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

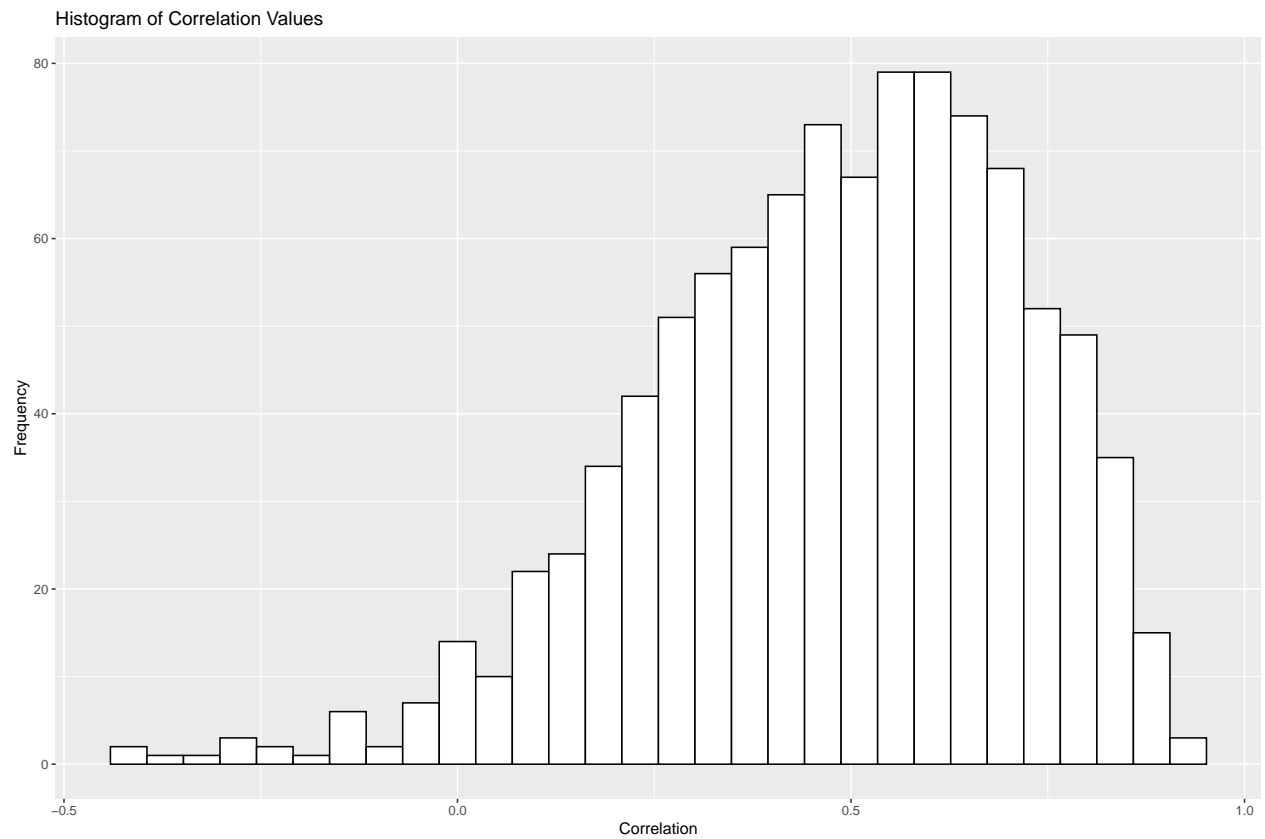**Boosting with stumps as base learner**

- mboost

- xgboost

**Feature selecting ensemble model**

RESULTS

Lasso

```
## number of models fitted: 1000
## Fraction of model fits with no selected genes: 0.004
##
## CORRELATIONS RESULTS
## Mean: 0.478871
## Median: 0.506616
## Variance: 0.05505632
## st.dev.: 0.2346408
```

Histogram of Correlation Values


```
## MSE RESULTS
## Mean: 0.1376413
## Median: 0.1168221
## Variance: 0.005542738
## st.dev.: 0.07444957
```

Histogram of MSE Values

Selection Frequency of Features

##
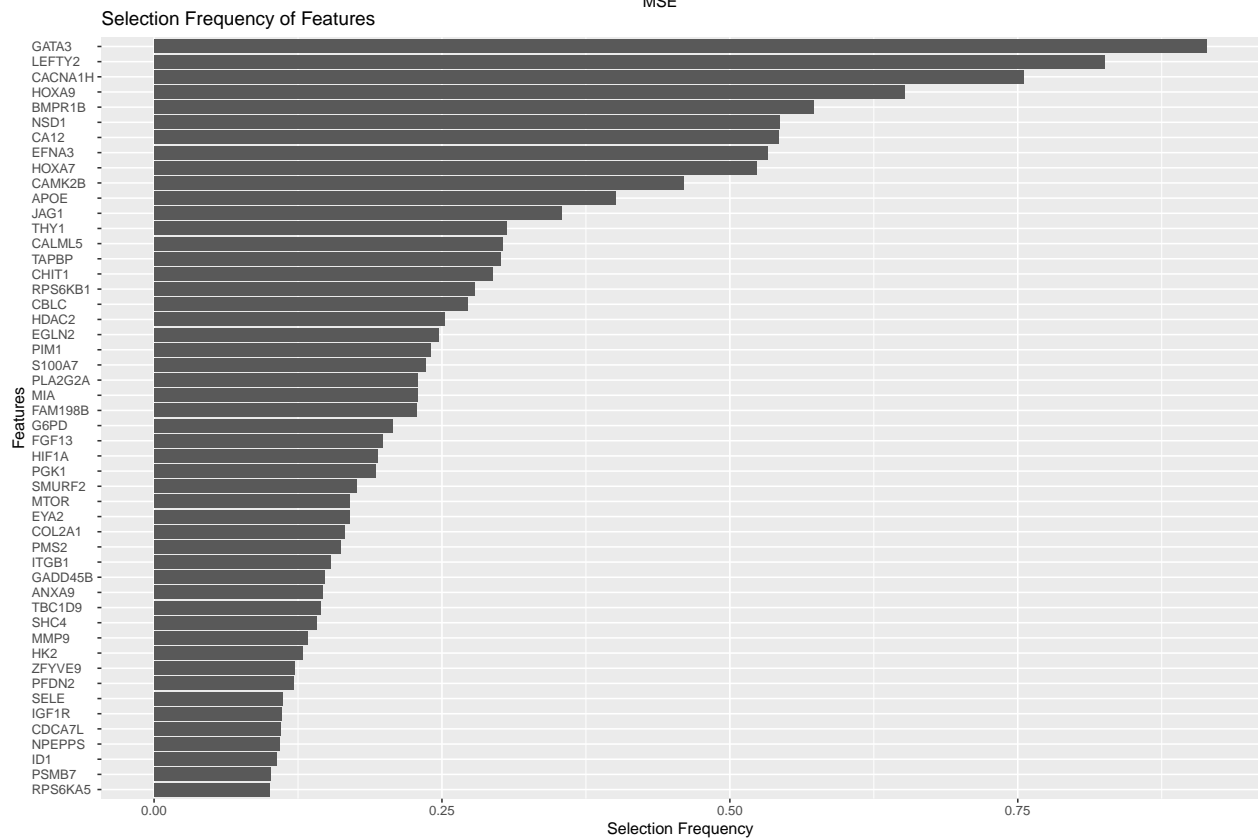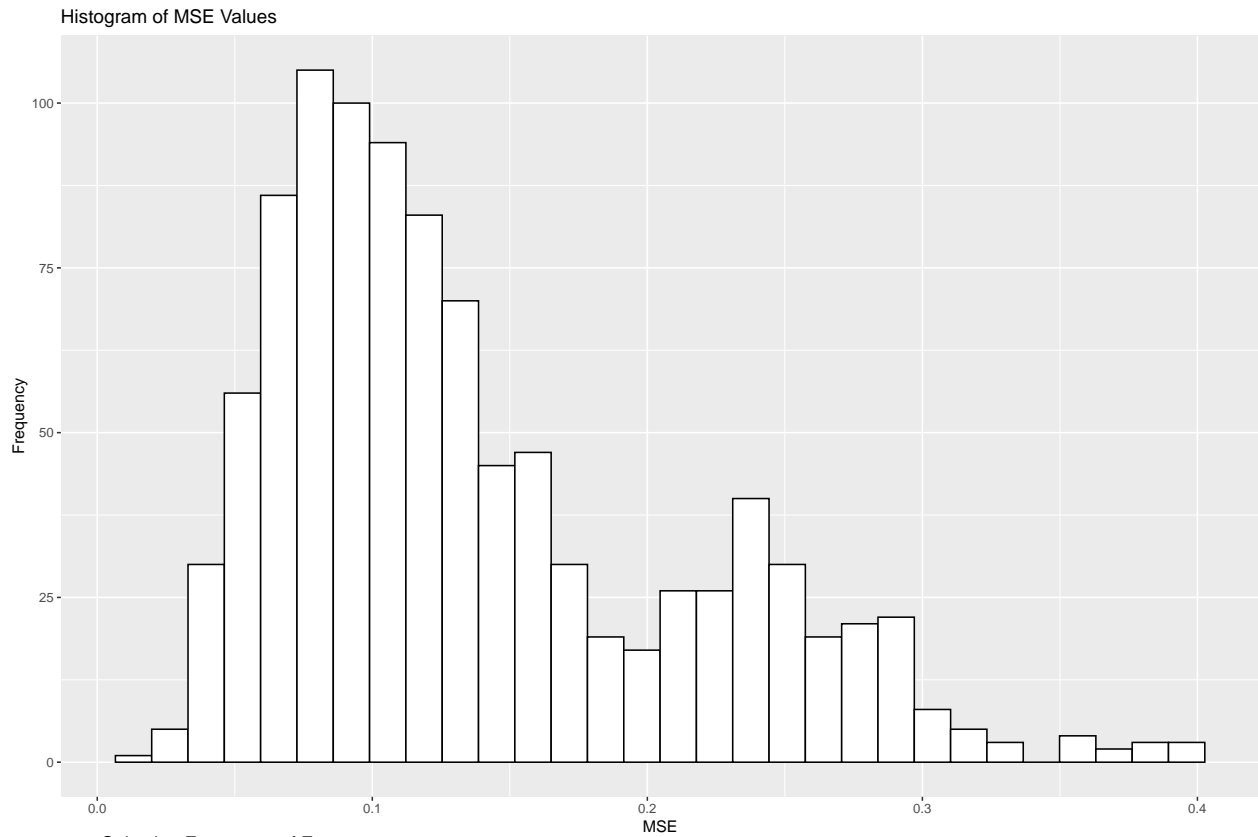
```
## Genes selected 50% or times:
## BMPR1B CA12 CACNA1H EFNA3 GATA3 HOXA7 HOXA9 LEFTY2 NSD1
## Top 10 featrues:
## GATA3 LEFTY2 CACNA1H HOXA9 BMPR1B NSD1 CA12 EFNA3 HOXA7 CAMK2B
```

Post Lasso

Ridge

Elastic Net

Boosting with stumps as base learner