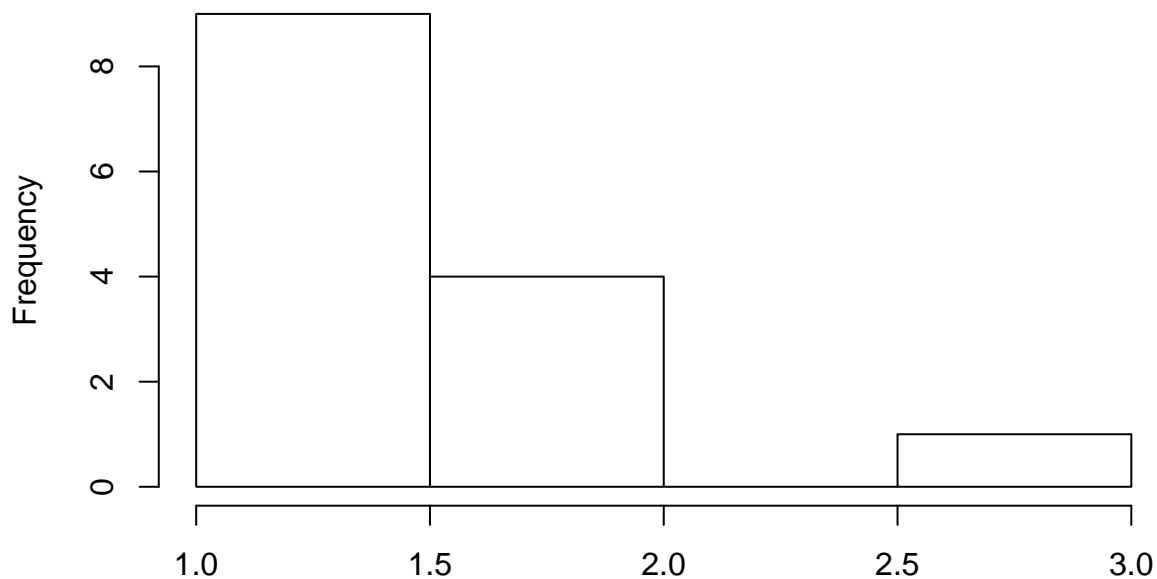


**prediction under seed 1001**



## 2. Important Features

### distribution across 3 seed x 3 CV



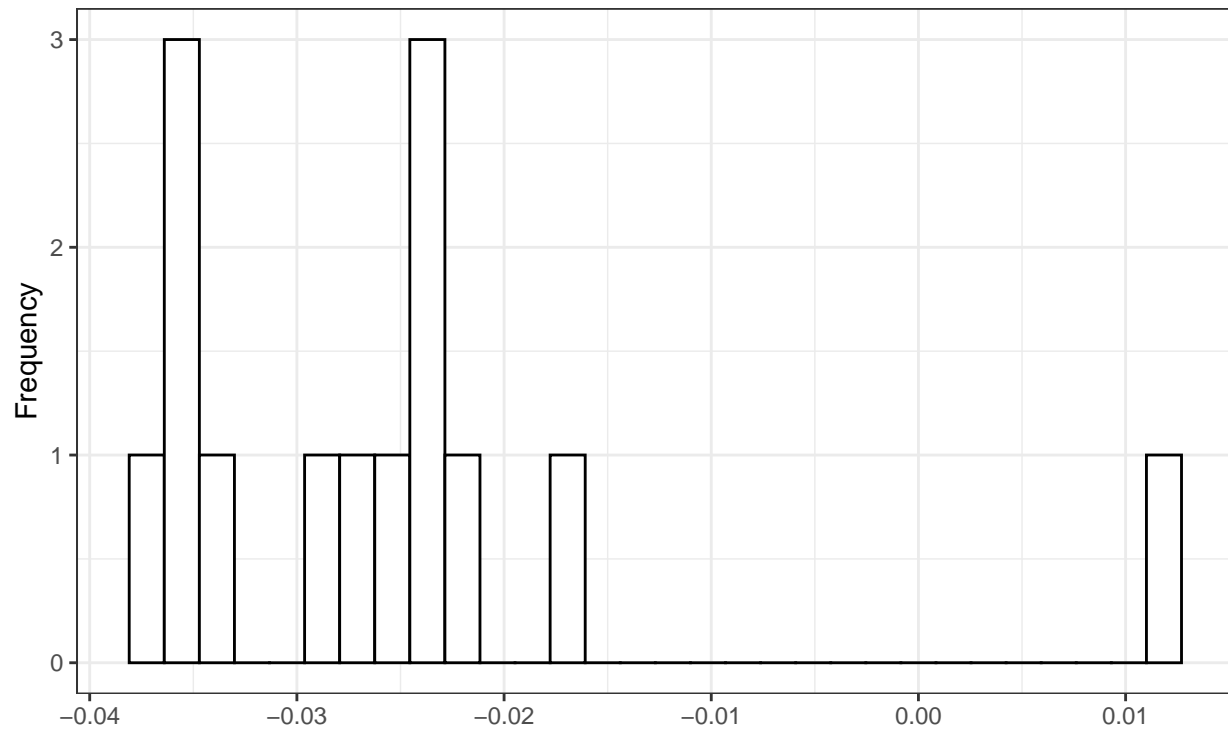
# of times a feature is selected by lasso (alpha= 1 )

```
## [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"

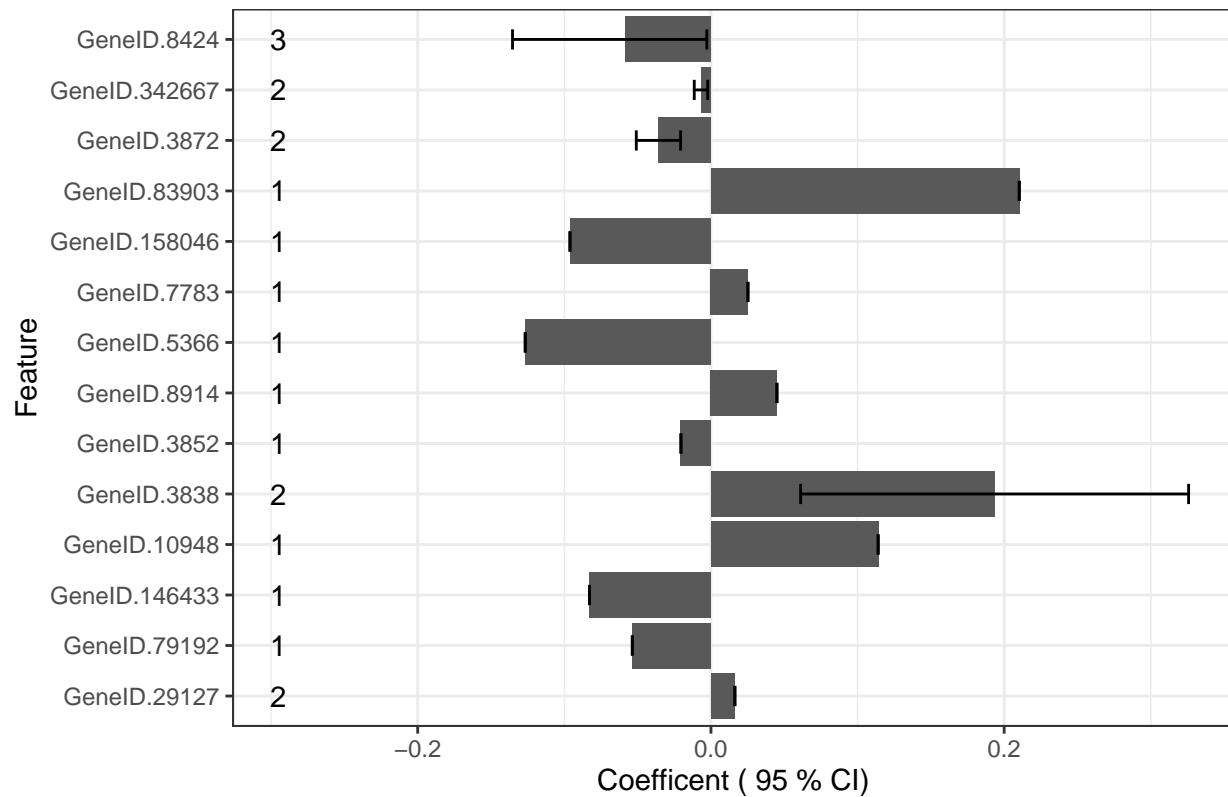
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000  1.000   2.000   2.222  3.000   6.000

## [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



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 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

