9890 GROUP PROJECT

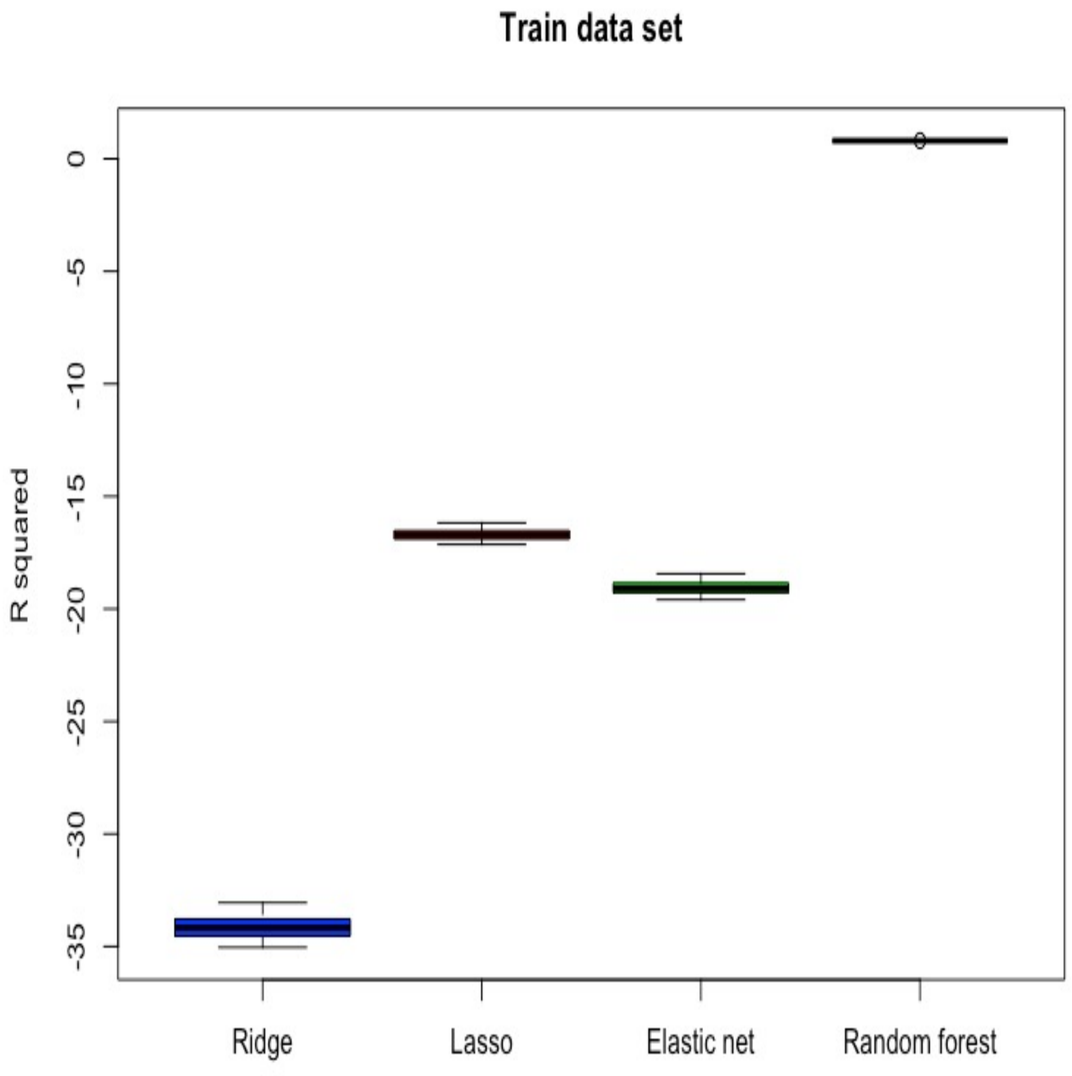
TEAM 15 –ANNA BAE, AUGUSTIN NARE

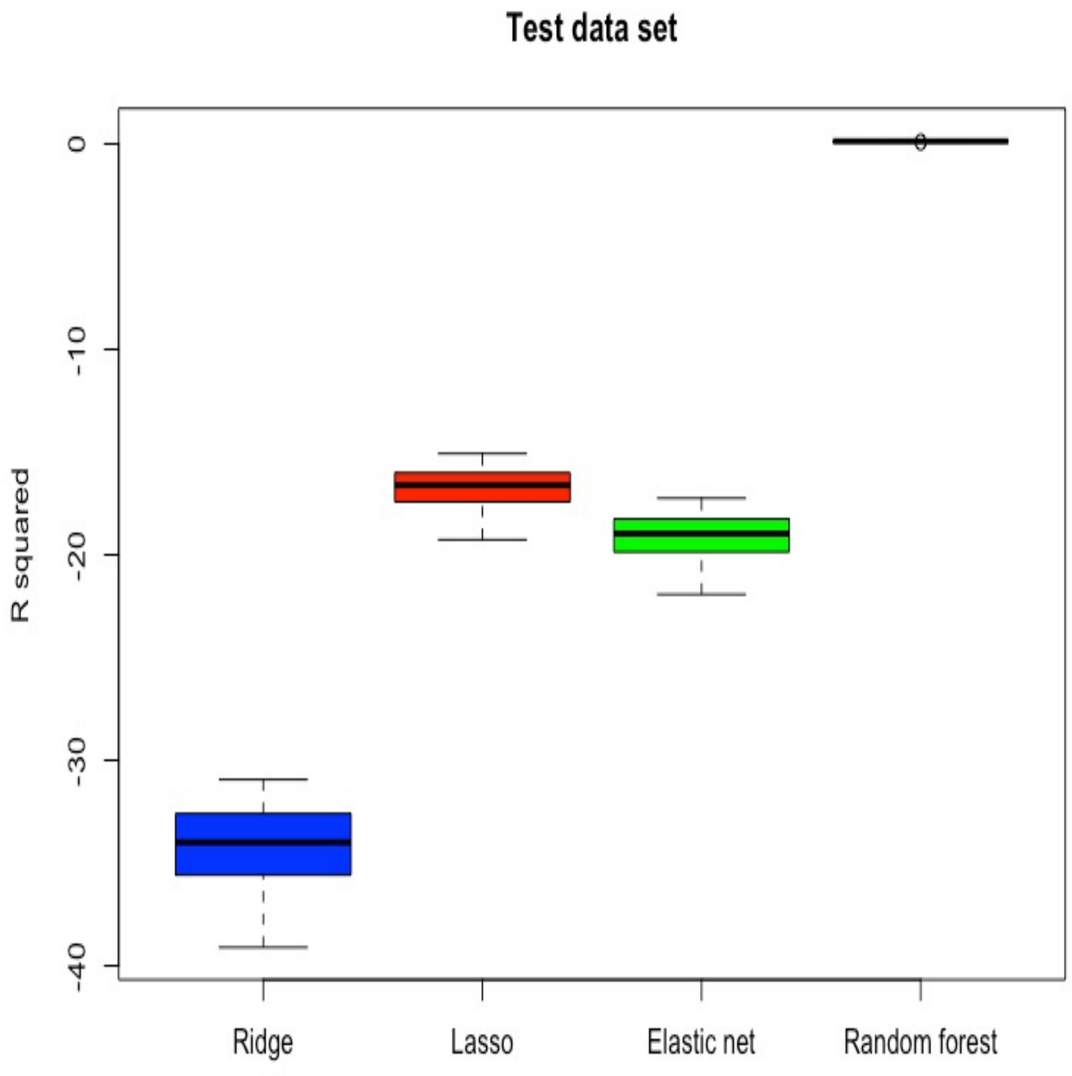
INTRODUCTION TO DATA SET

- The data set, US_birth_2018, contains Information about the natality in the United States in 2018 from https://www.kaggle.com/des137/us-births-2018.
- N(sample size) = 5000
- P = 40
- 16 numerical variables, 24 categorical variables.
- y = Birth weight detail in gram.

- We'd like to predict the birth weight based on parents' information such as parents' education level, race, mothers' BMI, height, etc.

BOX PLOTS OF R^2 TRAIN AND R^2 TEST



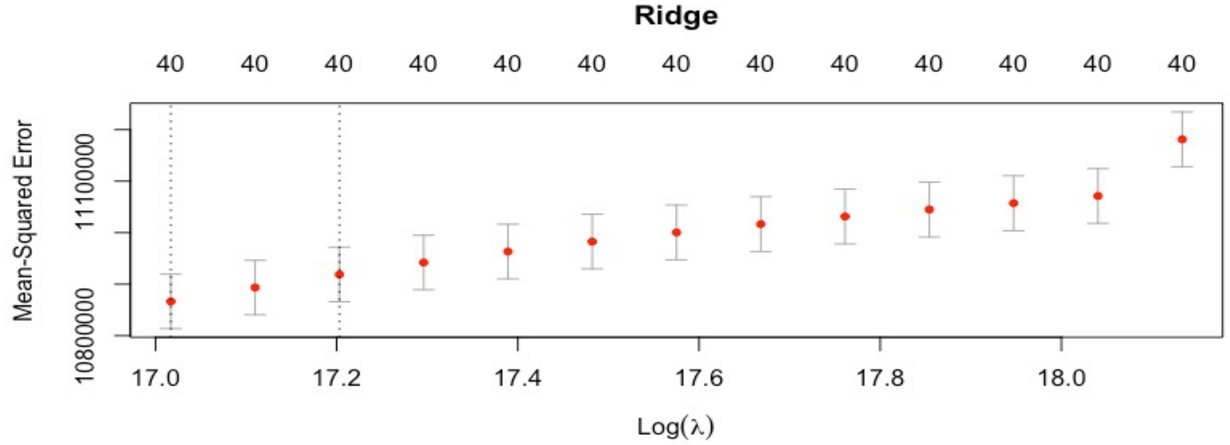


10-FOLD CV CURVES FOR ONE OF THE 100 SAMPLES

The smallest MSE for **Ridge regression** is 10,866,345

Lambda min: 24567164
Log(lambda min) : 17.01692

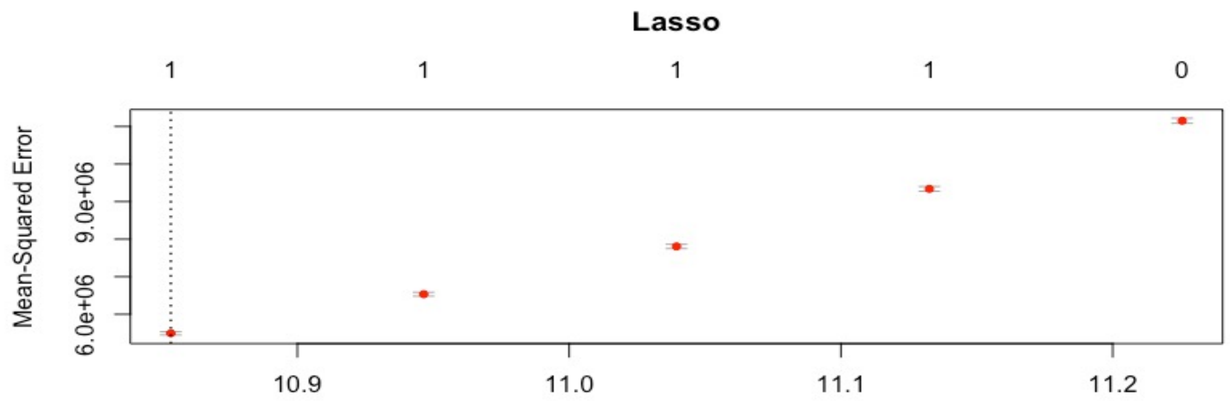
Avg time:
0.3285648 secs



The smallest MSE for **Lasso regression** is 5,495,843

Lambda min: 51711.53
Log(lambda min) : 10.85344

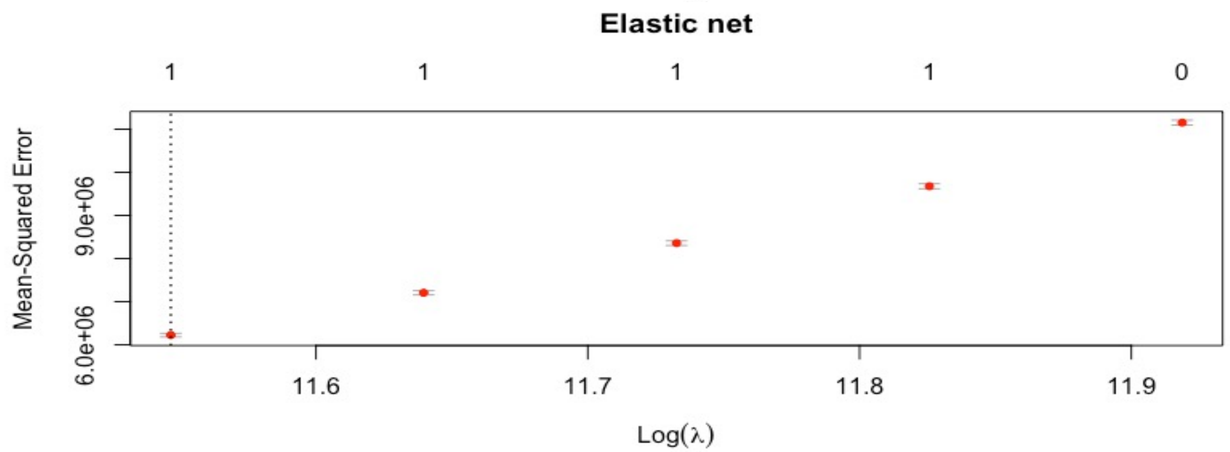
Avg time:
0.2163834 secs



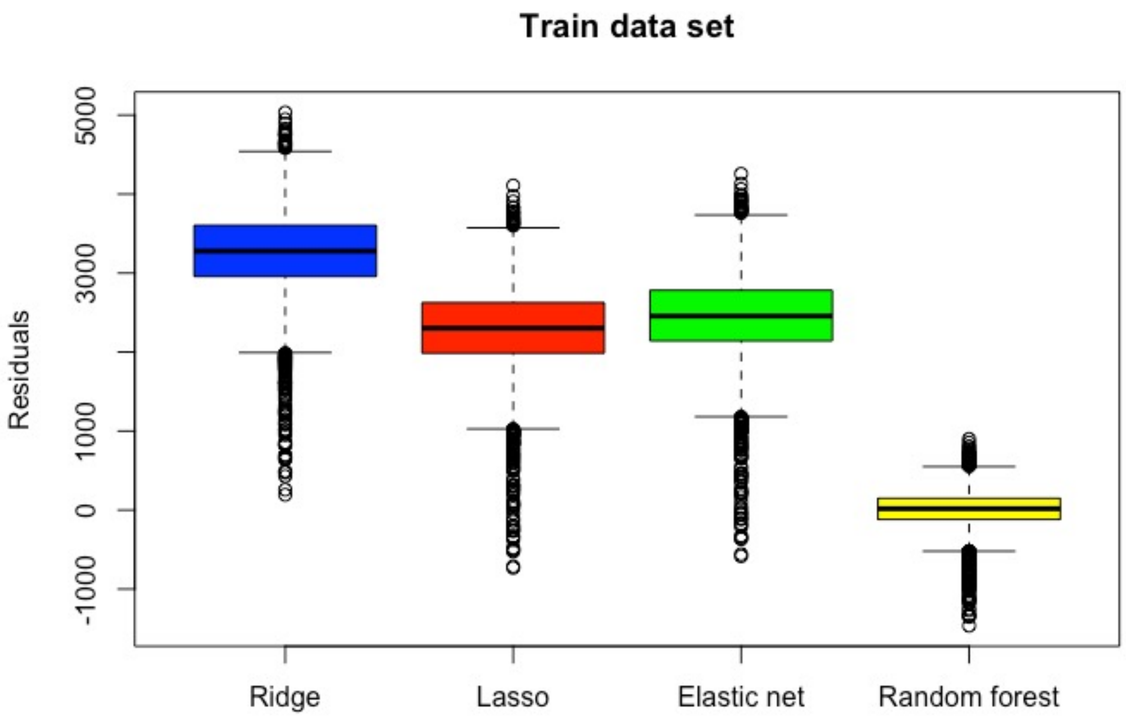
The smallest MSE for **Elastic net regression** is 6,217,850

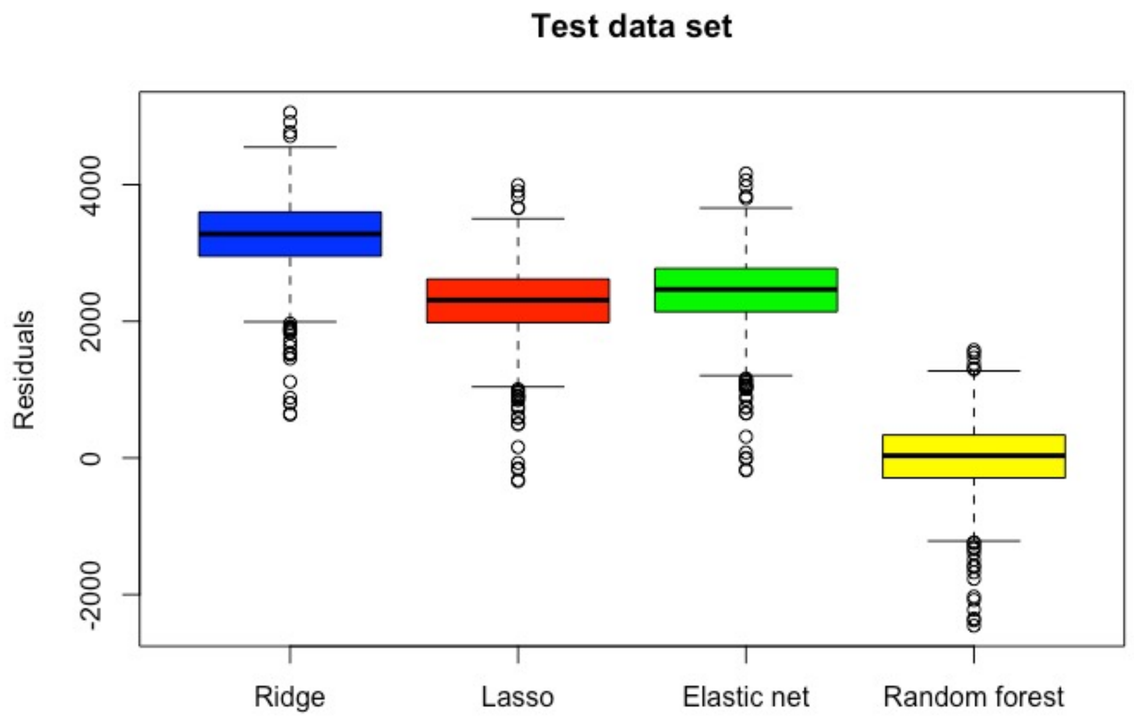
Lambda min: 103423.1
Log(lambda min) : 11.54658

Avg time:
0.2113414 secs



BOXPLOTS OF TRAIN AND TEST RESIDUALS OF ONE OF THE 100 SAMPLES





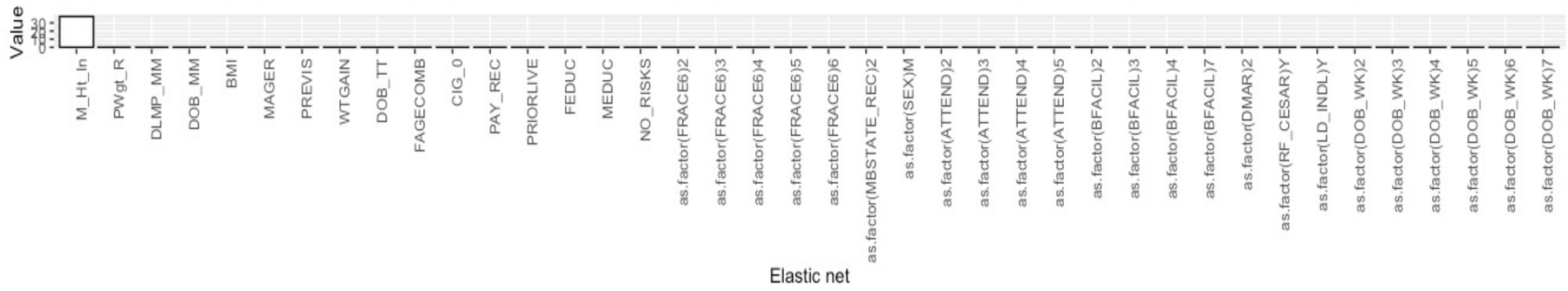
90% TEST R^2 INTERVALS AND TIME TO FIT MODELS

Methods	R^2 test 90% interval (5%, 95%)		Time(secs)
Ridge	-37.60701	-31.32745	0.4237831
Lasso	-18.47137	-15.26618	0.325855
Elastic Net	-21.06208.	-17.46165	0.4964371
Random Forest	0.1024872	0.1432783	49.32369

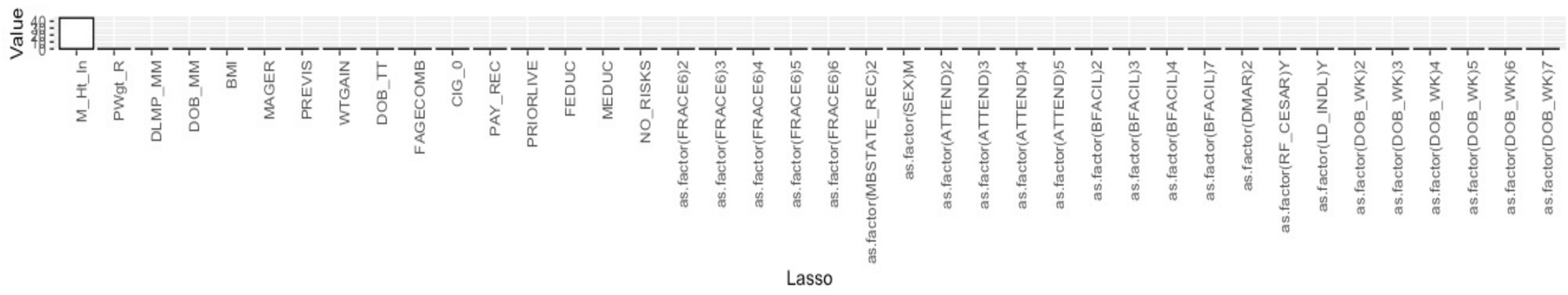
- Trade-off:
  Random forest's test R^2 has the highest 90% interval range of values than among the four models. However, it takes the longest time to fit the model.

BAR-PLOTS OF THE ESTIMATED COEFFICIENTS AND THE IMPORTANCE OF THE VARIABLE.

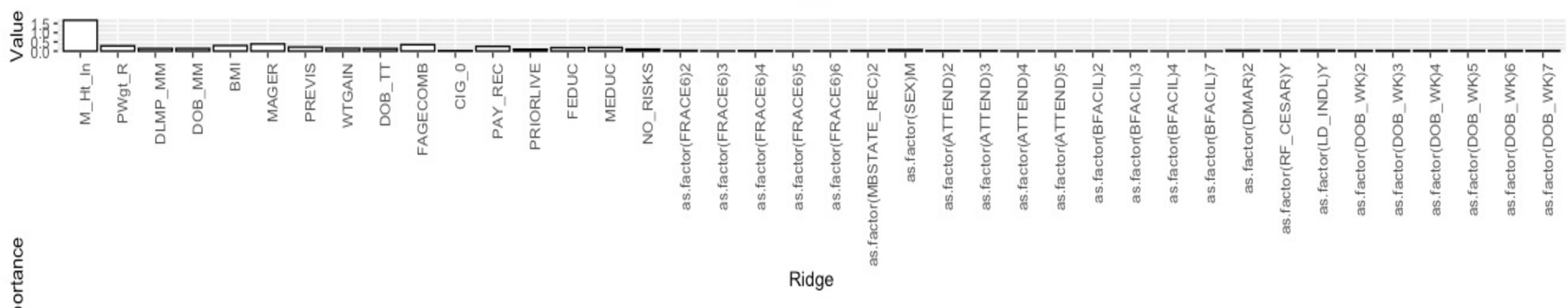
Elastic net
M_Ht_In: 37.94869



Lasso
M_Ht_In: 44.71495

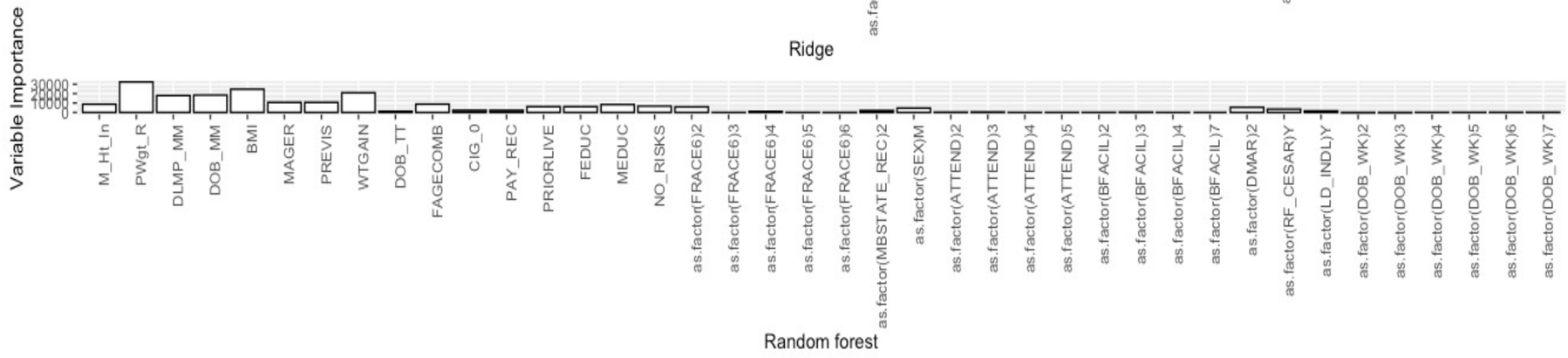


Ridge
M_Ht_In: 1.67369258
MAGER: 0.3895588
FAGECOMB: 0.3514663



Random Forest

	%IncMSE
PWgt_R:	32384.4232
BMI:	24638.0814
WTGAIN:	20913.8666



CONCLUSIONS

- The Random forest has the best predictive performance.
  - Its test data set residuals are closer to 0 than those of other models.
  - Its 90% test R^2 interval has higher values than those of other models.

- https://github.com/AnnaBae92/STA9890-Group-Project