Evaluate testing data (regression)

Andrew Chang

2018-11-07

Contents

Labels: os_time

0. Load Data

```
library(tidyverse)
library(ggrepel)
library(caret)
library(pROC)
source("~/ml-pipeline/ML_performanceCheck/PerformanceUtils.R")

home <- "~/ml-pipeline/tests"
project_name <- "test_regression_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
```

```
101 of full features
```

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

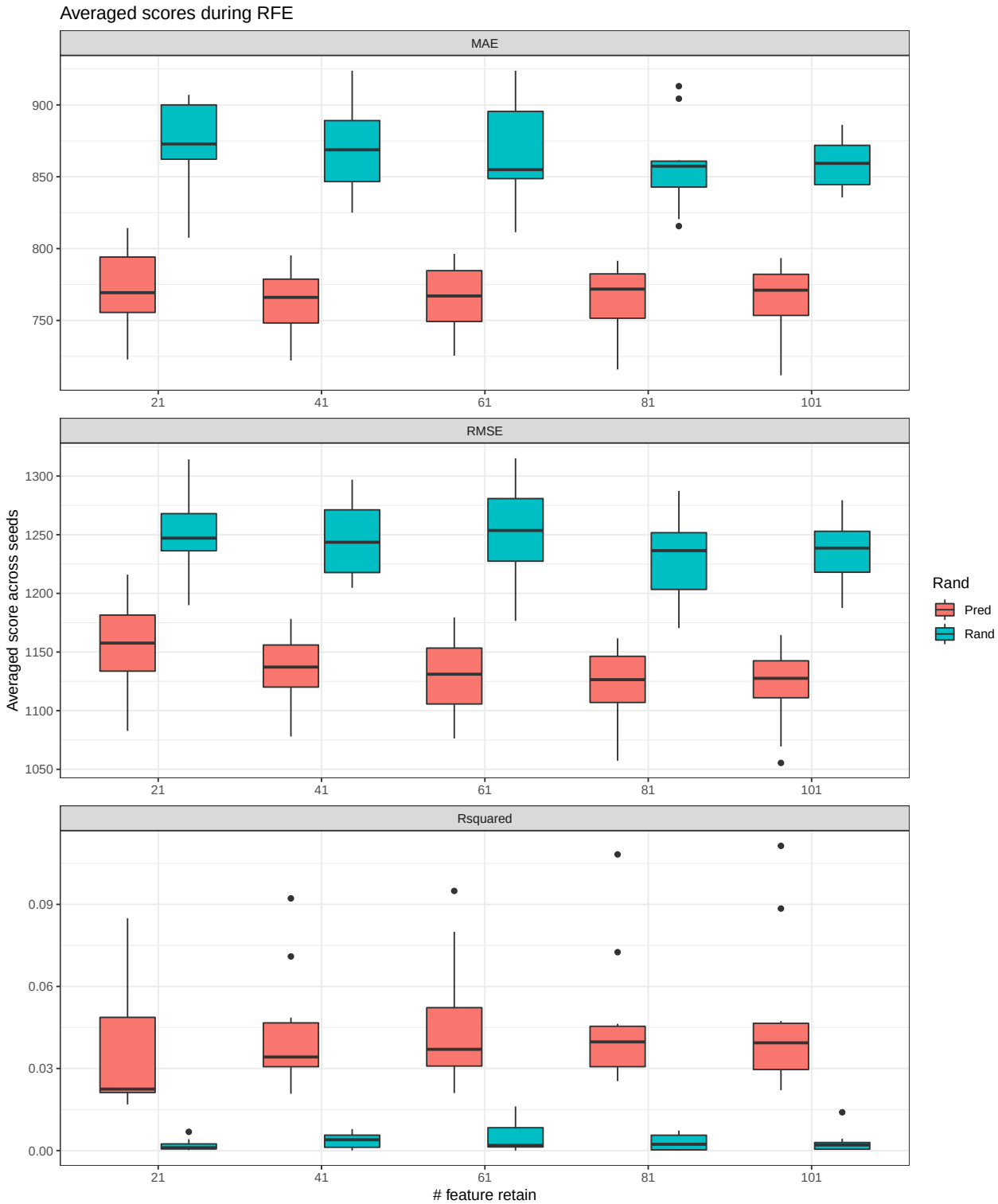
1. Scores

`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.

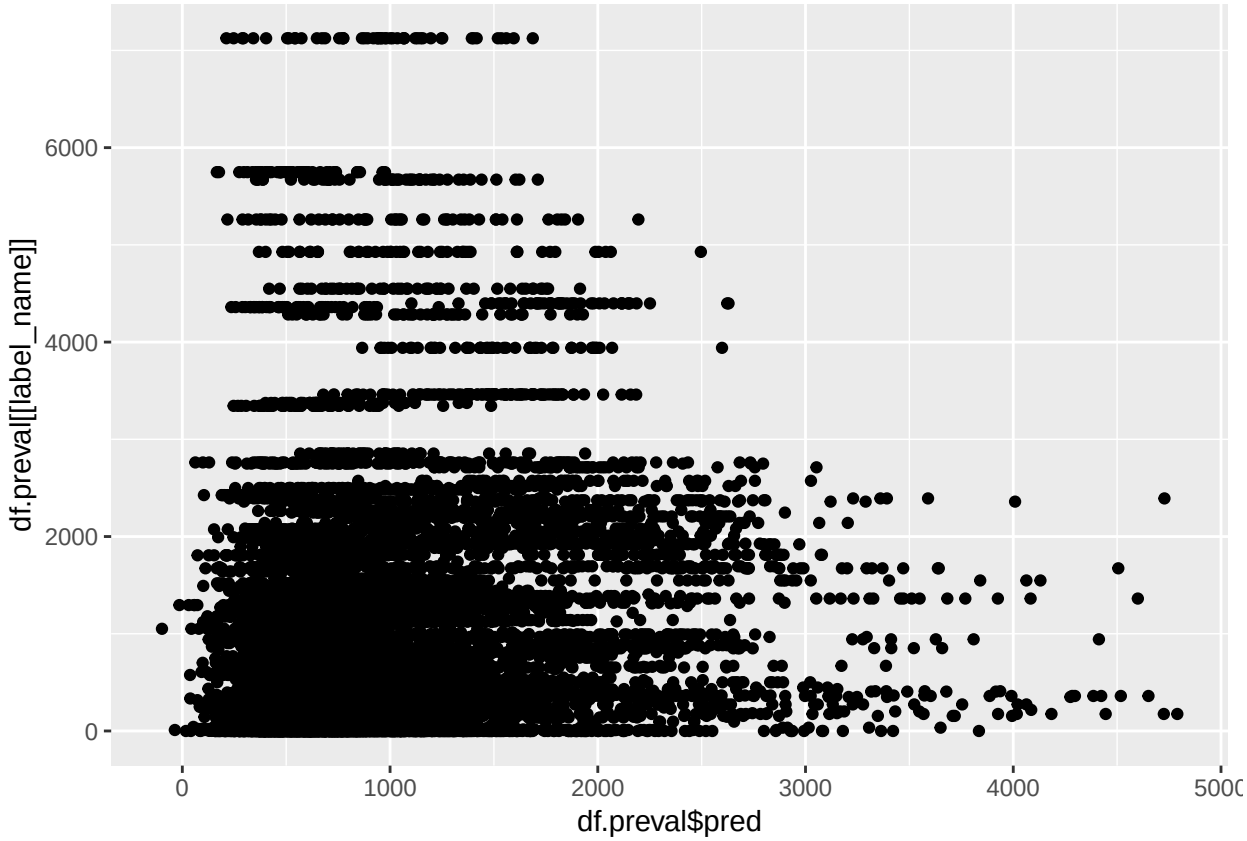
1. Scores

1.1 uncalibrated average

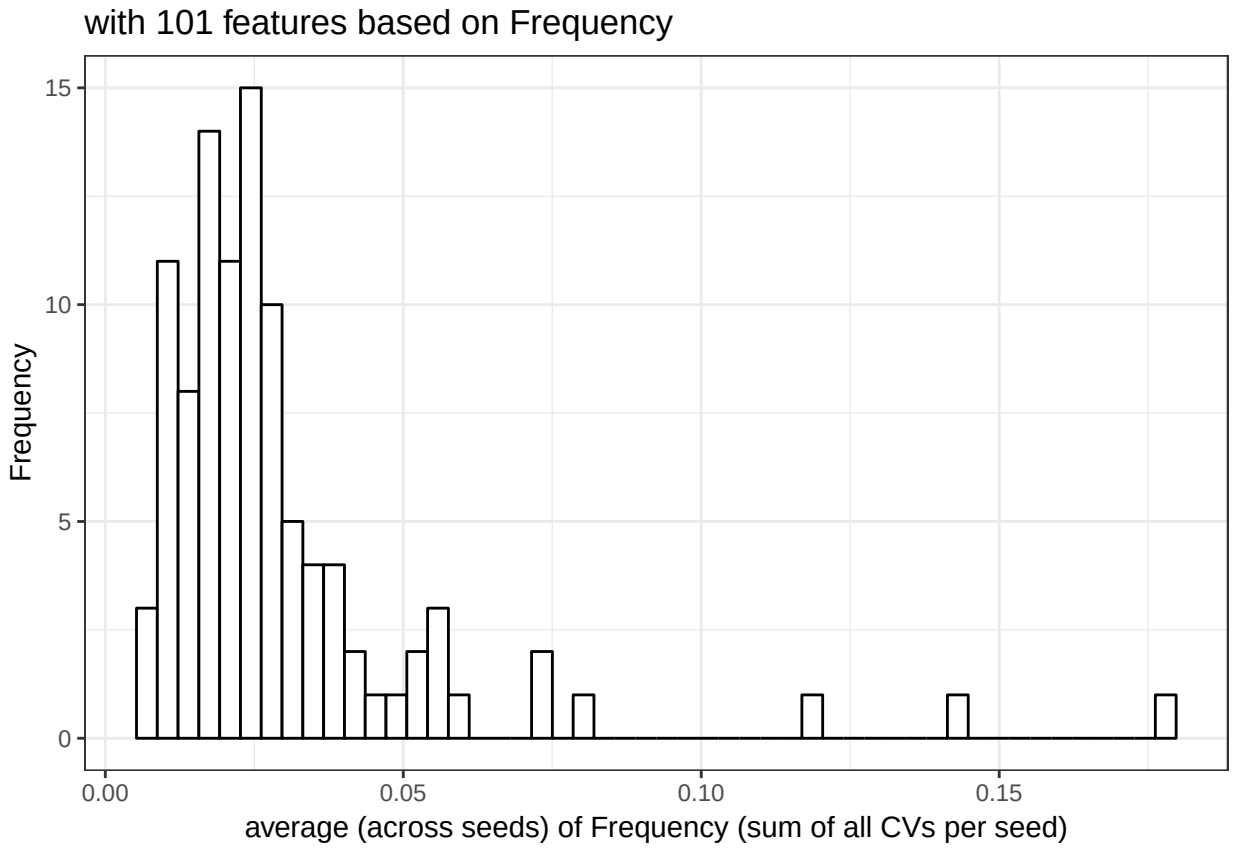


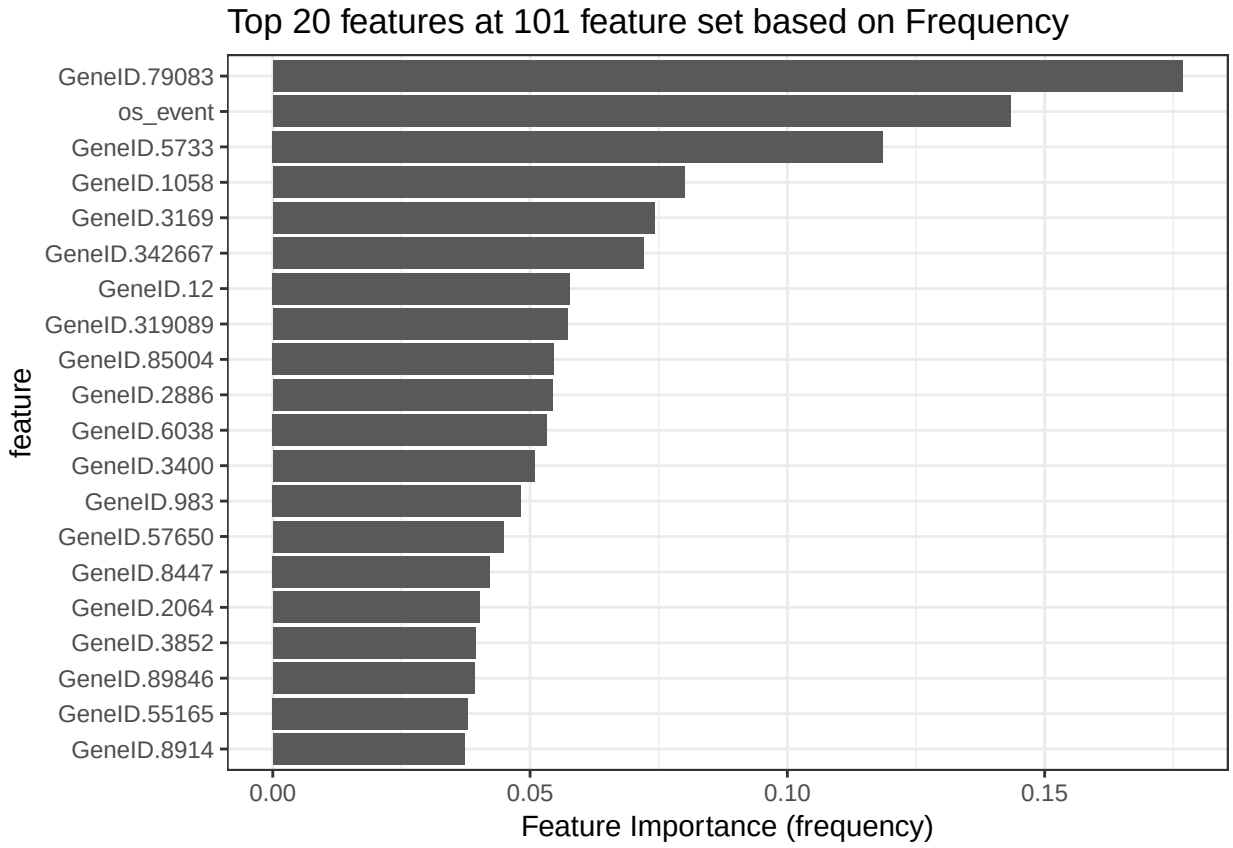

```
ggplot(data = df.preval, aes(x=df.preval$pred,
                             y = df.preval[[label_name]])) +
```

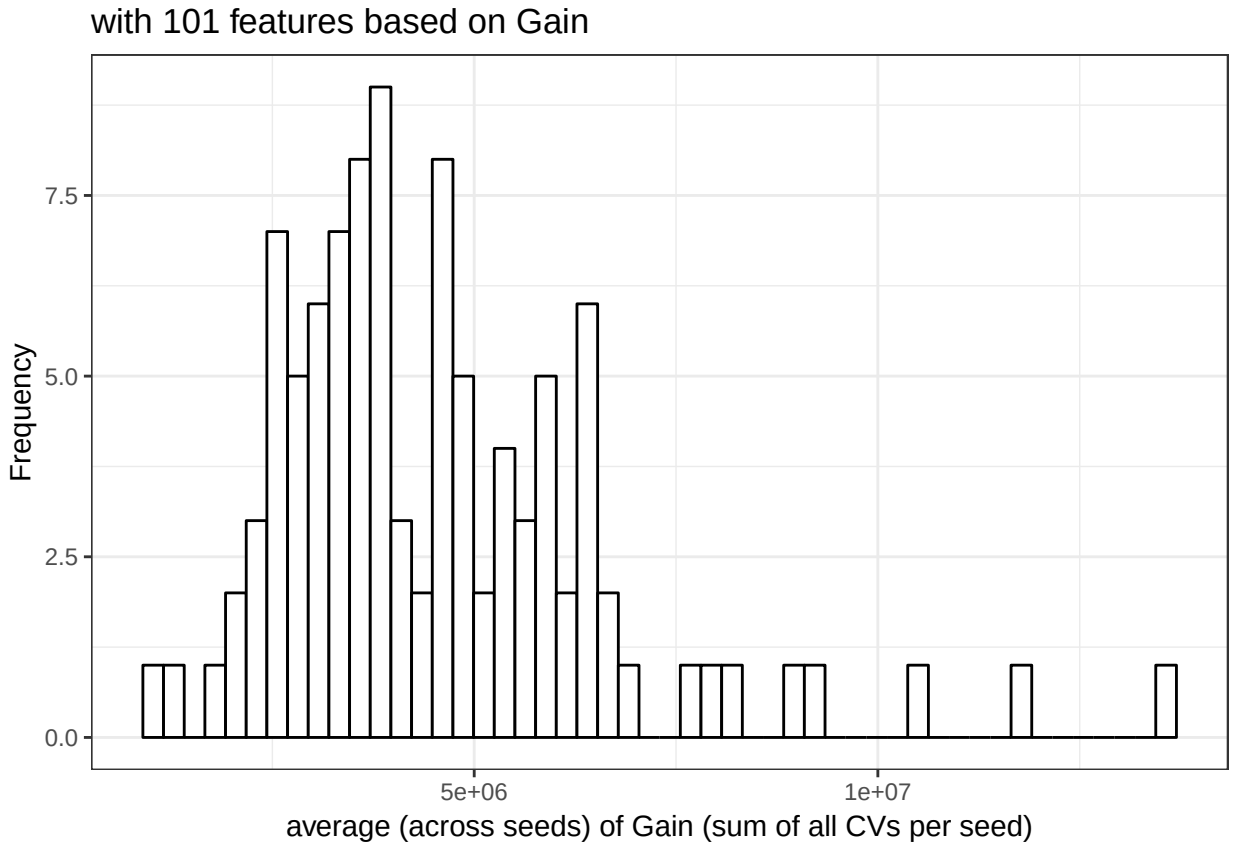
```
  geom_point()
```

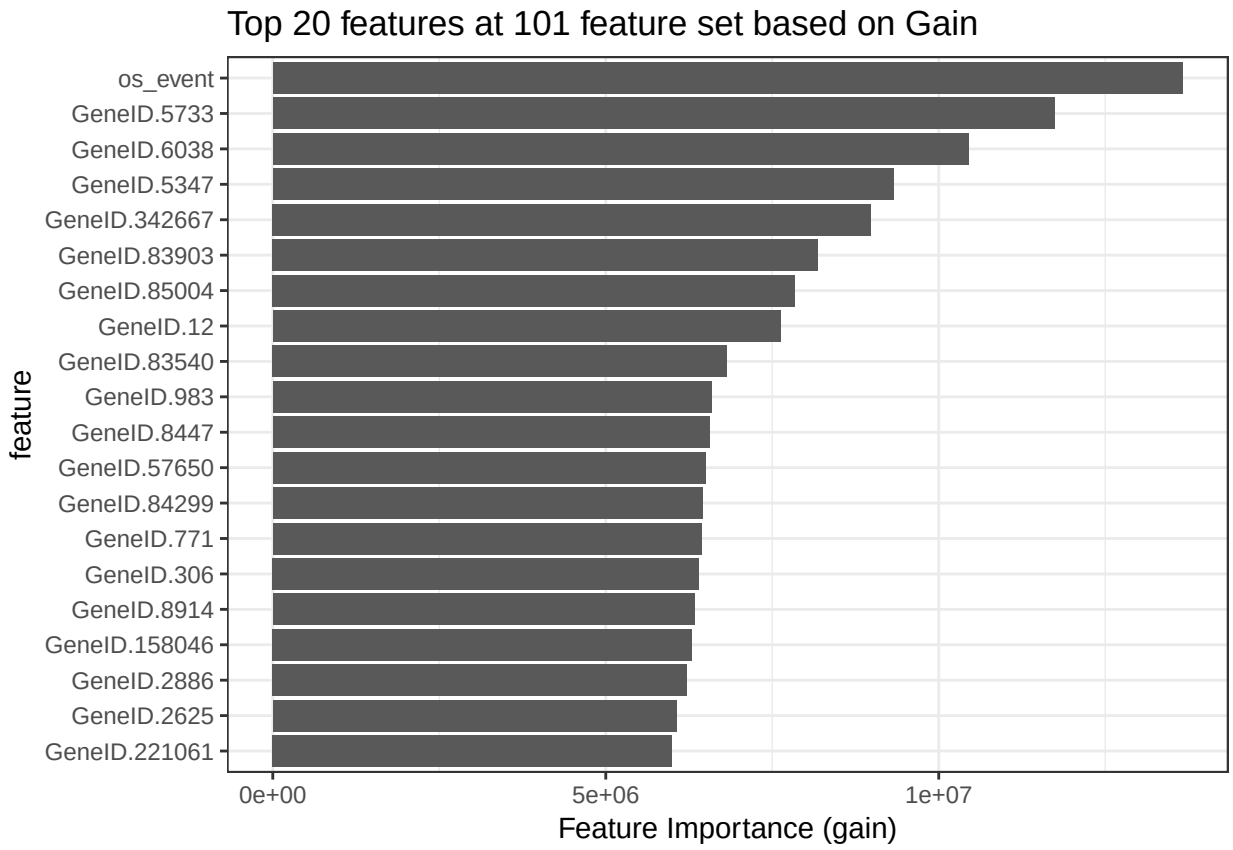

6000
4000
2000
0
df.preval[[label_name]]
0
1000
2000
3000
4000
5000
df.preval$pred

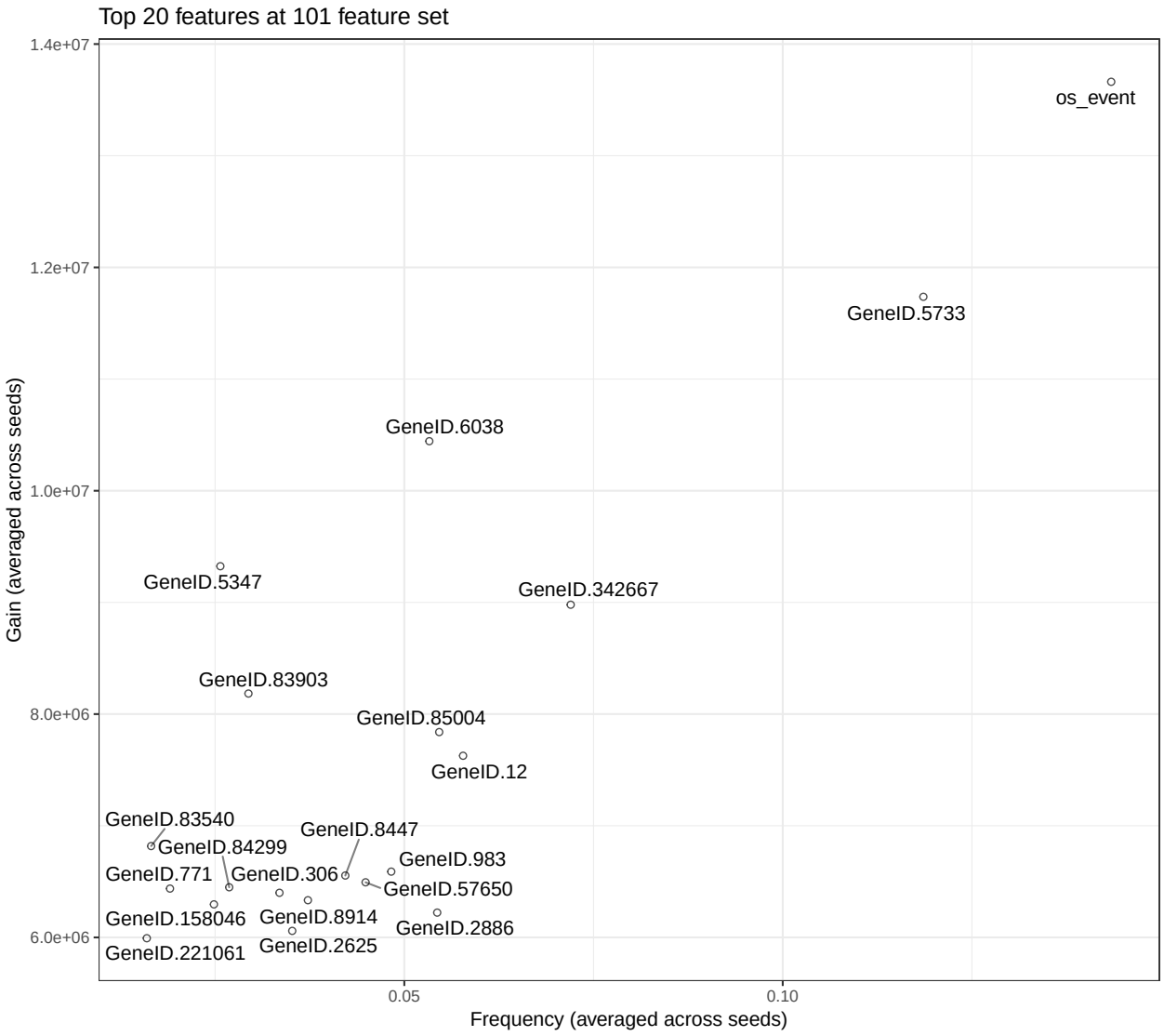
3. Important Features




Top 20 features at 101 feature set based on Frequency
GeneID.79083
os_event
GeneID.5733
GeneID.1058
GeneID.3169
GeneID.342667
GeneID.12
GeneID.319089
GeneID.85004
GeneID.2886
GeneID.6038
GeneID.3400
GeneID.983
GeneID.57650
GeneID.8447
GeneID.2064
GeneID.3852
GeneID.89846
GeneID.55165
GeneID.8914
feature
0.00
0.05
0.10
0.15
Feature Importance (frequency)


with 101 features based on Gain
7.5
5.0
2.5
0.0
Frequency
5e+06
1e+07
average (across seeds) of Gain (sum of all CVs per seed)


Top 20 features at 101 feature set based on Gain
os_event
GeneID.5733
GeneID.6038
GeneID.5347
GeneID.342667
GeneID.83903
GeneID.85004
GeneID.12
GeneID.83540
GeneID.983
GeneID.8447
GeneID.57650
GeneID.84299
GeneID.771
GeneID.306
GeneID.8914
GeneID.158046
GeneID.2886
GeneID.2625
GeneID.221061
feature
0e+00
5e+06
1e+07
Feature Importance (gain)


Top 20 features at 101 feature set
Gain (averaged across seeds)
1.4e+07
1.2e+07
1.0e+07
8.0e+06
6.0e+06
0.05
0.10
Frequency (averaged across seeds)
os_event
GeneID.5733
GeneID.6038
GeneID.5347
GeneID.342667
GeneID.83903
GeneID.85004
GeneID.12
GeneID.83540
GeneID.84299
GeneID.8447
GeneID.983
GeneID.771
GeneID.306
GeneID.57650
GeneID.158046
GeneID.8914
GeneID.2886
GeneID.221061
GeneID.2625

4. Hyper-parameters