Evaluate testing data (survival) - lasso

EVE W.

2020-07-25

Contents

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

0. Load Data

```
300 of samples were used
```
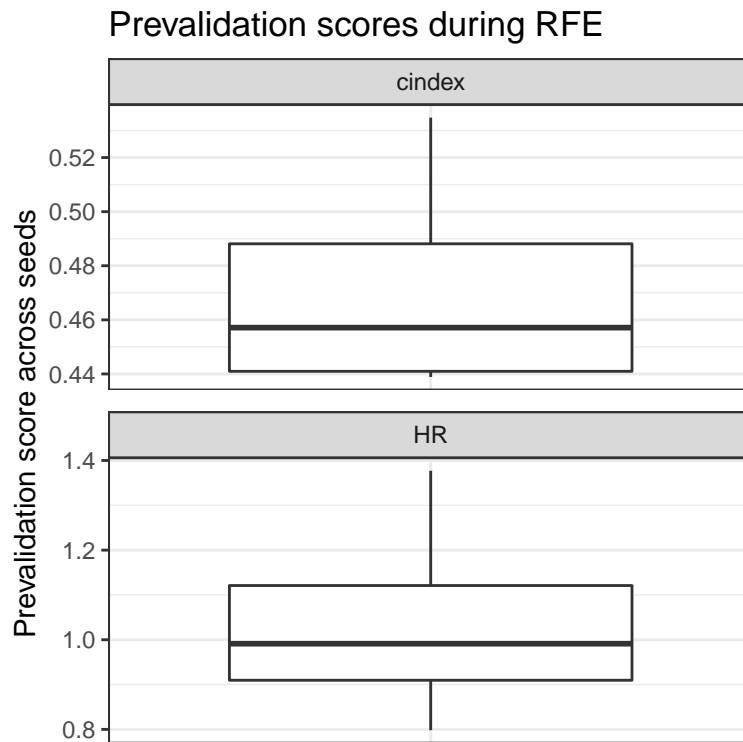
```
100 of full features
```

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

run with lasso.r with alpha = 1.

1. Scores

Prevalidation scores during RFE



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.
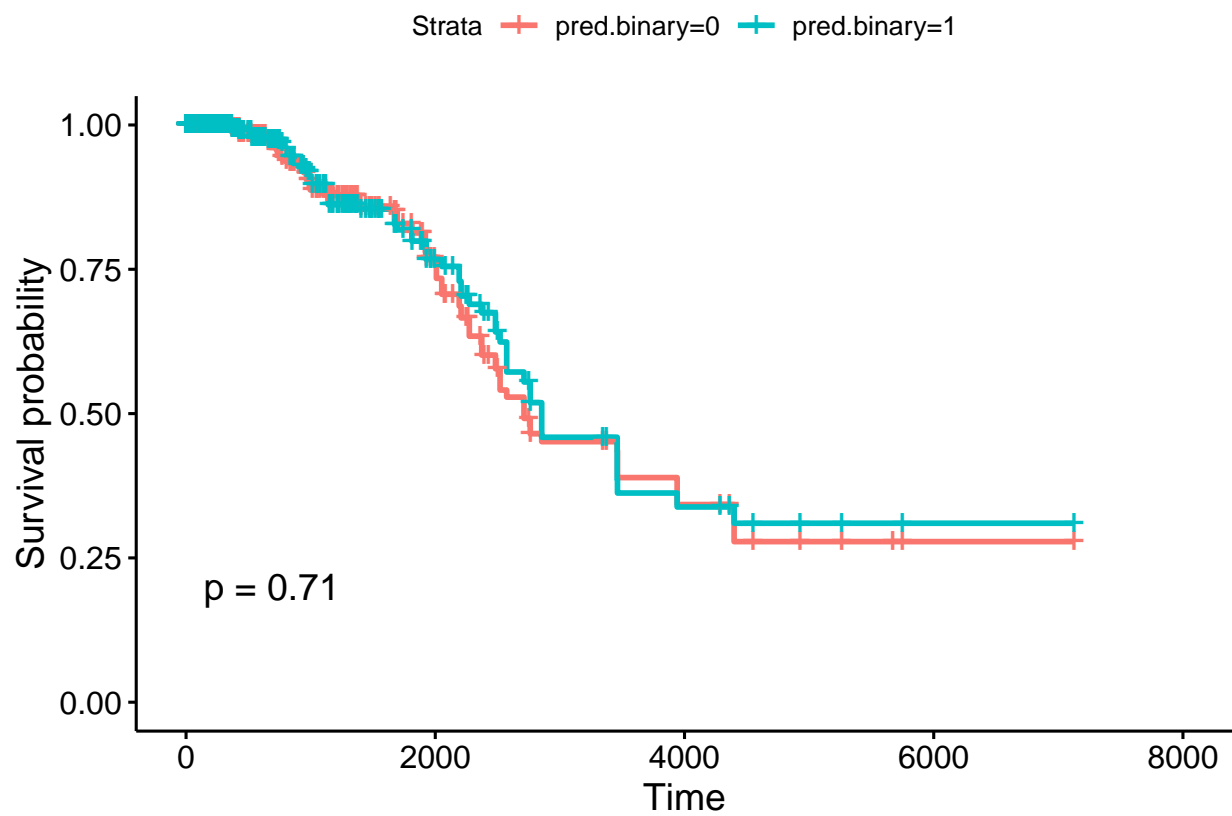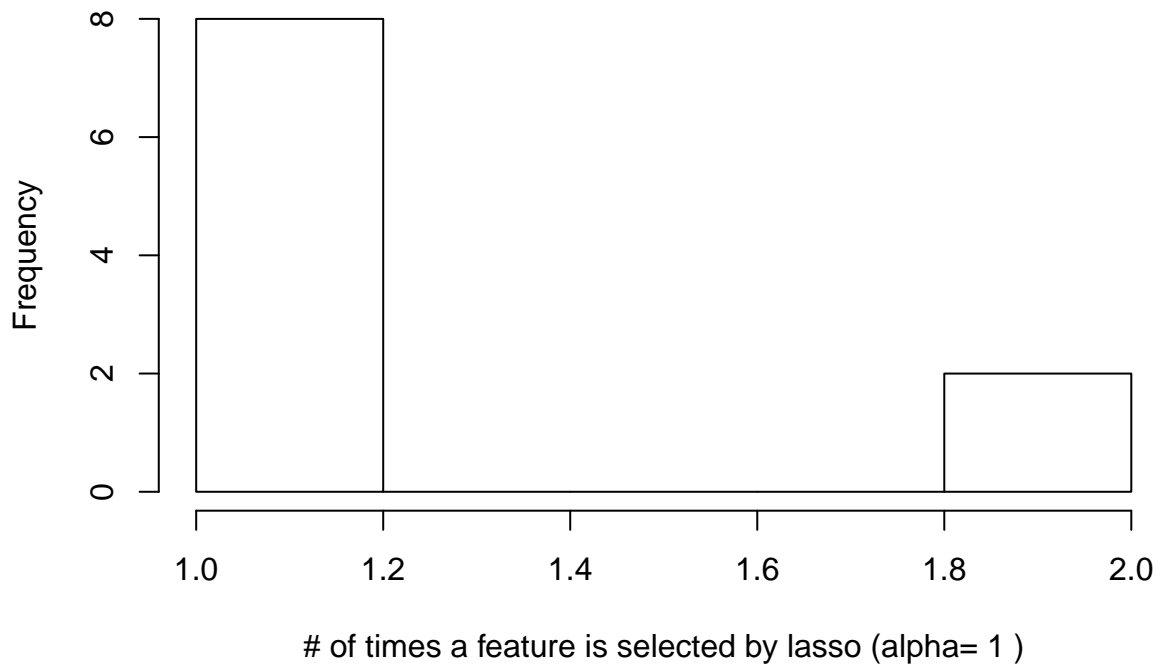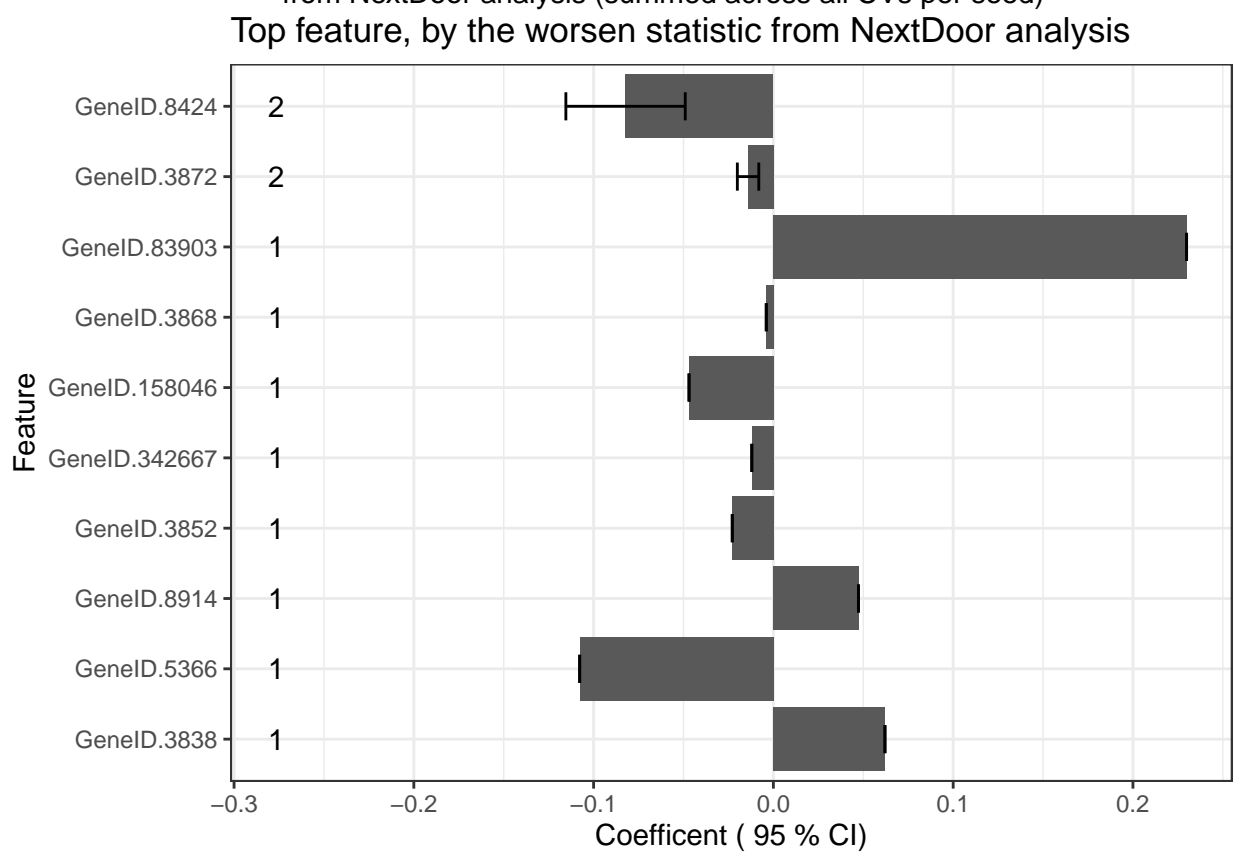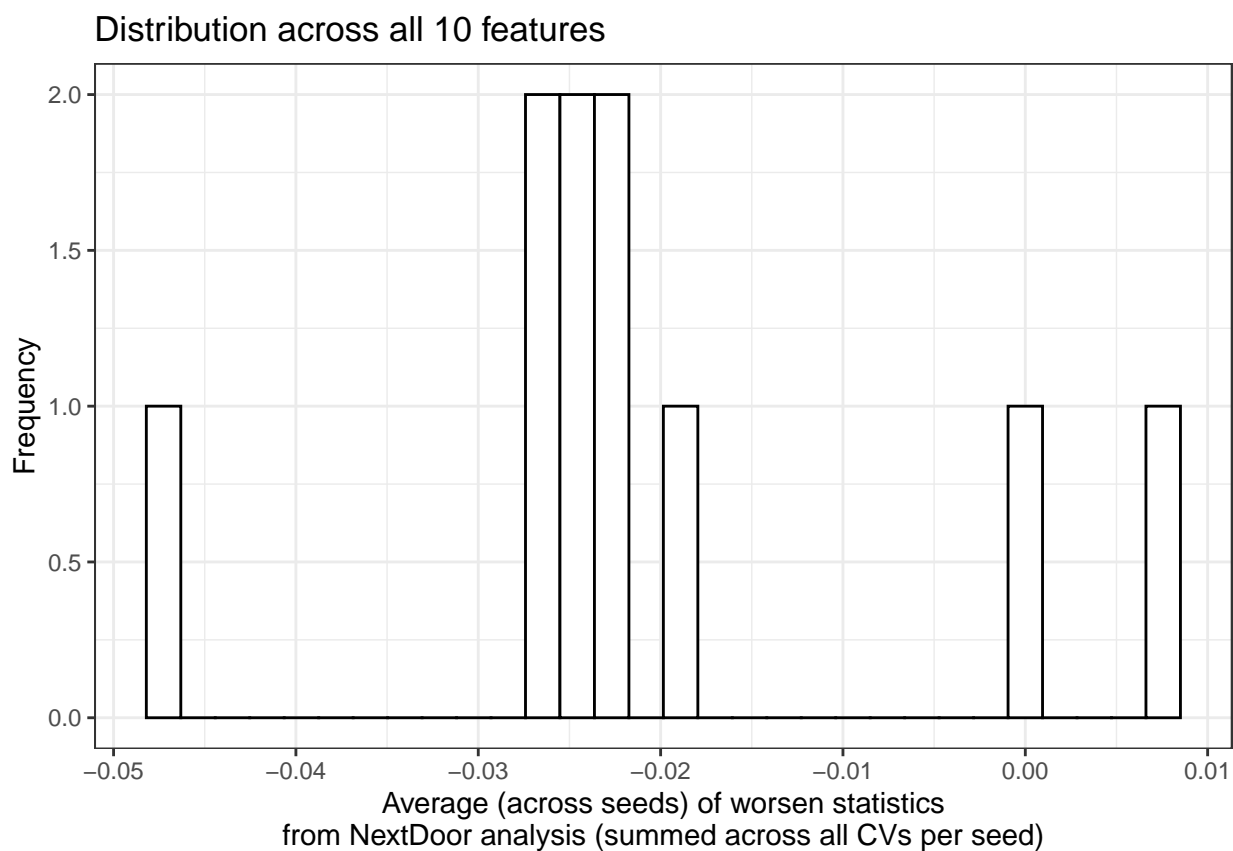
prediction under seed 1001
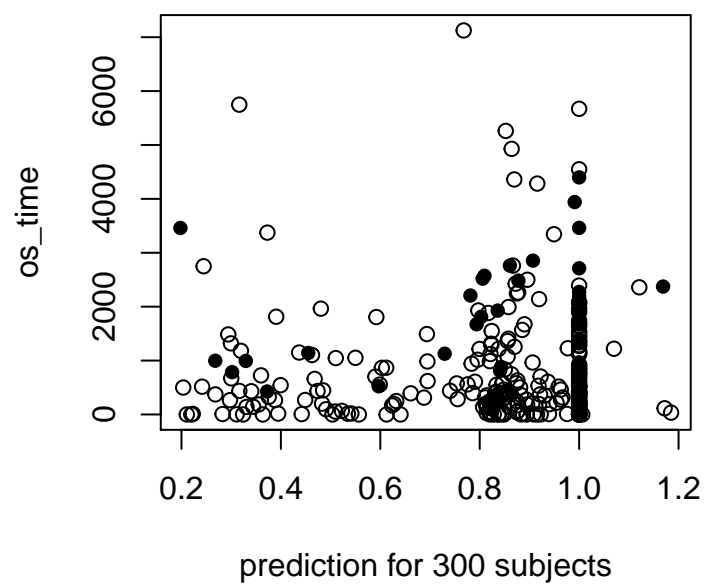
2. Important Features

distribution across 2 seed x 3 CV

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

```
## [1] "there are 10 unique features used from the 100 feature set"
## [1] "summary of number of features used in each run under 2 seeds and 3 CVs"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       2       2       3       3

## [1] "there are 8 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 10 features

Top feature, by the worsen statistic from NextDoor analysis

```
## [1] "there are 8 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 10 features

Frequency of use across CVs and seeds

Top feature, by usage frequency

top 10 features