Predicting The Risk of Defaulting Against Credit Risk Variables

Group 9: Sabah Javaid, Augustin Nare and Iftikar Ahmed

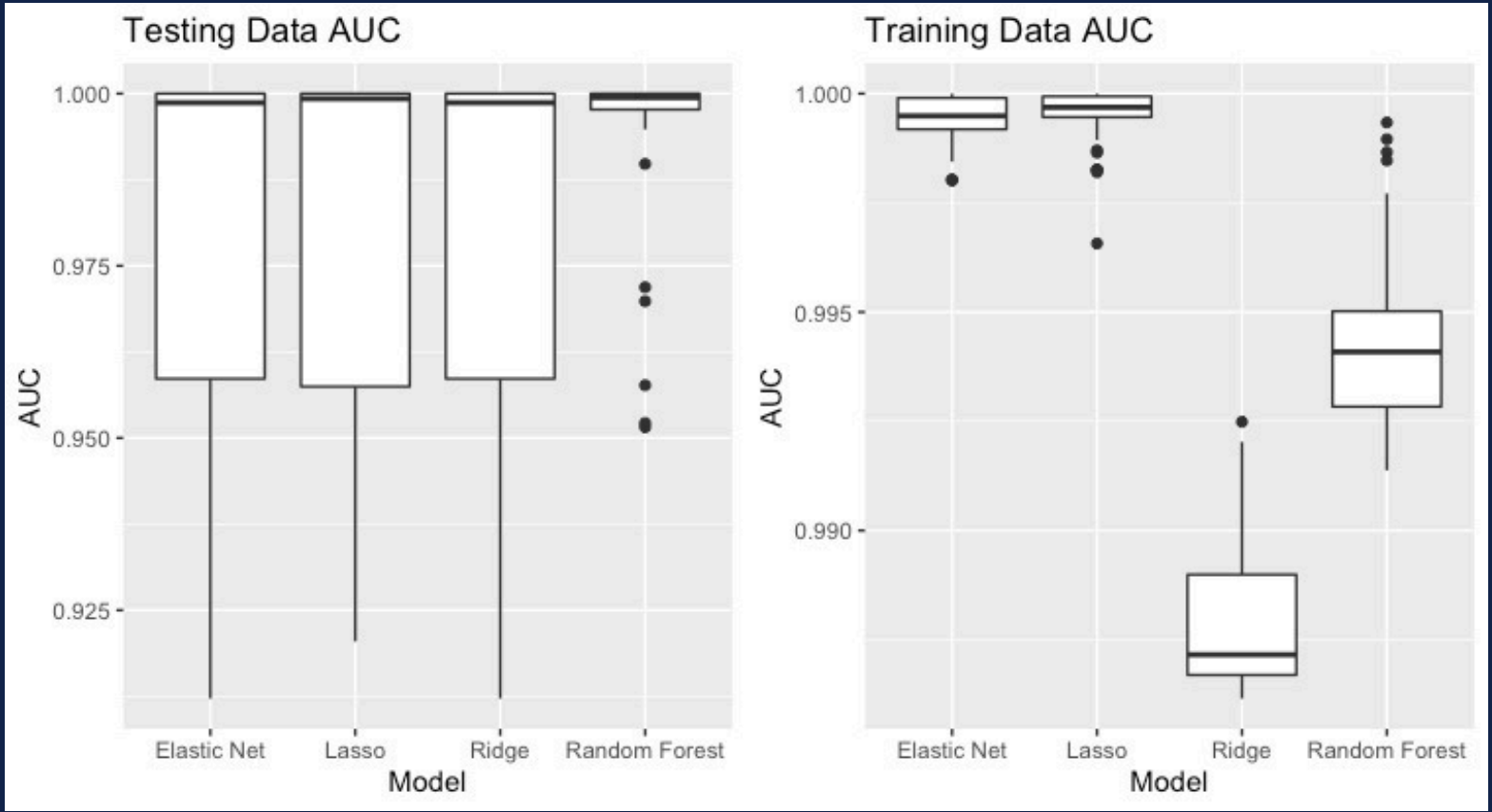
Kaggle Dataset: https://www.kaggle.com/rameshmehta/credit-risk-analysis

GitHub Link: https://github.com/iftahmed/STA-9891-Project

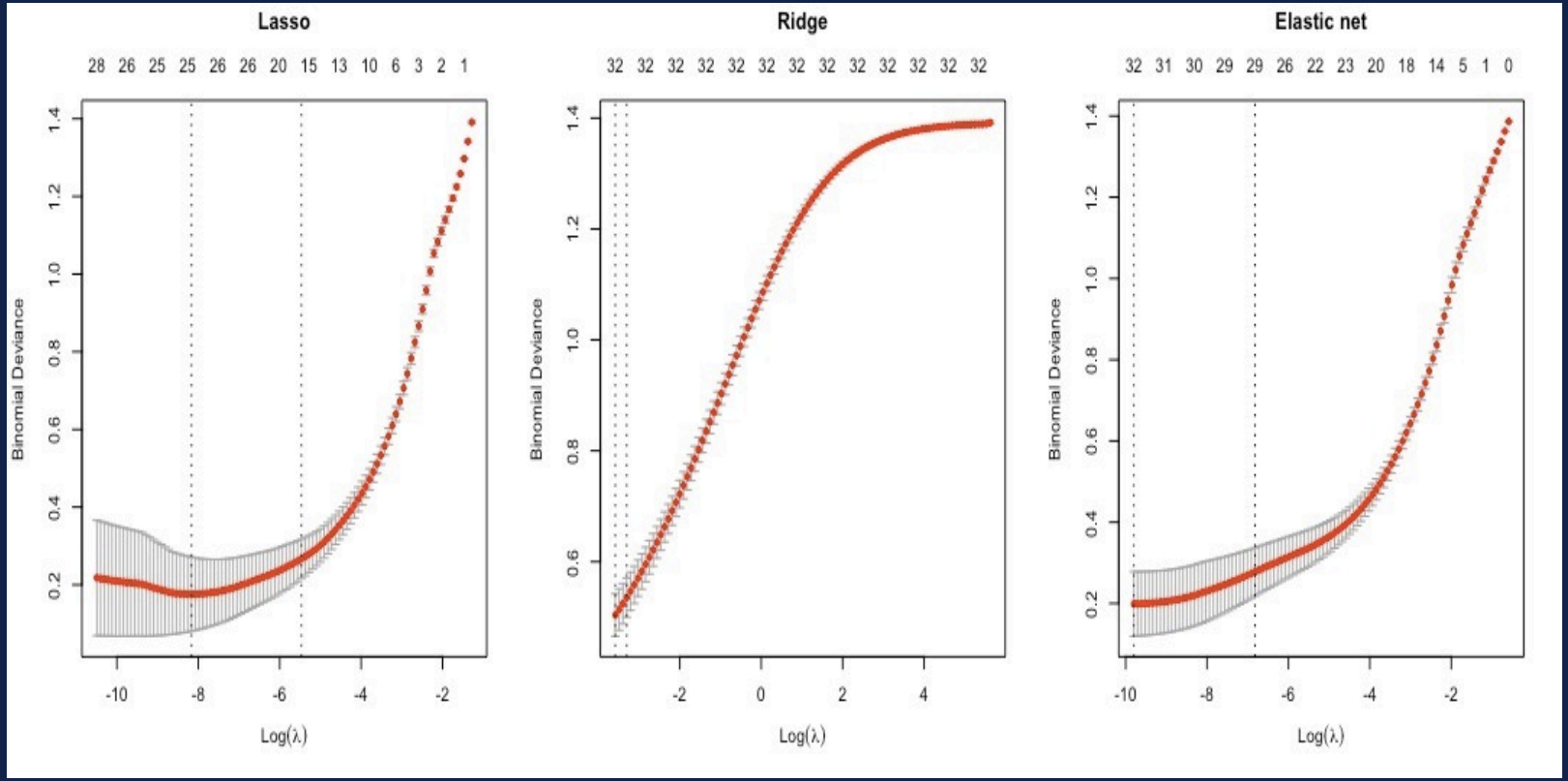
Data Description

- **Dataset:** Credit Risk Analysis(Kaggle)
- **Response Variable:** default_ind
- The original number of features: p = 73 while sample size equated to n = 855,969
- After cleaning and organizing the data, we were left with p = 38 and n = 1000
- Imbalance Ratio: 17.7%
  - N(+)= 177
  - N(-)= 745
- Predictors: loanamnt, annual_inc, grade, installments, etc.

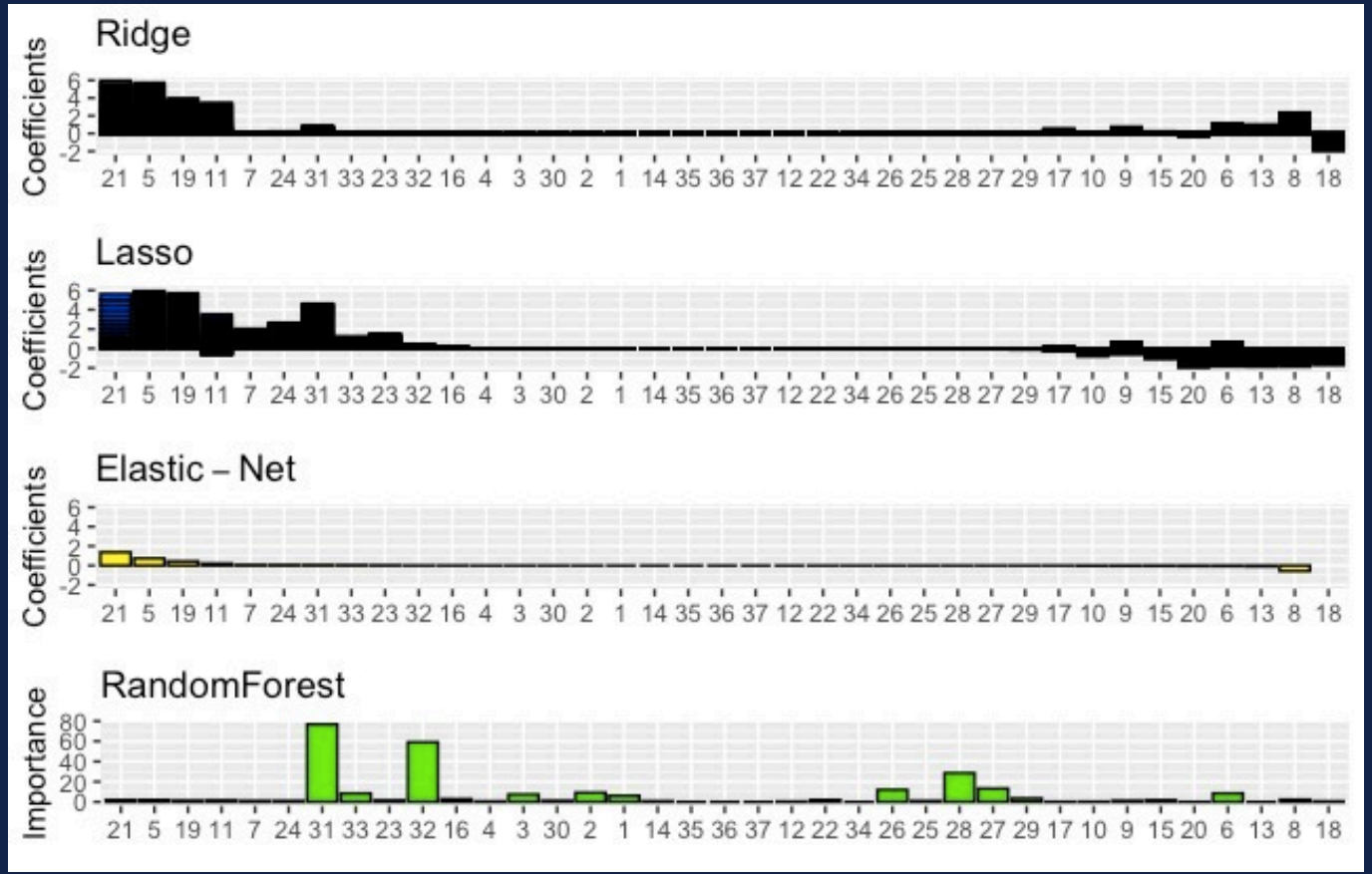
Boxplots: AUC of 50 samples



10-Fold Cross Validation Curves



Bar Plots of Estimated Coefficients



Fitting Time & Performance

Model	Median Testing AUCS	Time (CV parameter tuning + fitting the model)
Ridge	0.9986727	7.46
Lasso	0.9992535	1.33
Elastic Net	0.9986727	4.78
Random Forest	0.9994214	1.05

*** The AUC medians of the Lasso and Random Forest are the greatest, respectively.

*** Ridge and Elastic Net have the same results in testing median AUC however Ridge had a longer runtime in comparison to Elastic net.

***Random Forest has the largest AUC value while its runtime was the shortest.

Closing Remarks

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

- Revolving Balance
- Interest Rate

(negative):

- Last 6 months inquiries
- sub grade

Two most important variables for Random Forest model:

- collection recovery fee
- recoveries

Conclusion:

- While all four models (Lasso, El-Net, Ridge and Random Forest) show very high levels of test AUC, the Random Forest is valued the highest.
- In addition, the runtimes range from 1.05 to 7.46 seconds.
- Regression method that fits the data the best and was high performing:
  - Random Forest with a runtime of 1.05 seconds.