STA9891 – Final Project

A binary classification of Diabetes Patients and the prediction of their Hospital Readmittance Outcomes

By Group 13 - Brian Contreras and Troy Whittemore

https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#
https://github.com/BrianContreras317/9891-Project

Data Overview

This data covers diabetic, inpatient encounters in hospitals with length of stays between 1 and 14 days.

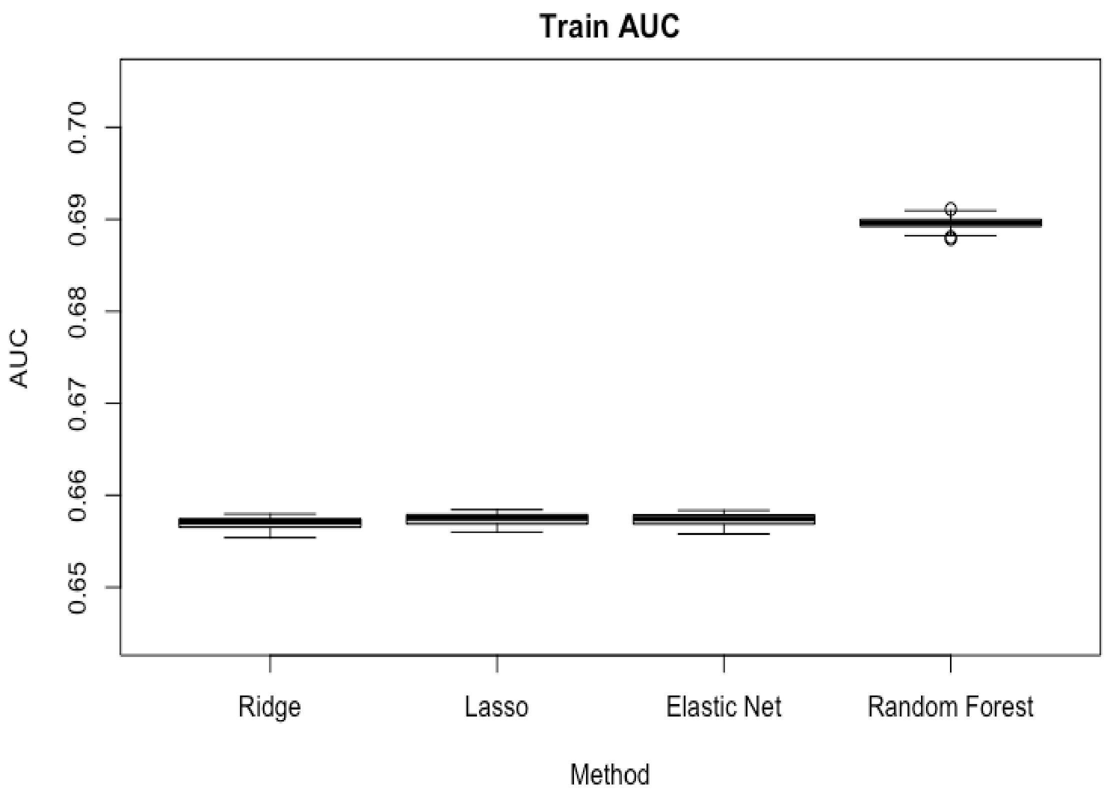
The importance of the analysis of this data is to understand the interaction between the variables within hospitals and the readmission rates of the patients.

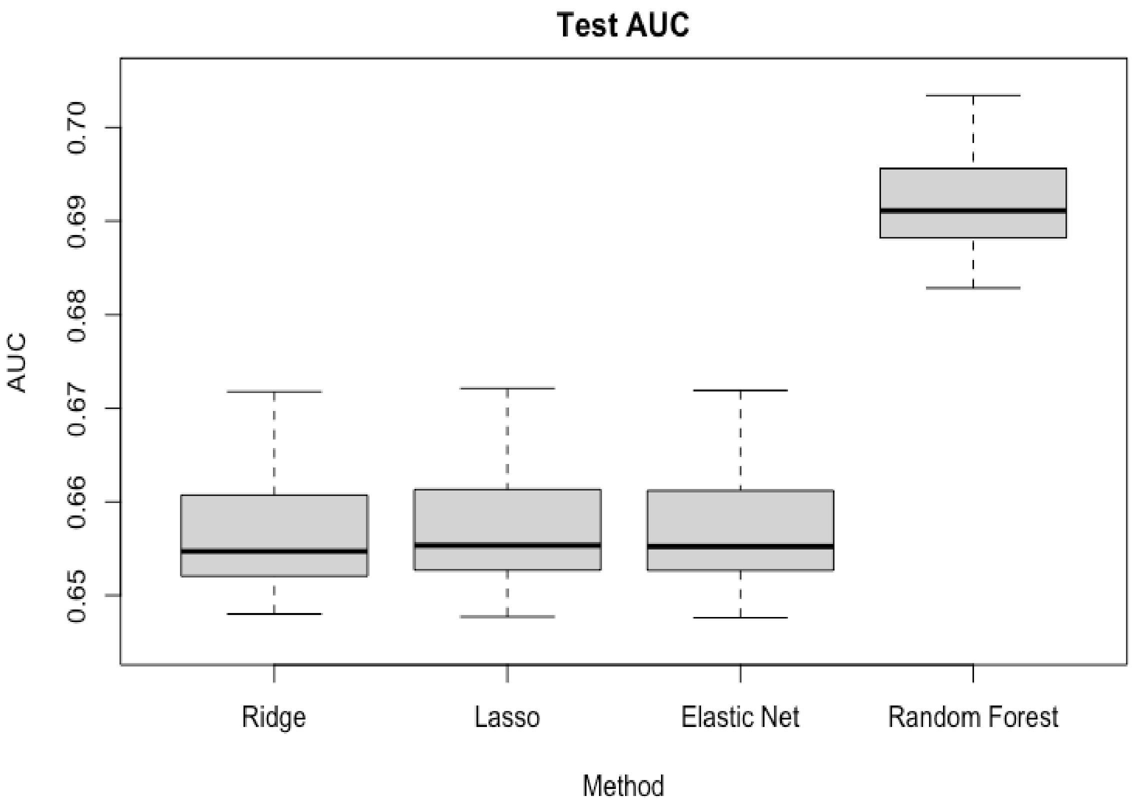
n = 83,845 p = 41

n+ = 38,955 instances of readmission, n- = 44,790 instances of no readmission, roughly a 7:8 ratio

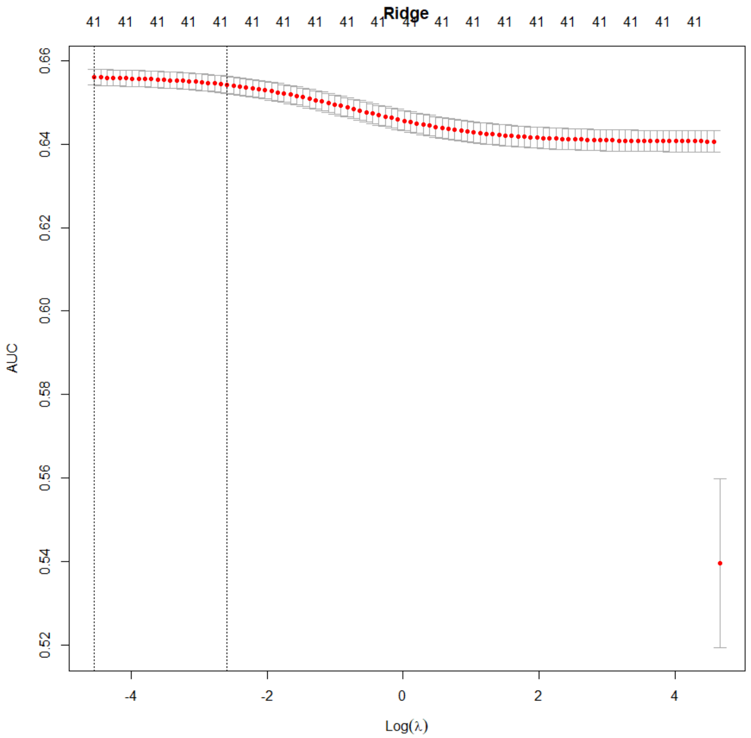
features include; age, race, gender, time in hospital, medical information, results of different diagnostic tests, medications, and more.

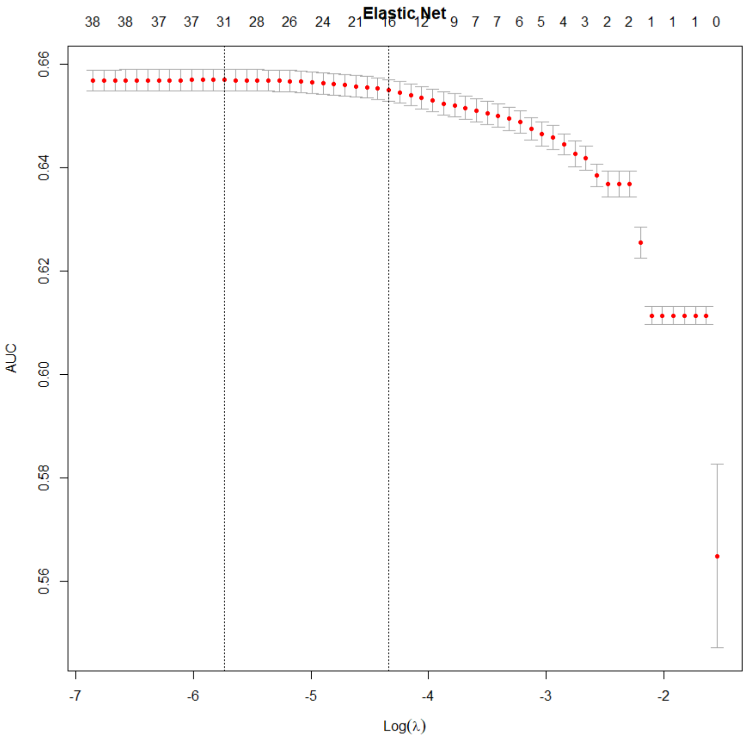
AUC Boxplots

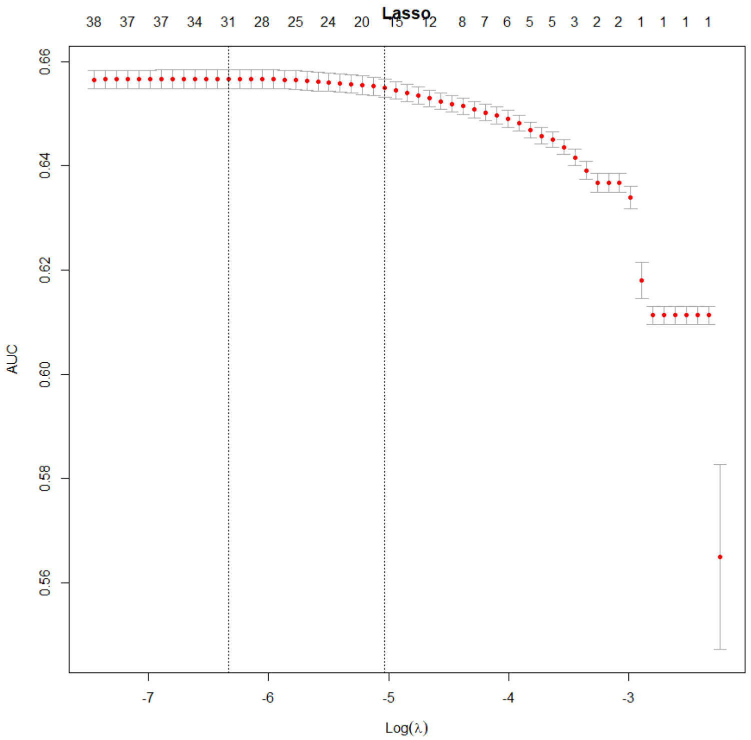




10-fold CV curves and Time







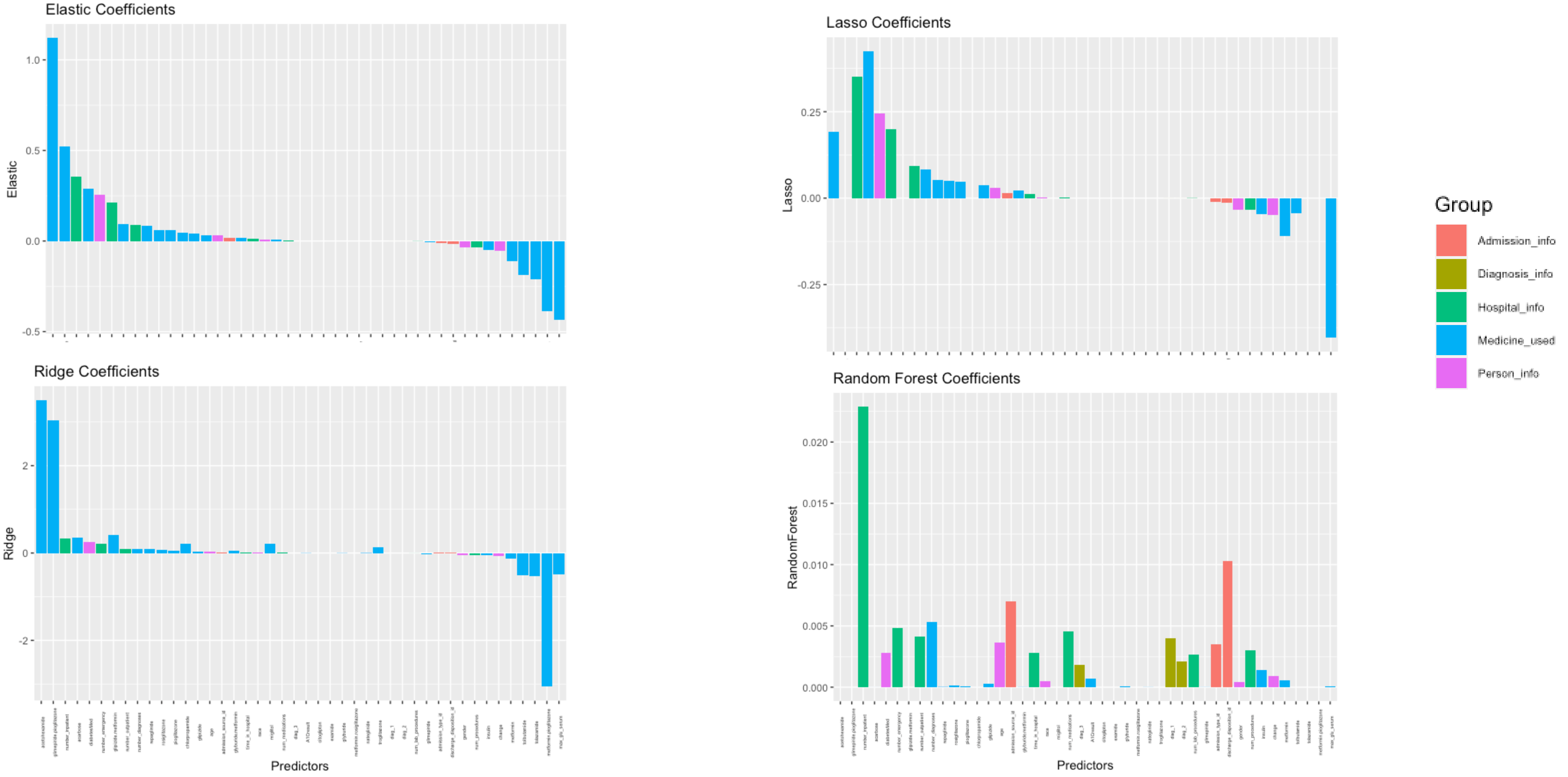
Ridge	Elastic Net	Lasso
83.67 seconds	38.19 seconds	37.55 seconds

Time to complete regressions

Method	Test AUC	Time (seconds)
Ridge	0.654728	86.52
Lasso	0.6552557	38.67
Elastic Net	0.6552481	39.97
Random Forest	0.6911152	575.43

- Trade off between time and performance?
- Lasso and Elastic Net both performed significantly better than Ridge on this data set.
- Random Forest did have the best Test AUC at the cost of time.

Importance of Parameters



Concluding Remarks

Random Forest performed the best, but at the cost of higher runtime

Our models generally have an AUC of around .65-.70

Elastic-net and Lasso generally keep most of the predictors

Number of inpatient and emergency visits in the last year, and diabetes medicine used generally had the most importance on readmittance chance