Milestone 2: Core results 01

Anders

25/1/2023

Data

One clinical trails on breast cancer (advanced HR+/HER2- and HER2-E breast cancer) using two different drug combination; and a cohorts study (here used as test data set). / Both data set have 771 genes. This genes are specifically selected based likely potential role in breast cancer. / The gene set is dived into X numbers of "signature genes"; which are thought to represent unities with respect to cancer biology...

Trail

Two treatments which differ with respect to drug combination - Target: ribociclib and endocrine therapy (letrozole) - Chemotherapy: doxorubicin, cyclophosphamide and paclitaxel. approx. 50 patients in each group. Endpoints: proliferation score, ROR score

Cohort

The primary objective of this study is to compare two cdk4/6 targeted drugs (Palbociclib, n=36; Abemaciclib, n=3 in combination with endocrine therapy (tamoxifen, fulvestrant or aromatase inhibitors, I think?)

Endpoints: progression free survival (months), OS?, and status of the two former (dont know what that means)

Major goal

1. Find best model to predict outcome of cancer treatment with genetic profile as predictive features
2. Features selection in order to understand cancer biology

Major challanges

Preliminary experiments (on trail 1) showed instability in prediction and feature selection between bootstrap samples of Lasso. I believe this is a classical problem of high-dim data?

Approch

Test all thinkable models to see if some is surperior

Evaluation of models

Two levels of evaluation is scheduled:

1. Relative comparison of models

1000 bootstrap models are fitted and then evaluated on the original sample. This gives a relative comparison of the various models with respect to data similar to the given data set. In addition to Correlation and MSE, frequency of selected features is compared.

2. Expected outcome of future patients

3 strategies are considered:
1. Repeated cross-validations (100 rep)
2. Bootstrap models with 0.632 (or 0.632?) adjustment
3. Test data set from a second trail (This trail have different responses)

Models

Lasso

Post Lasso

Ridge

Elastic Net

Boosting with stumps as base learner

- mboost

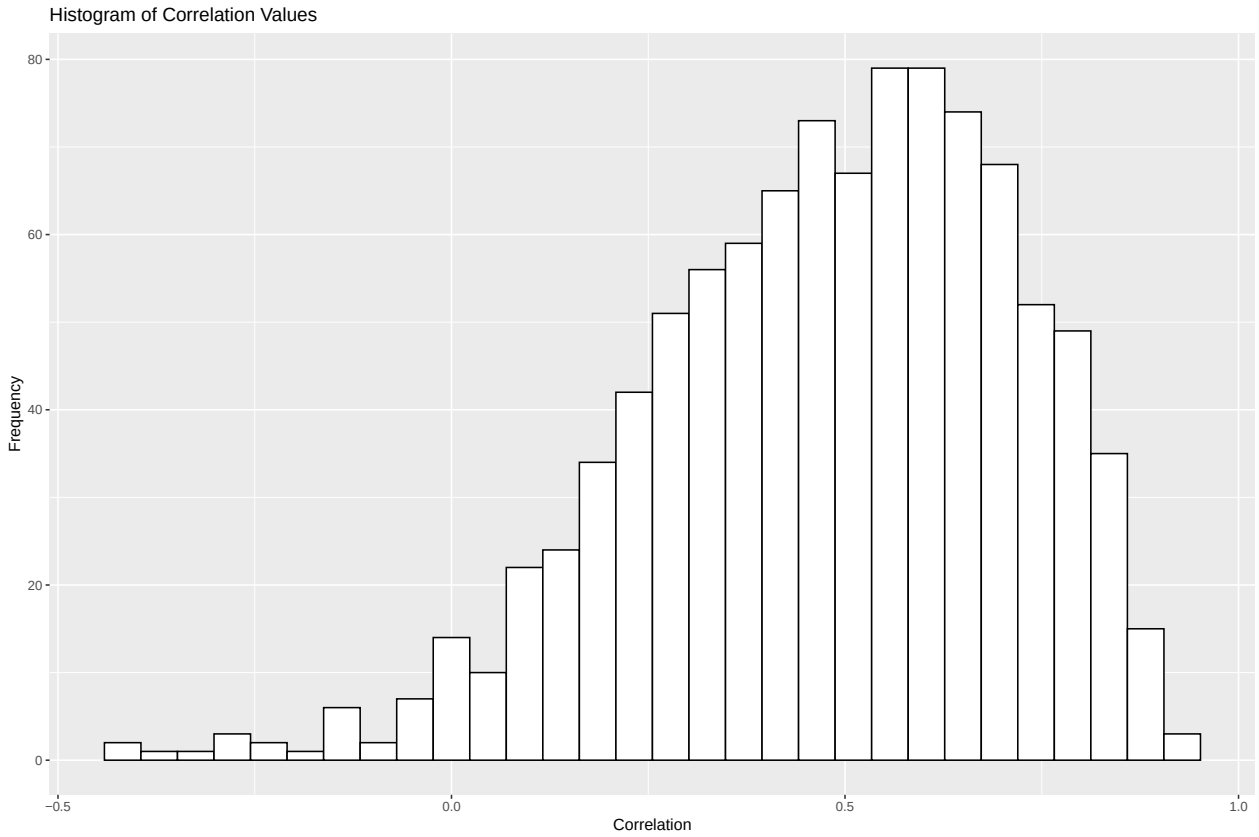
- xgboost

Feature selecting ensemble model

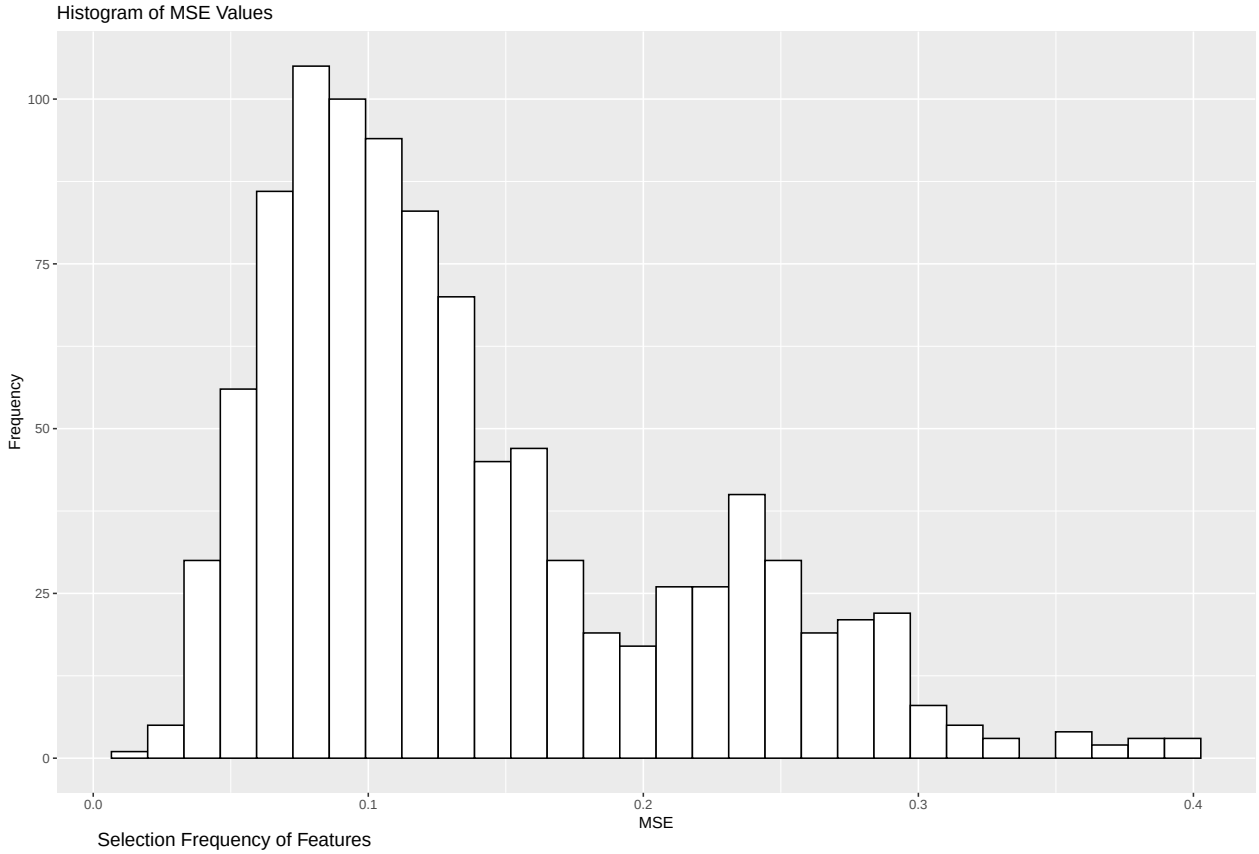
RESULTS

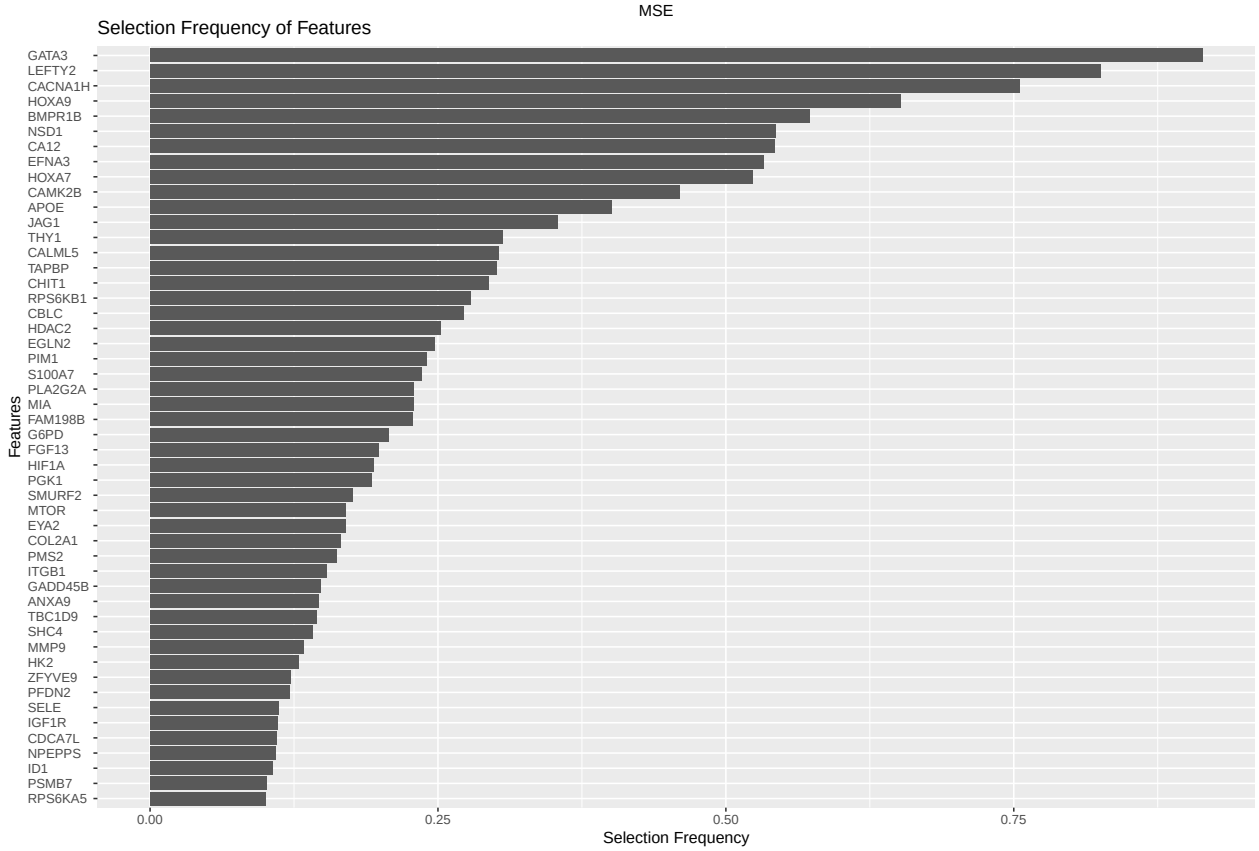
Lasso

```
## Fraction of model fits with no selected genes: 0.004
## [1] "Correlation results:"
## Mean correlation: 0.478871
## Median correlation: 0.506616
## Variance 0.05505632
## [1] "st.dev.:"
## [1] 0.2346408
```




```
## [1] 0.1376413
## [1] 0.07444957
```


Histogram of MSE Values
Frequency
MSE
0
25
50
75
100
0.0
0.1
0.2
0.3
0.4


Selection Frequency of Features
Features
Selection Frequency
GATA3
LEFTY2
CACNA1H
HOXA9
BMPR1B
NSD1
CA12
EFNA3
HOXA7
CAMK2B
APOE
JAG1
THY1
CALML5
TAPBP
CHIT1
RPS6KB1
CBLC
HDAC2
EGLN2
PIM1
S100A7
PLA2G2A
MIA
FAM198B
G6PD
FGF13
HIF1A
PGK1
SMURF2
MTOR
EYA2
COL2A1
PMS2
ITGB1
GADD45B
ANXA9
TBC1D9
SHC4
MMP9
HK2
ZFYVE9
PFDN2
SELE
IGF1R
CDCA7L
NPEPPS
ID1
PSMB7
RPS6KA5
0.00
0.25
0.50
0.75

```
## [1] "GATA3"
```

```
##  [1] "GATA3"   "LEFTY2"  "CACNA1H" "HOXA9"   "BMPR1B"  "NSD1"    "CA12"
##  [8] "EFNA3"   "HOXA7"   "CAMK2B"
```

Post Lasso

Ridge

Elastic Net

Boosting with stumps as base learner