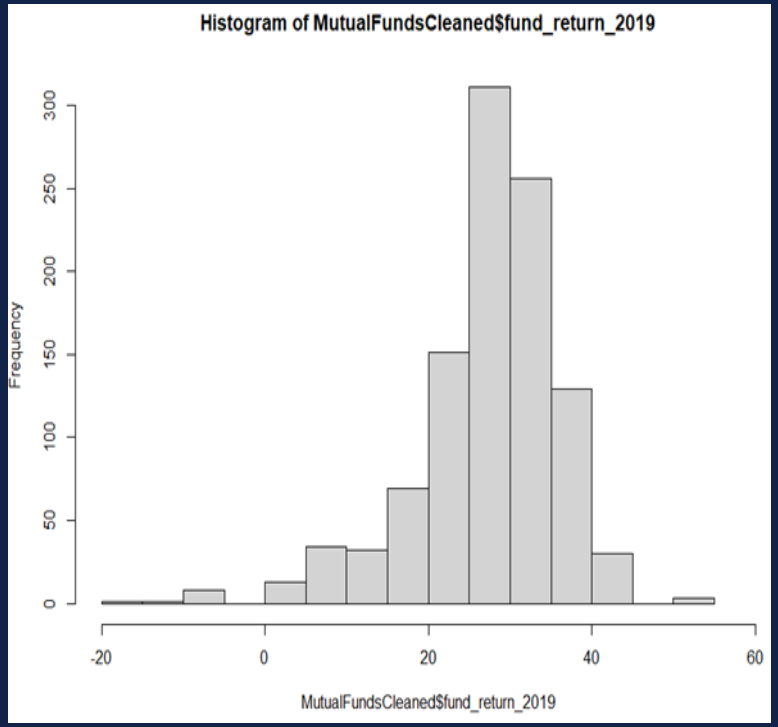
Predicting the 2019 Return of Funds Against US Funds Data

Group 12: Sabah Javaid and Yulia Starovoytova

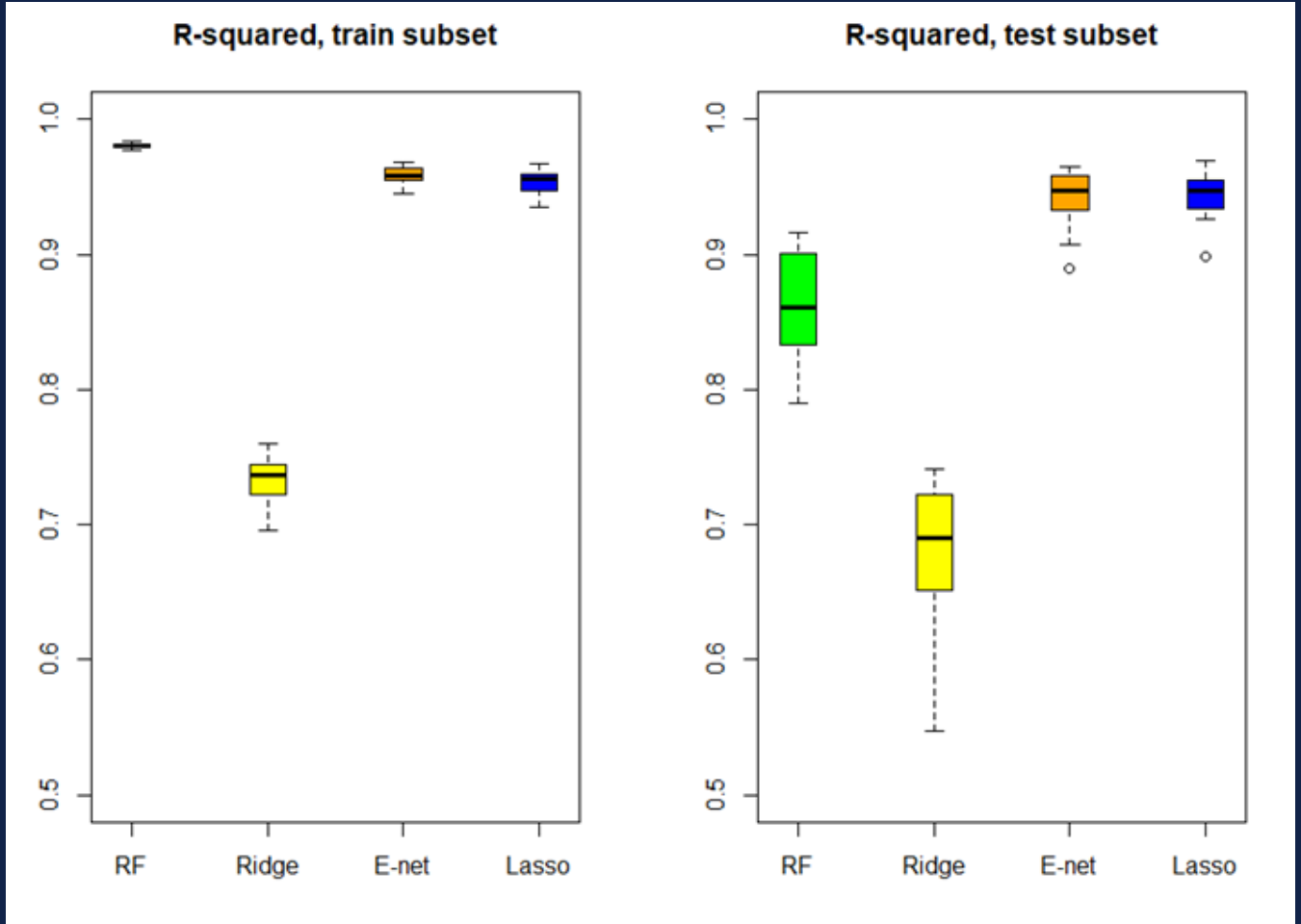
GitHub Link: https://github.com/SabahJavaid/STA-9890

Data Description

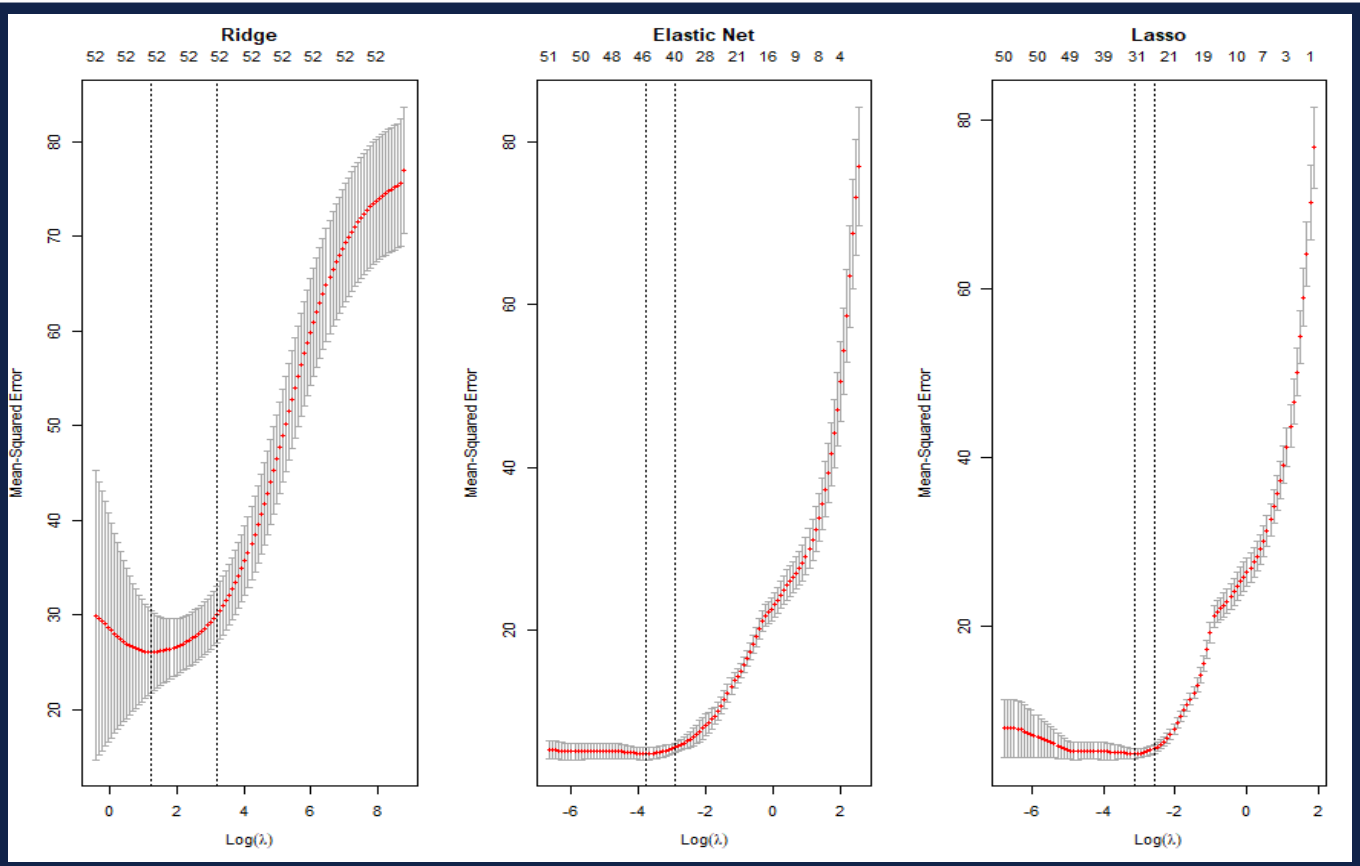
- **Dataset:** US Funds Dataset from Yahoo Finance (Kaggle)
- **Response Variable:** Return of Funds 2019
- The original number of features: p = 173 while sample size equated to n = 24822
- After cleaning and organizing the data, we were left with p = 53 and n = 1308
  - Of p = 53, 45 are numerical and 8 are categorical
- Predictors: Asset_cash, Asset_stock, Price_earnings_ratio, etc.



SIDE-BY-SIDE BOXPLOTS OF R^2_{train} & R^2_{test}

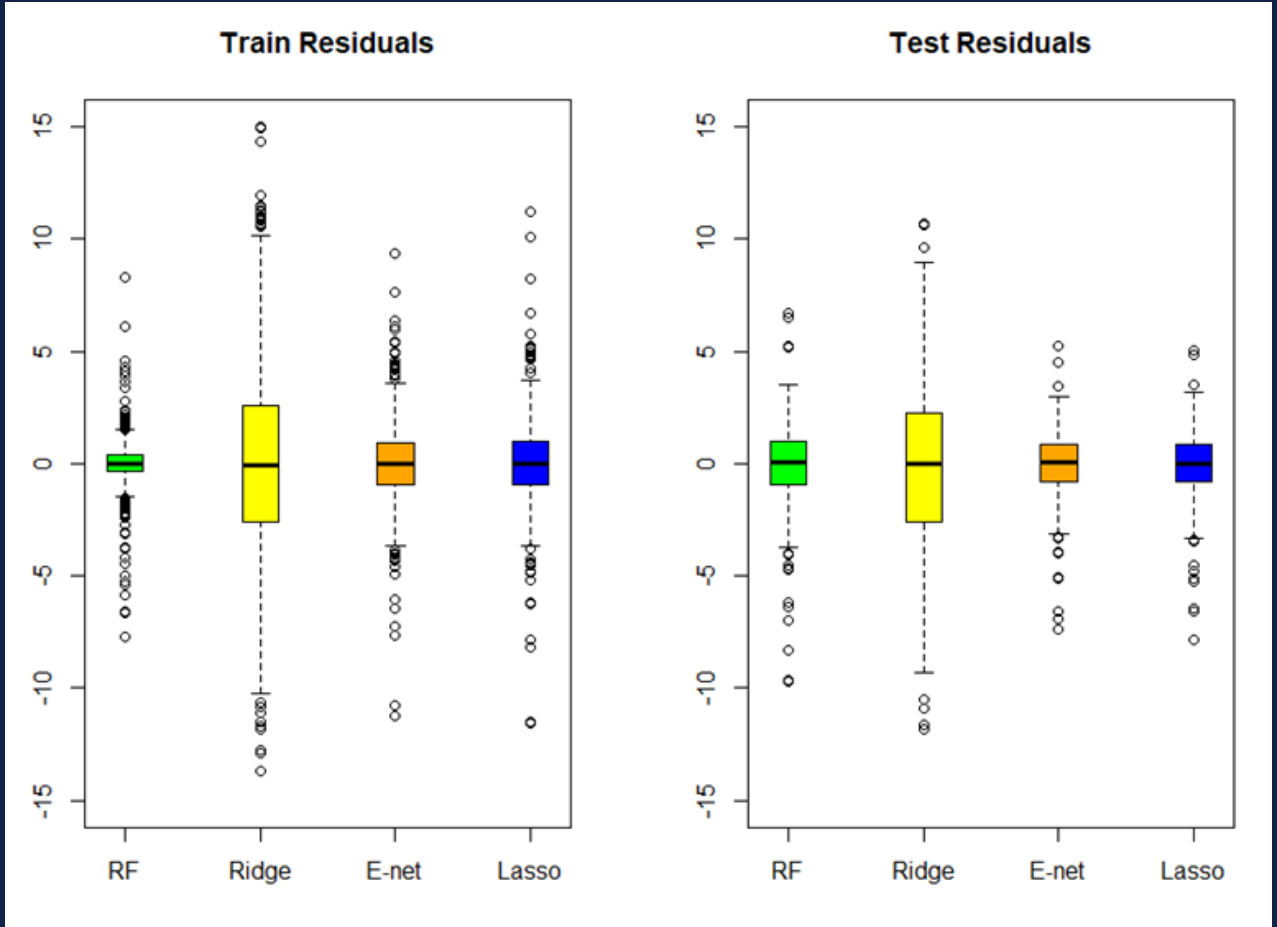


10-Fold Cross Validation Curves



Time (sec)	Ridge: 0.17	Elastic Net: 0.20	Lasso: 0.31

Residual Boxplots of a Random Sample



Models' Fitting Time & Performance

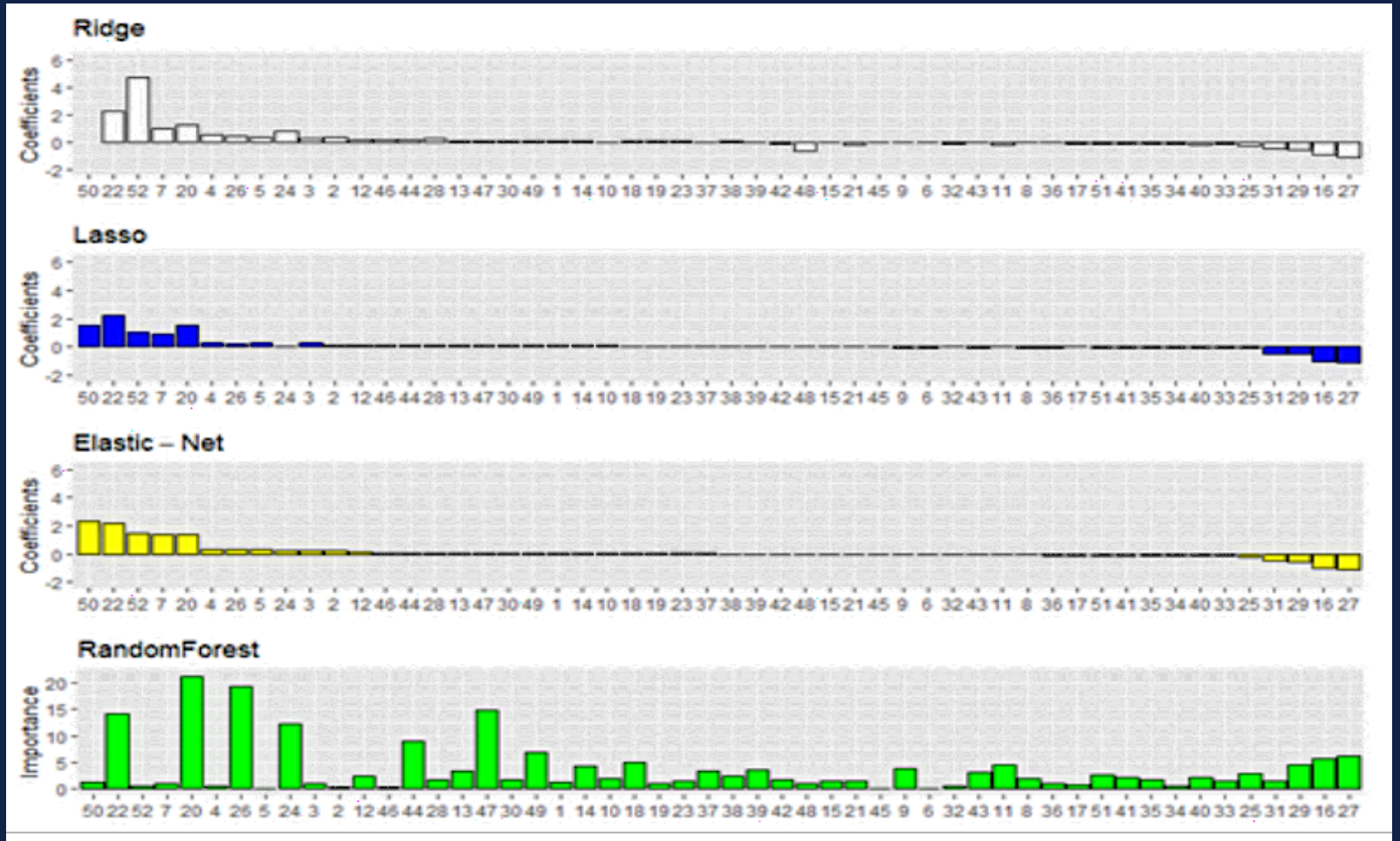
Model	90% test R-squared Confidence Interval	Time (CV parameter tuning + fitting the model)
Ridge	(0.6419336, 0.7364171)	0.24
Lasso	(0.9195771, 0.9497027)	0.32
Elastic Net	(0.9093348, 0.9520546)	0.35
Random Forest	(0.8501102, 0.9105110)	11.41

*** The CI range of the Ridge is much wider comparing to other models

*** Lasso and El-Net showing almost same results in test R-squared and runtime

*** Random Forest takes the longest runtime with still high R-squared value

Bar Plots of Estimated Coefficients



Closing Remarks

Two largest coefficients for El-Net, Lasso and Ridge (positive):

- Category Treynor ratio for 5 years (highlights the risk-adjusted profits)
- Fund return for 5 years

(negative):

- Fund return YTD (to assess the performance of a portfolio)
- Fund return 2018

Two most important variables for Random Forest model:

- Median point of market capitalization (an indicator of the size of companies in which fund invests)
- Category return 2019

Conclusion:

- Three of the models (Lasso, El-Net and Random Forest) show very high levels of test R-squared, over 90% and relatively short runtime (from 0.32 to 11.41 sec).
- Our Mutual funds dataset includes many features with trends over time (ex. a fund or a category return percentage for different years or range of years), which in our opinion could be an explanation for the inflated R-squared values.
- Based on the estimated coefficients and importance of the parameters analysis there are also other good predictors in this dataset.