STA 9890
Department of Statistics

Regression Analysis To Predict Financial Distress

TANAY MUKHERJEE

FALL, 2020

Data Dictionary

Y variable: Financial Distress [Column 3]

X1 net profit / total assets
X2 total liabilities / total assets
X3 working capital / total assets
X4 current assets / short-term liabilities
X5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
X6 retained earnings / total assets
X7 EBIT / total assets
X8 book value of equity / total liabilities
X9 sales / total assets
X10 equity / total assets
X11 (gross profit + extraordinary items + financial expenses) / total assets
X12 gross profit / short-term liabilities
X13 (gross profit + depreciation) / sales
X14 (gross profit + interest) / total assets
X15 (total liabilities * 365) / (gross profit + depreciation)
X16 (gross profit + depreciation) / total liabilities
X17 total assets / total liabilities
X18 gross profit / total assets
X19 gross profit / sales
X20 (inventory * 365) / sales
X21 sales (n) / sales (n-1)
X22 profit on operating activities / total assets

X23 net profit / sales
X24 gross profit (in 3 years) / total assets
X25 (equity - share capital) / total assets
X26 (net profit + depreciation) / total liabilities
X27 profit on operating activities / financial expenses
X28 working capital / fixed assets
X29 logarithm of total assets
X30 (total liabilities - cash) / sales
X31 (gross profit + interest) / sales
X32 (current liabilities * 365) / cost of products sold
X33 operating expenses / short-term liabilities
X34 operating expenses / total liabilities
X35 profit on sales / total assets
X36 total sales / total assets
X37 (current assets - inventories) / long-term liabilities
X38 constant capital / total assets
X39 profit on sales / sales
X40 (current assets - inventory - receivables) / short-term liabilities
X41 total liabilities / ((profit on operating activities + depreciation) * (12/365))
X42 profit on operating activities / sales
X43 rotation receivables + inventory turnover in days
X44 (receivables * 365) / sales

X65 sales/ equity
X66 total defaulters/ total loans sanctioned
X45 net profit / inventory
X46 (current assets - inventory) / short-term liabilities
X47 (inventory * 365) / cost of products sold
X48 EBITDA (profit on operating activities - depreciation) / total assets
X49 EBITDA (profit on operating activities - depreciation) / sales
X50 current assets / total liabilities
X51 short-term liabilities / total assets
X52 (short-term liabilities * 365) / cost of products sold)
X53 equity / fixed assets
X54 constant capital / fixed assets
X55 working capital
X56 (sales - cost of products sold) / sales
X57 (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X58 total costs /total sales
X59 long-term liabilities / equity
X60 sales / inventory
X61 sales / receivables
X62 (short-term liabilities *365) / sales
X63 sales / short-term liabilities
X64 sales / fixed assets

X65 sales/ equity
X66 total defaulters/ total loans sanctioned

Dataset: https://www.kaggle.com/shebrahimi/financial-distress

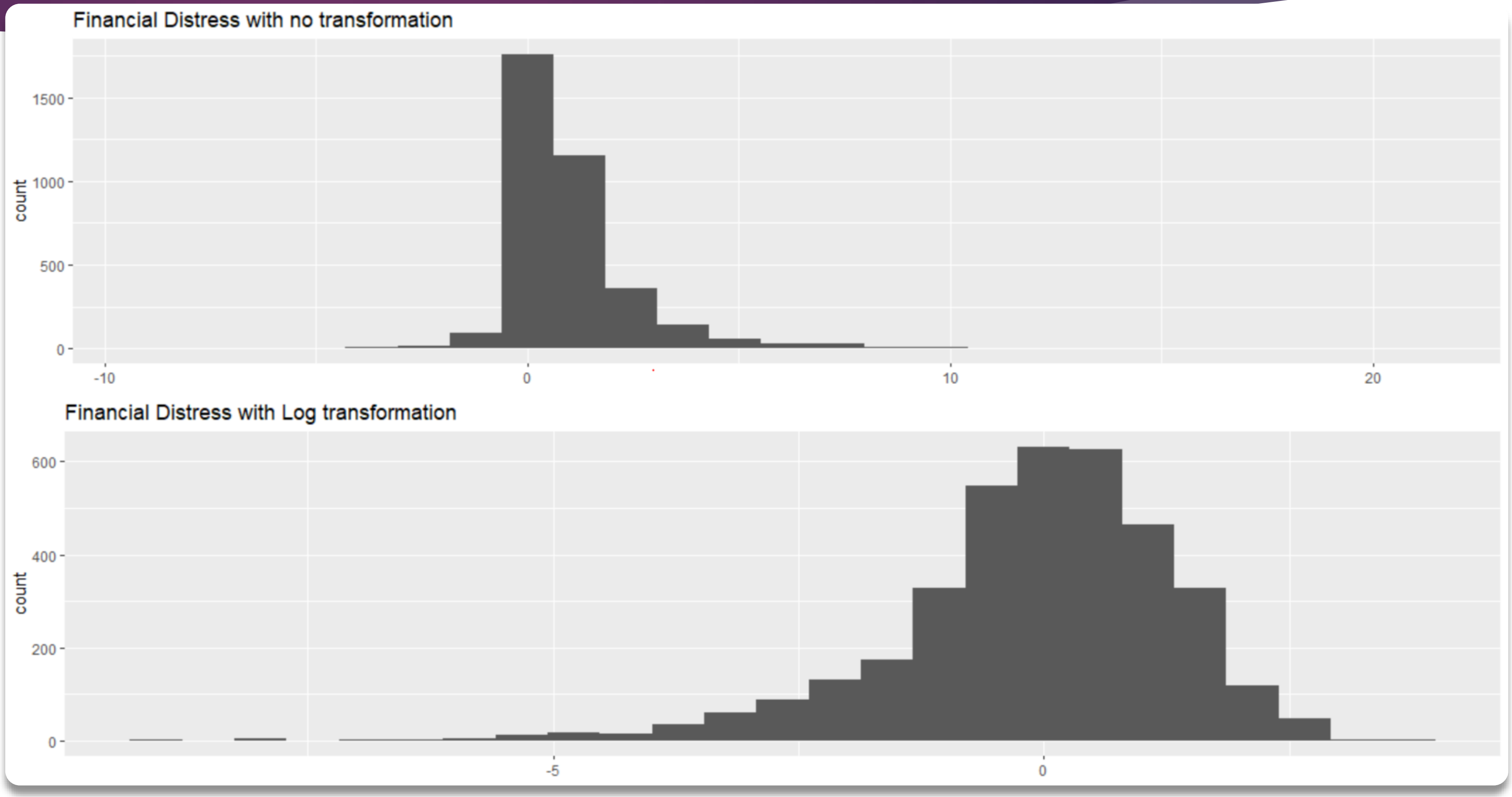
Understanding the data

1. The financial distress is our y-variable that we will try to predict.
2. It is mostly between -2 to +2 and is given right skewed.
3. We try to transform he variable by taking a log transformation of our response variable.
4. Total features are 66, and total observations are 3671.
5. The log transformation equation:

   y = log(y + 1 – min(y))

6. To further normalize it make it more close to the standard normal curve we can also do:

   y = log(square-root(y^2)) + c

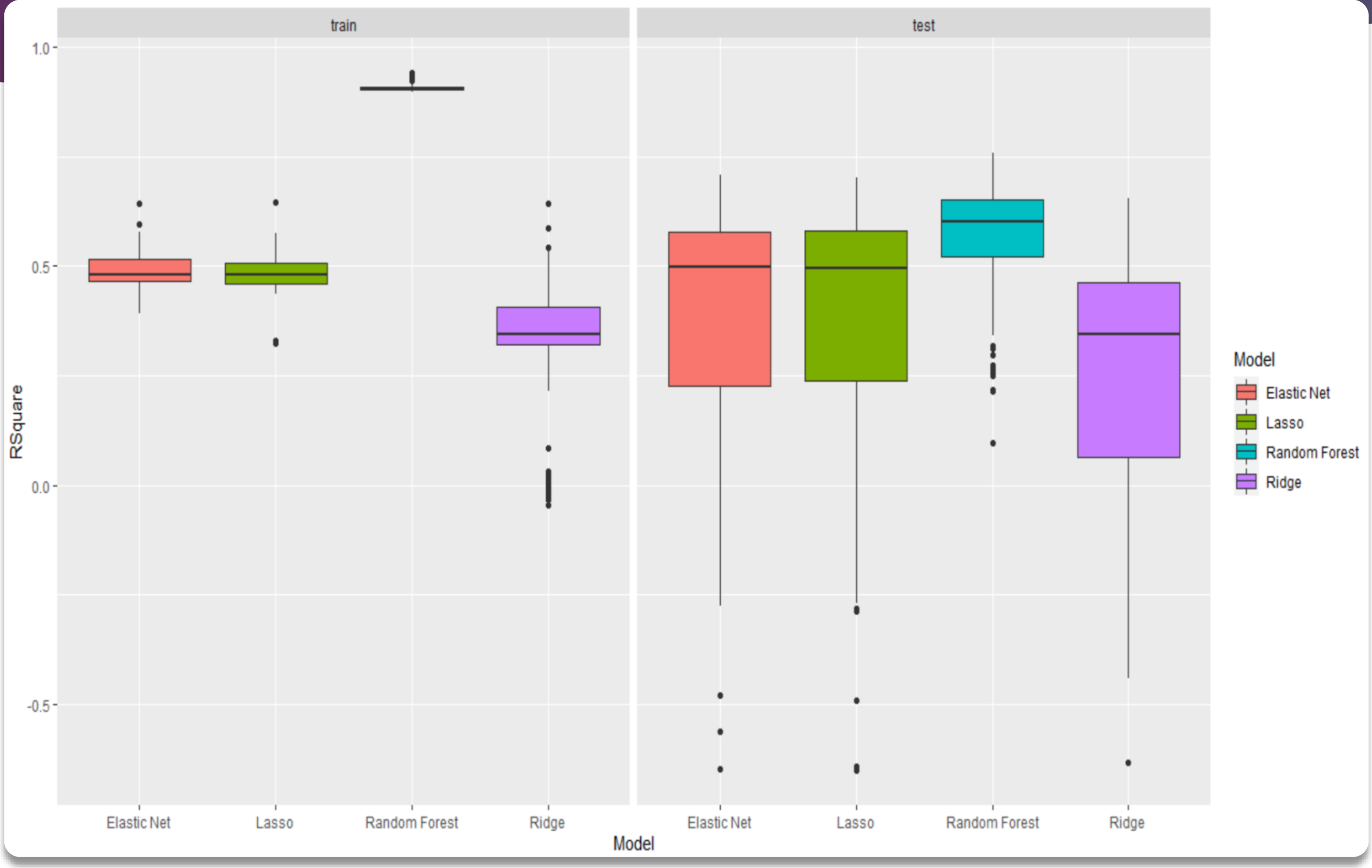


Box Plots of R-Square Value for all 4 models

1. The data was split into training and test data set. Training set is 80% of all the observations and test set is the remaining 20%
2. Run a simulation of 100 samples
3. We fit the data for 4 regression models – Elastic Net, Lasso, Random Forest and Ridge.
4. We use the following equation to calculate our R-square value:

$$R^2_{test} = 1 - \frac{\frac{1}{n_{test}} \sum_{i \in D_{test}} (y_i - \hat{y}_i)^2}{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Model	R-Square (median)	
	Train	Test
Elastic Net	0.46	0.48
Lasso	0.44	0.48
Random Forest	0.9	0.63
Ridge	0.38	0.36

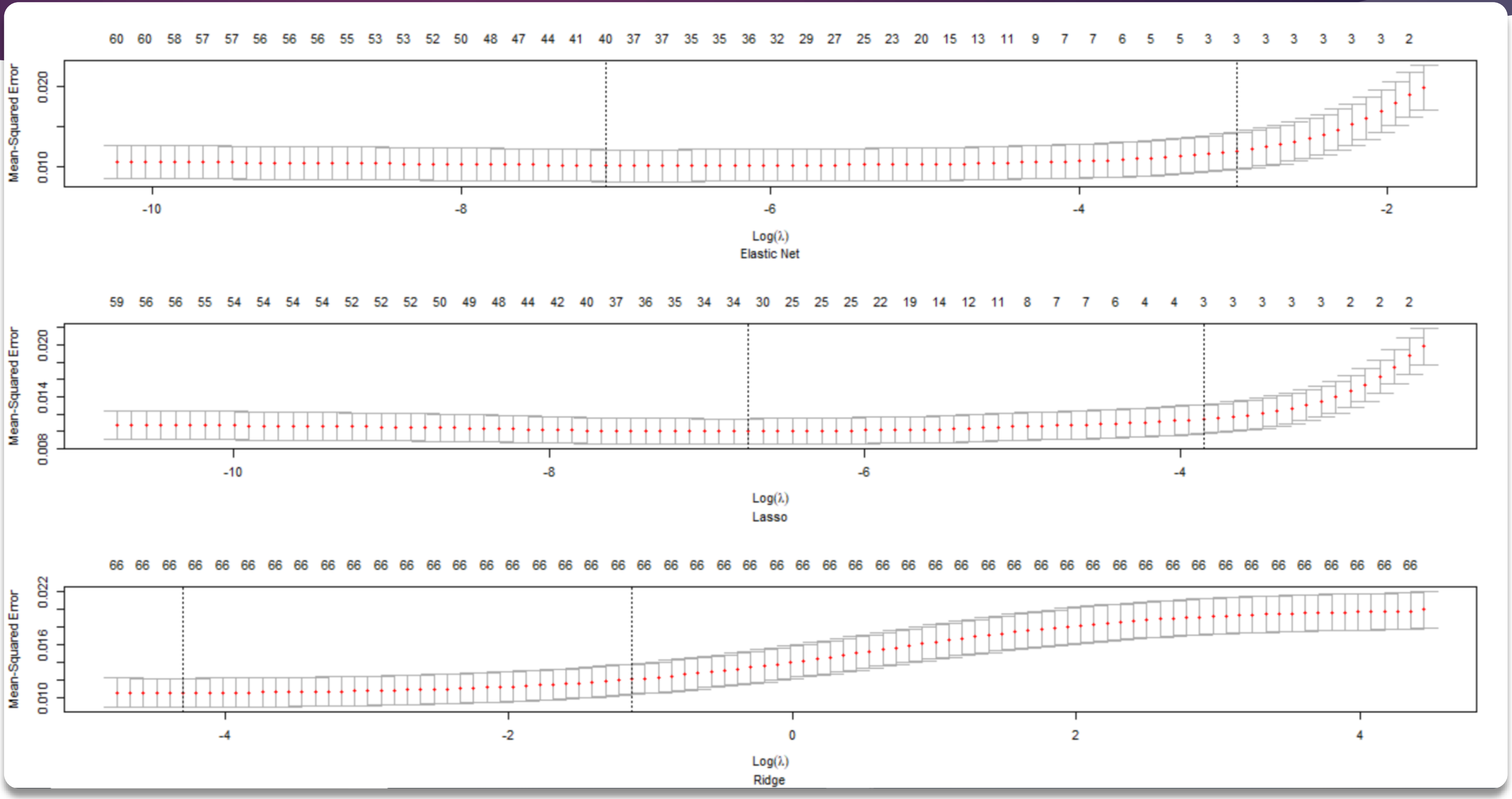


10-fold CV curves for Elastic Net, Lasso, Ridge

1. We use 10-fold cross validation to tune in all the lambdas.

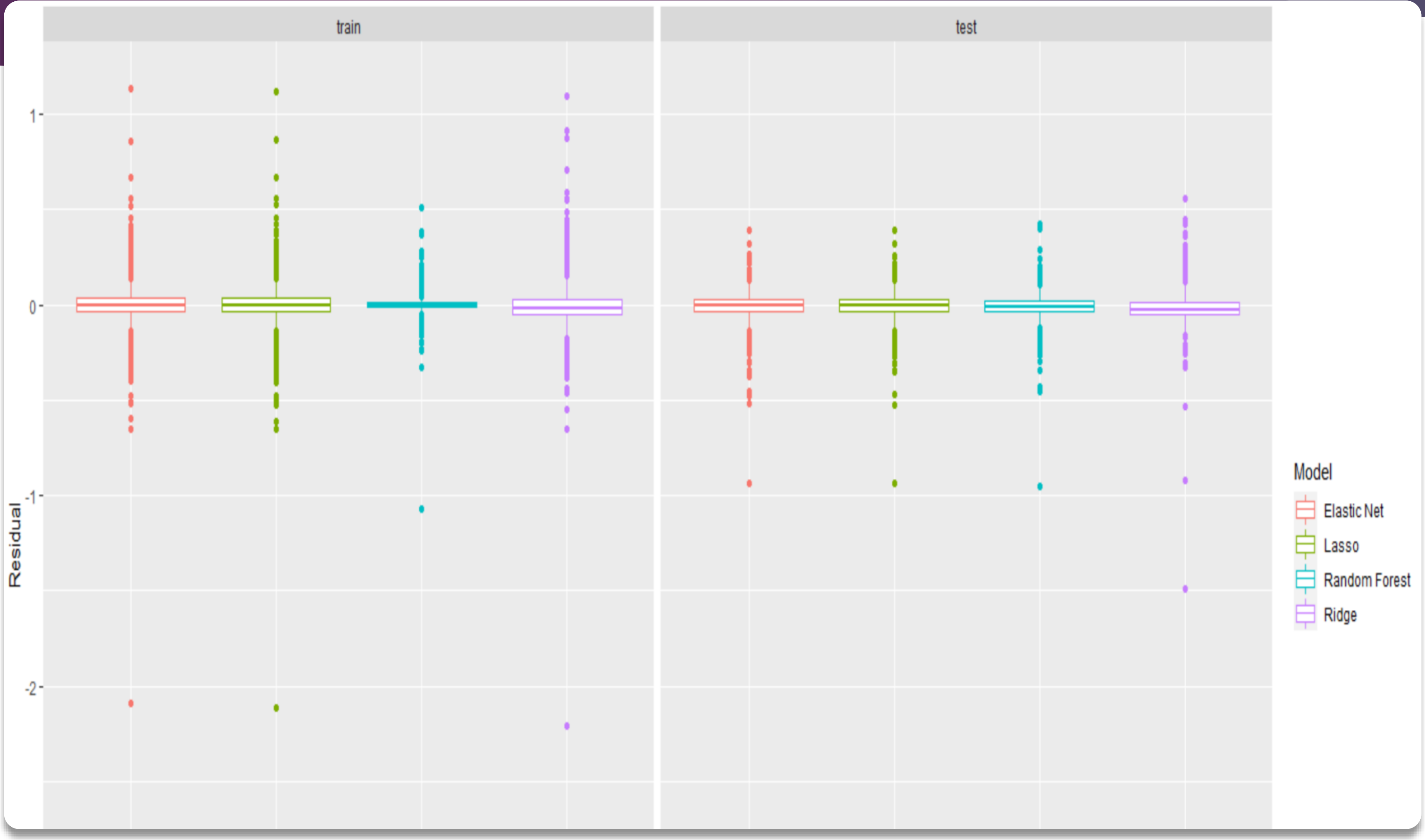
2. Elastic Net uses 40 features, Lasso uses 32 features whereas Ridge uses all 66 features.

Model	Minimum lambda
Elastic Net	0.0006014624
Lasso	0.0006947281
Ridge	0.05255364



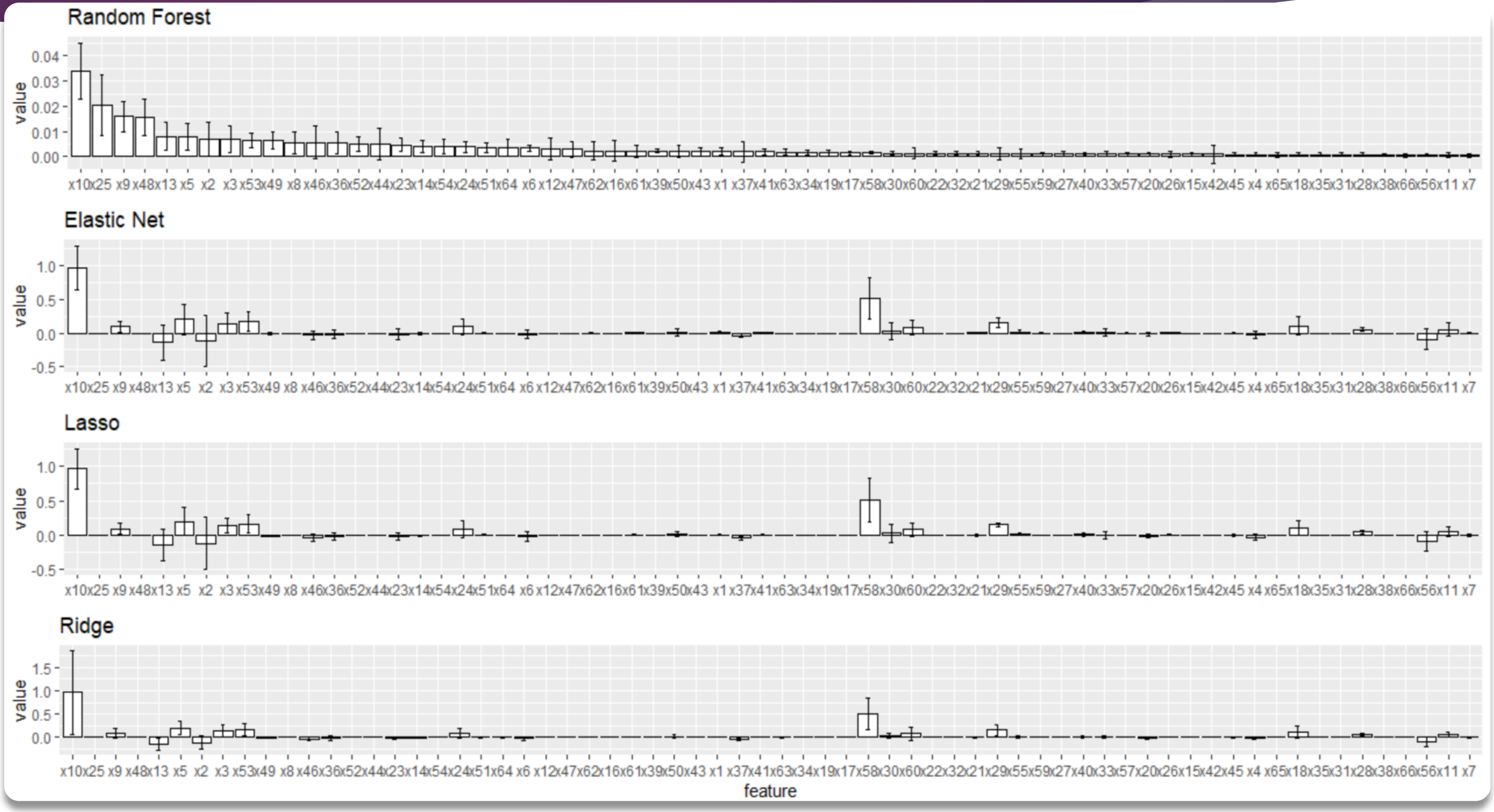
Box Plots of Residuals for all 4 models

1. Residual plot has similar results for all the 4 models and the median is close to 0.
2. The observations are not much different between train and test set.
3. Random Forest has more compact results with the smallest interquartile range. This can also be attributed to the fact that Random Forest is probably overfitting the variables. That means the model has extra capacity to pick the random noise in the observation.



Bar plots with bootstrapped error bars

1. Ran a simulation of 100 for bootstrap exercise.
2. Arranged all the bar plots with errors in descending order of output from random forest.
3. Variable X10 appears to be a prominent feature for all the models which is equity/total assets.
4. Also, in general, the observation says, any feature that talks about total expense divided by overall asset value helps in deciding the financial distress.
5. The results from Elastic Net, Ridge and Lasso are similar.



Summary

1. In the preceding slides, we saw that from R-square box-plot Radom Forest seems to be the best model for prediction.
   1. Lasso and Elastic – Net almost show same results without much to differentiate.
   2. Ridge seems to have the maximum variance with quite a few negative R-square and thus is not a great model for predicting financial distress.
2. The residual box plot kind of showed similar results for all the 4 models with median around 0 but the quartile range was lowest for random forest.
   - The residual plots for test show lesser variance and thus are a good reflection of our prediction.
   - More outliers are seen towards positive side of 0 than the negative side of 0, which means the model is more biased with positive values in the dataset.
3. Through bootstrap bar plot, we can see that Random Forest is best at picking the important features that will help us predict the financial distress.
   - It has more non – zero coefficients and is best at picking the important features. In this case X variables – 10, 25 and 9 are the best predictors. Those features basically are:

     x10 – equity/total assets, x25 - (equity - share capital) / total assets and x9 - sales / total assets.
   - For Elastic Net, Ridge and Lasso, again X10 seems to be the most important features whereas they also identify X58 - total costs /total sales as an important feature for prediction.
4. Lastly, looking at the time required for tuning each model, we see that Random Forest takes a lot more time in analyzing the features than other models for prediction. The break down of time is shared on the right-hand side.

Model	Time (in secs)
Elastic Net	0.24
Lasso	0.28
Random Forest	27.46
Ridge	0.27

THANK YOU!