Evaluate testing data (survival)

Andrew Chang

2018-11-07

Contents

Labels: os_time

0. Load Data

```r
library(tidyverse)
library(ggrepel)
library(survival)
library(survminer)
library(randomForestSRC)
library(broom)
source("~/ml-pipeline/ML_performanceCheck/PerformanceUtils.R")

home <- "~/ml-pipeline/tests"
project_name <- "test_survival_1"
training_data <- "test_regr_surv_tcga_brca.csv"
label_name <- "os_time"

## output path for analyzed data
outfile <- paste0(home, "/results/", project_name)
```

```
300 of samples were used

100 of full features

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

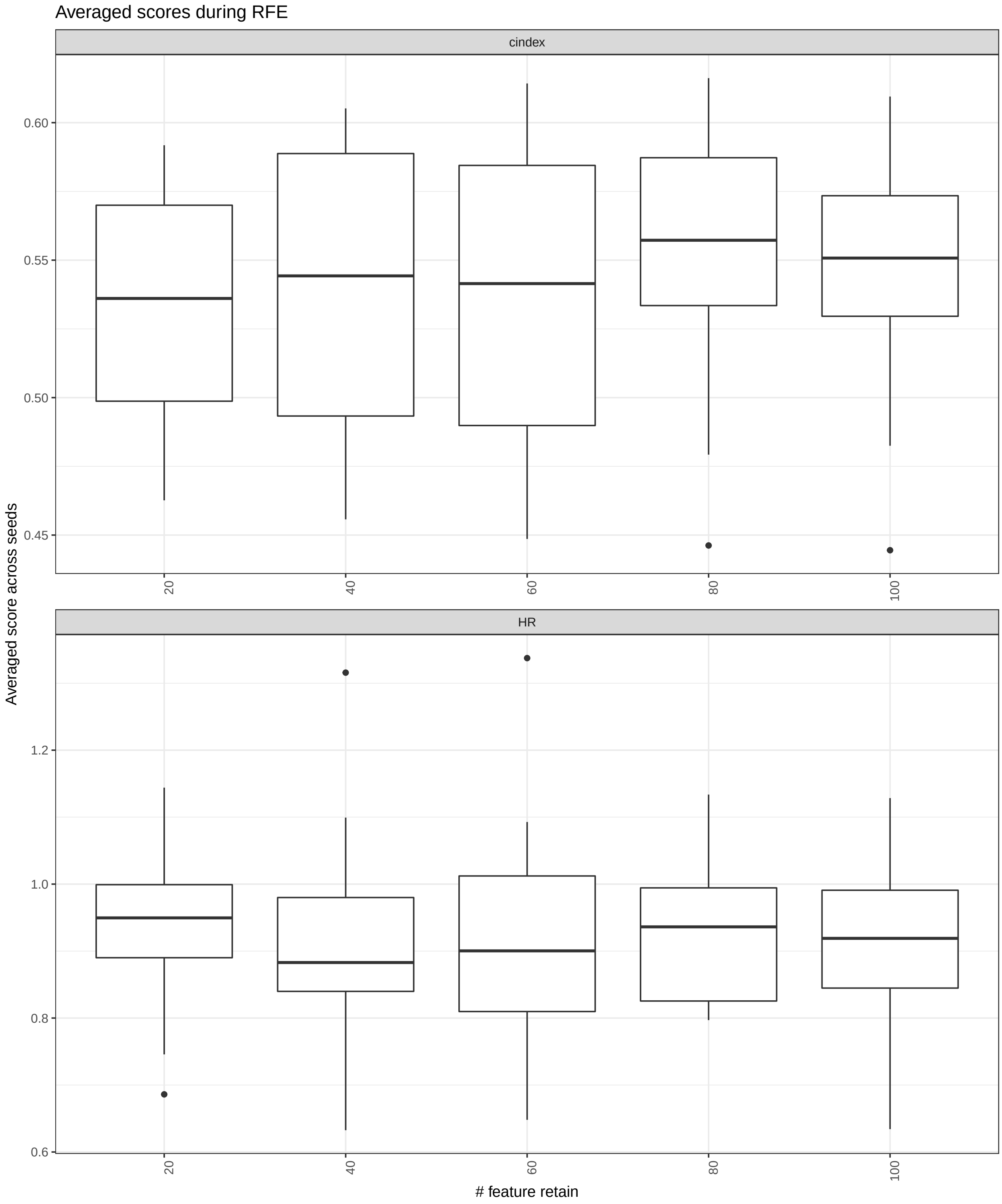
1. Scores

`average = T`: scores are based on pre-validation (combine predictions from all CVs per job and then calculate a single score per job).

`average = F`: report all the scores from each CV across entire jobs.

1. Scores

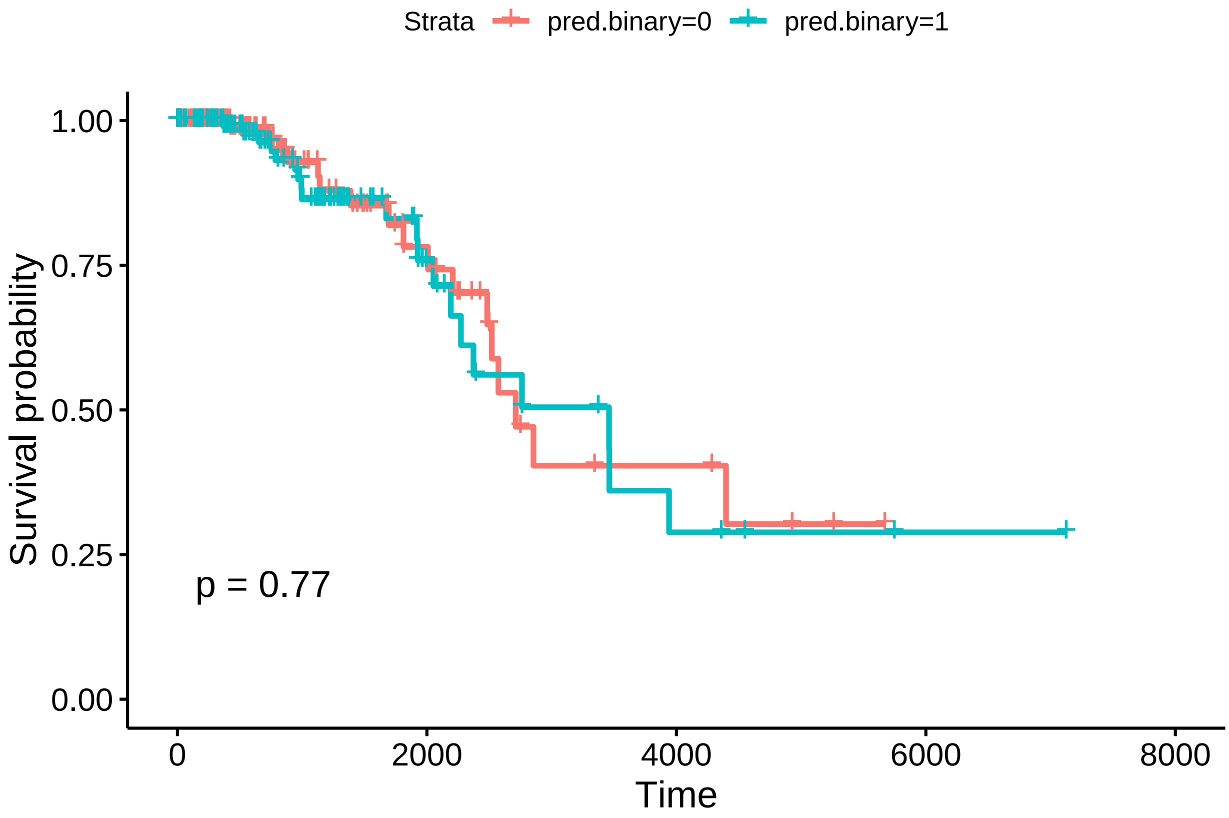
1.1 uncalibrated average



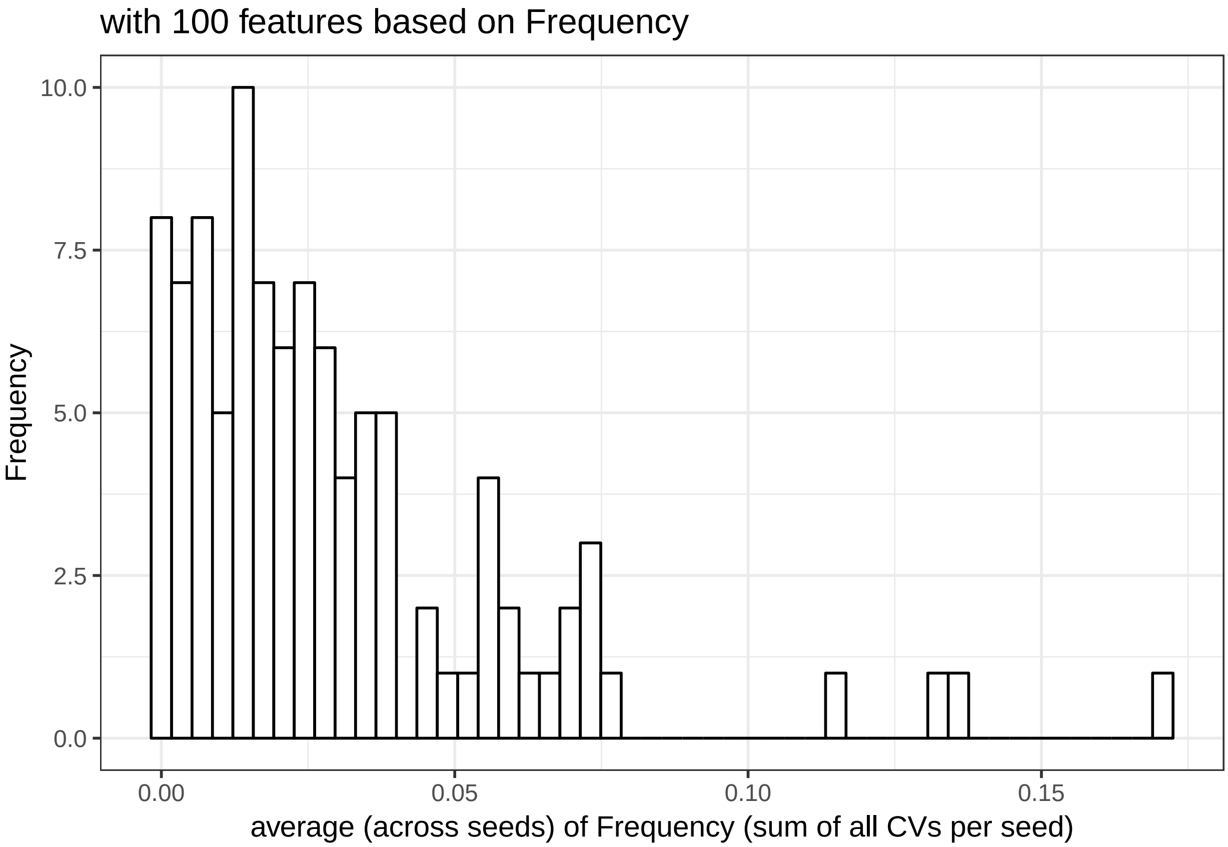
```
score.med <- plt$df.scores %>%
  gather("metrics", "score", -one_of(c("seed", "size"))) %>%
  group_by(seed, size, metrics) %>%
  summarize(avg.score = mean(score)) %>%
  filter(metrics %in% c("HR", "cindex")) %>%
  group_by(size, metrics) %>%
  summarize(med = round(median(avg.score), 3))

cmax <- score.med %>%
  filter(metrics == "cindex") %>%
  arrange(desc(med)) %>%
  head(1)
hrmin <- score.med %>%
  filter(metrics == "HR") %>%
  arrange(med) %>%
  head(1)

knitr::kable(rbind(cmax, hrmin))
```

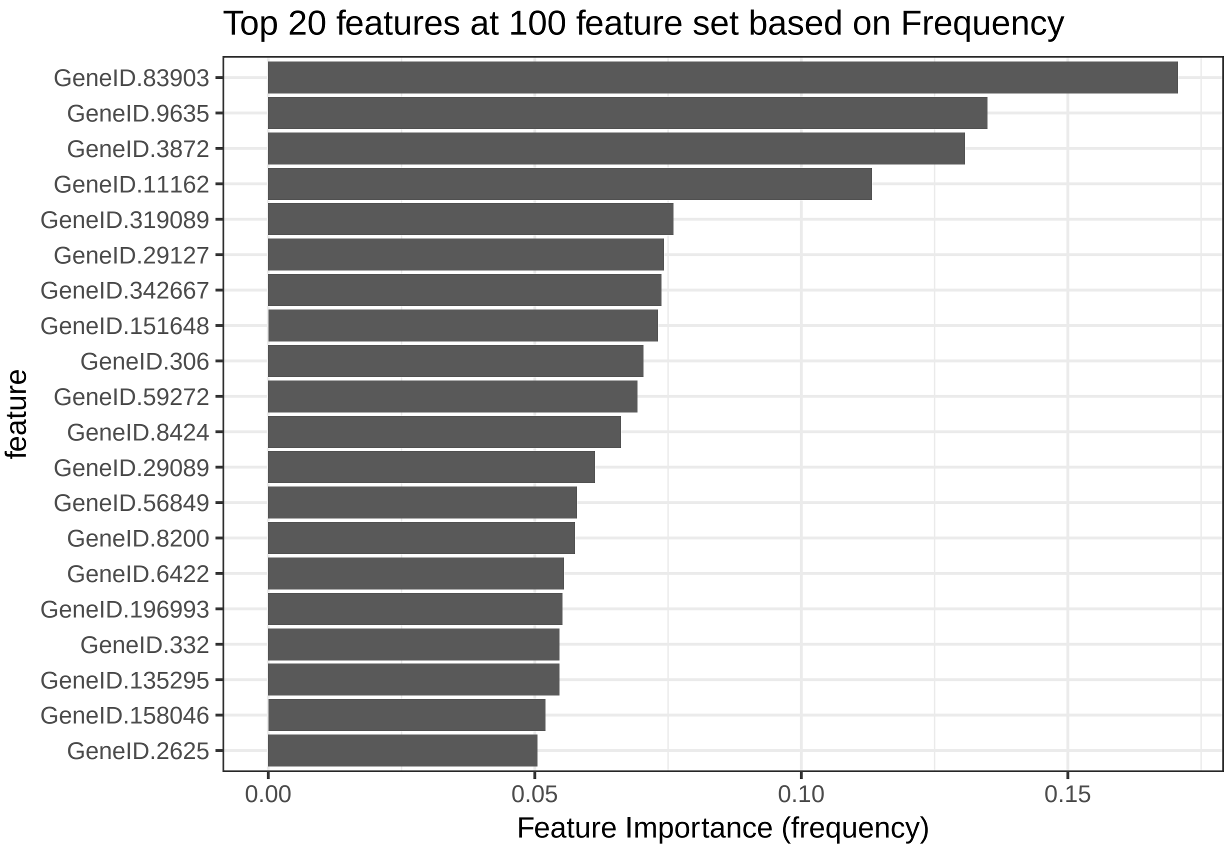
size	metrics	med
80	cindex	0.557
40	HR	0.883

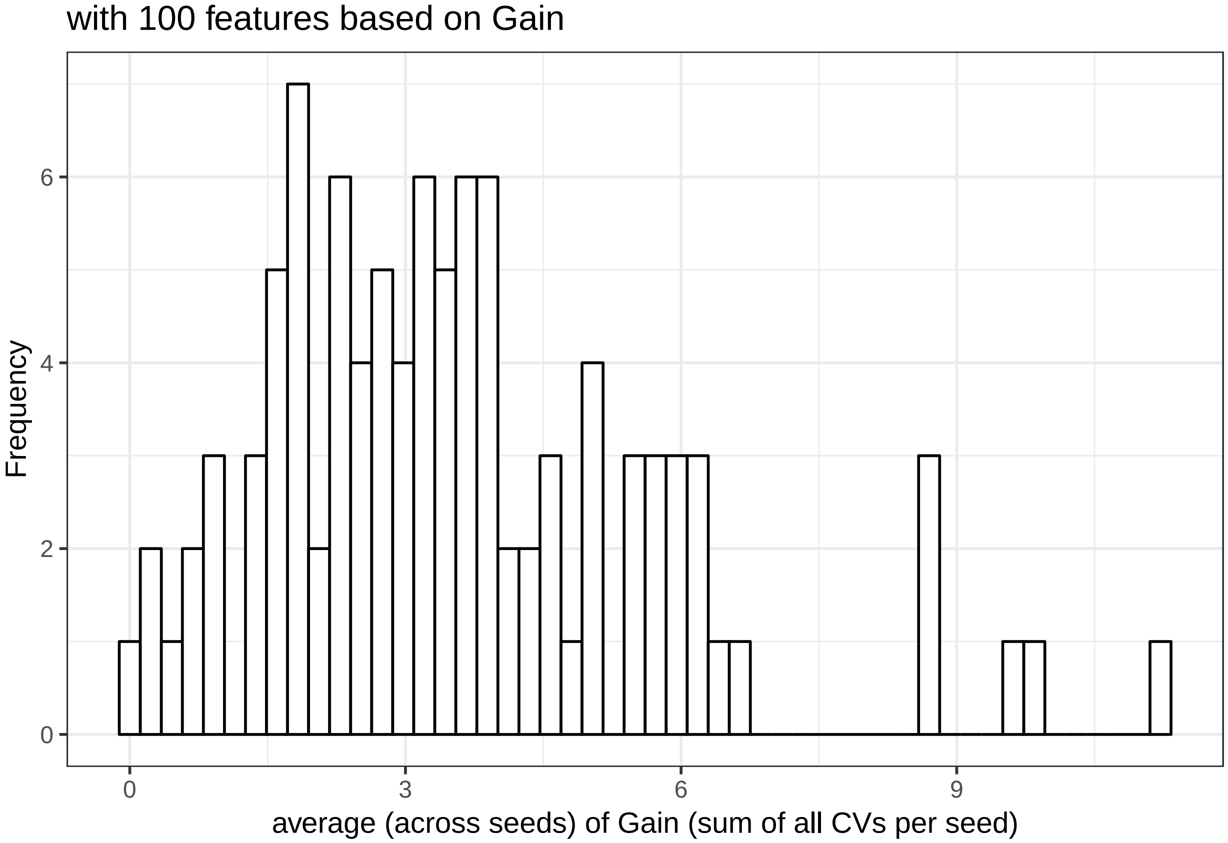
```
df.in <- df.preval %>%
  filter(size == 40, seed==1001) %>% # seed==1001
  mutate(pred.binary = ifelse(pred < median(pred), 1, 0))

fit1 <- survfit(Surv(col_surv, col_event) ~ pred.binary, data = df.in)
ggsurvplot(fit1, data = df.in, pval = TRUE)
```

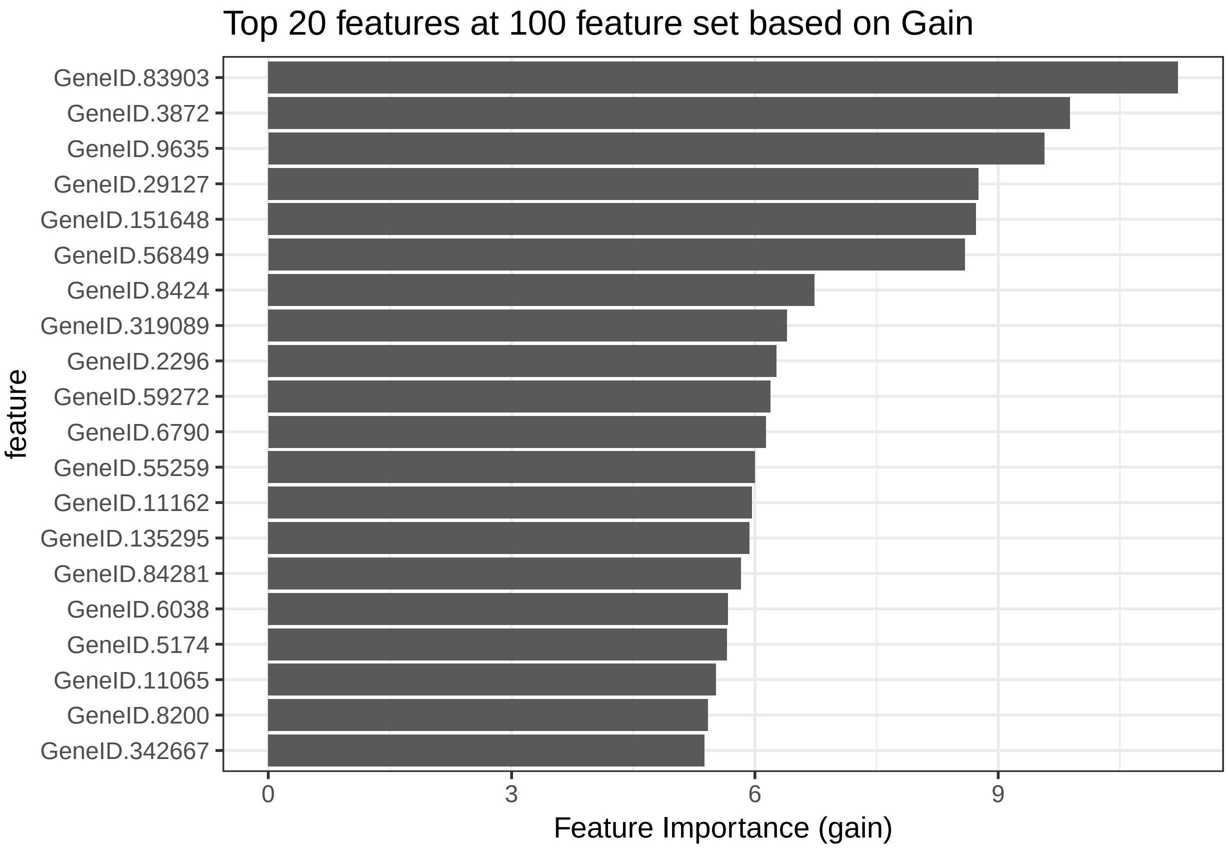

Strata
pred.binary=0
pred.binary=1
1.00
0.75
0.50
0.25
0.00
Survival probability
p = 0.77
0
2000
4000
6000
8000
Time

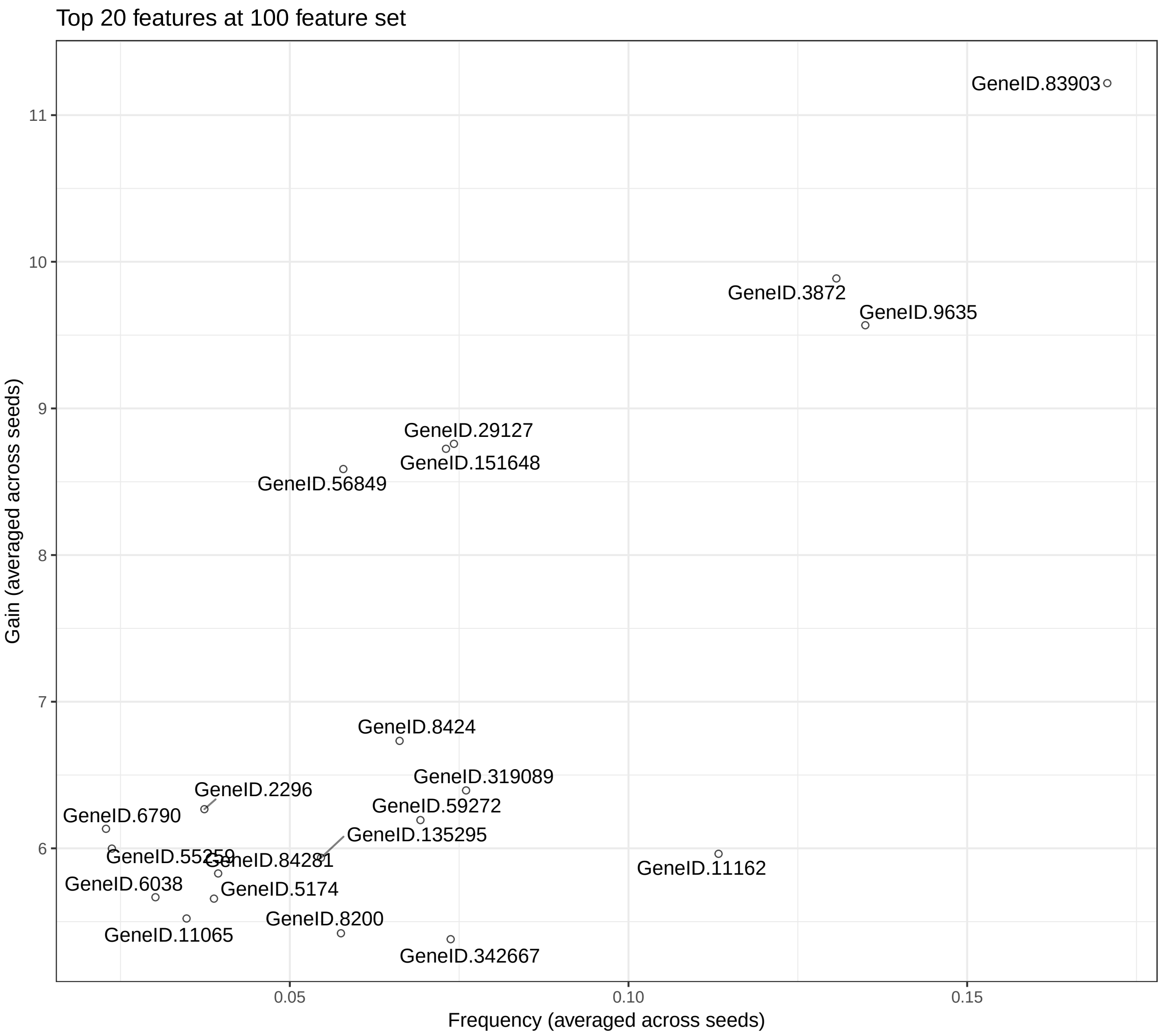
3. Important Features



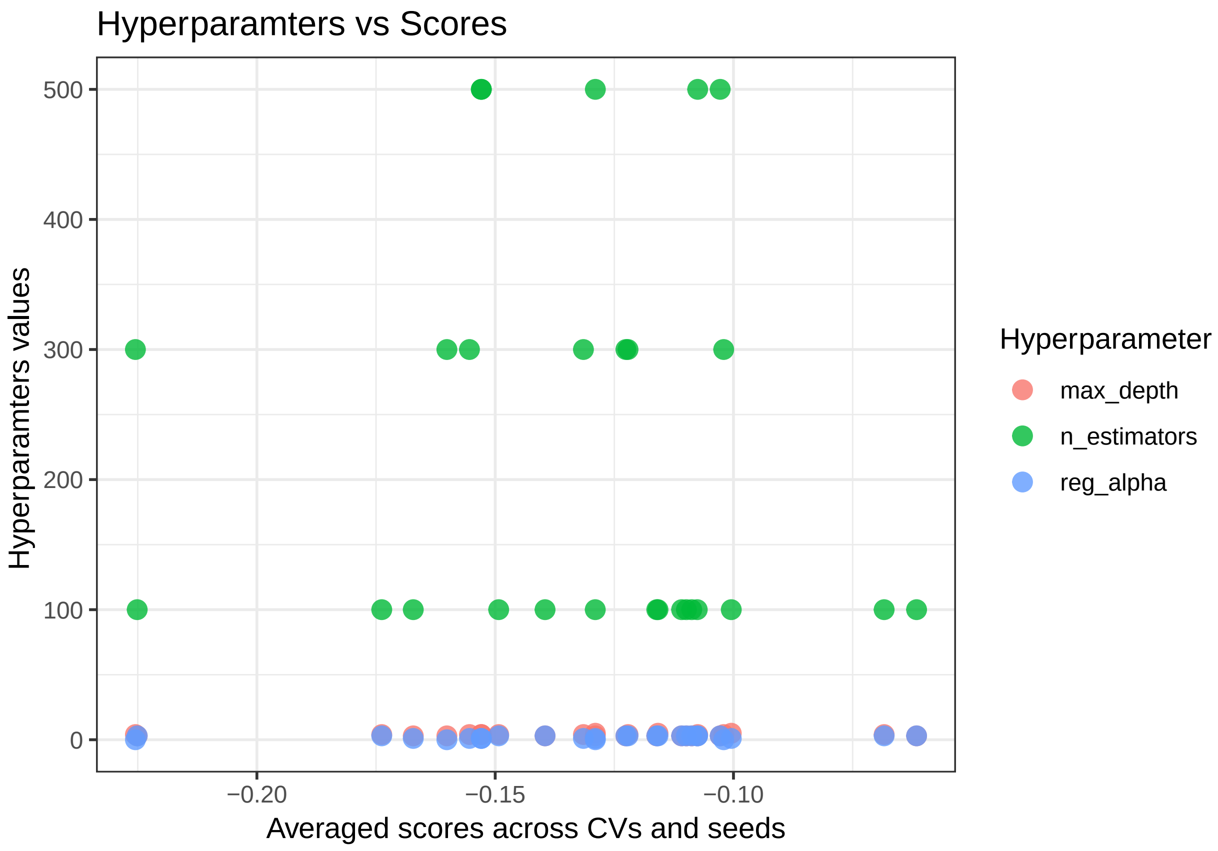

Top 20 features at 100 feature set based on Frequency
feature
GeneID.83903
GeneID.9635
GeneID.3872
GeneID.11162
GeneID.319089
GeneID.29127
GeneID.342667
GeneID.151648
GeneID.306
GeneID.59272
GeneID.8424
GeneID.29089
GeneID.56849
GeneID.8200
GeneID.6422
GeneID.196993
GeneID.332
GeneID.135295
GeneID.158046
GeneID.2625
0.00
0.05
0.10
0.15
Feature Importance (frequency)


with 100 features based on Gain
Frequency
0
2
4
6
0
3
6
9
average (across seeds) of Gain (sum of all CVs per seed)


Top 20 features at 100 feature set based on Gain
feature
GeneID.83903
GeneID.3872
GeneID.9635
GeneID.29127
GeneID.151648
GeneID.56849
GeneID.8424
GeneID.319089
GeneID.2296
GeneID.59272
GeneID.6790
GeneID.55259
GeneID.11162
GeneID.135295
GeneID.84281
GeneID.6038
GeneID.5174
GeneID.11065
GeneID.8200
GeneID.342667
0
3
6
9
Feature Importance (gain)


Top 20 features at 100 feature set
Gain (averaged across seeds)
Frequency (averaged across seeds)
11
10
9
8
7
6
0.05
0.10
0.15
GeneID.83903
GeneID.3872
GeneID.9635
GeneID.29127
GeneID.151648
GeneID.56849
GeneID.8424
GeneID.319089
GeneID.2296
GeneID.59272
GeneID.6790
GeneID.135295
GeneID.55259
GeneID.84281
GeneID.11162
GeneID.6038
GeneID.5174
GeneID.8200
GeneID.11065
GeneID.342667

4. Hyper-parameters


Hyperparamters vs Scores
Hyperparamters values
Averaged scores across CVs and seeds
Hyperparameter
max_depth
n_estimators
reg_alpha