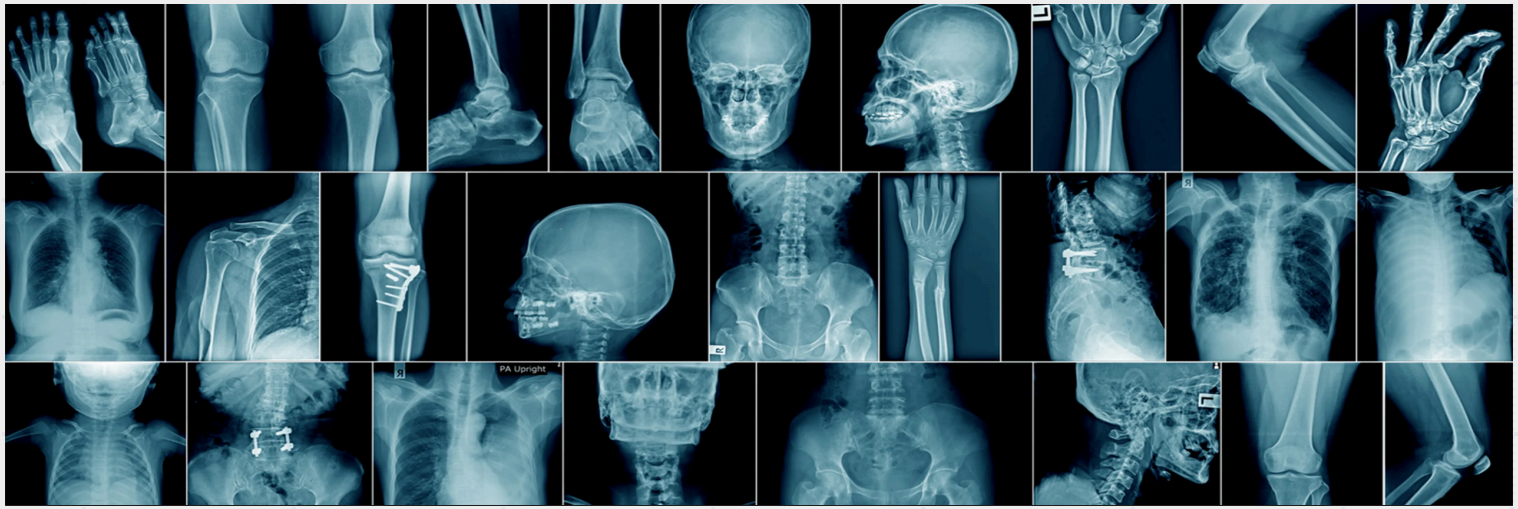
Chest radiograph classification



My Vien

Introduction

→ X ray images are useful with diagnosis of diseases.

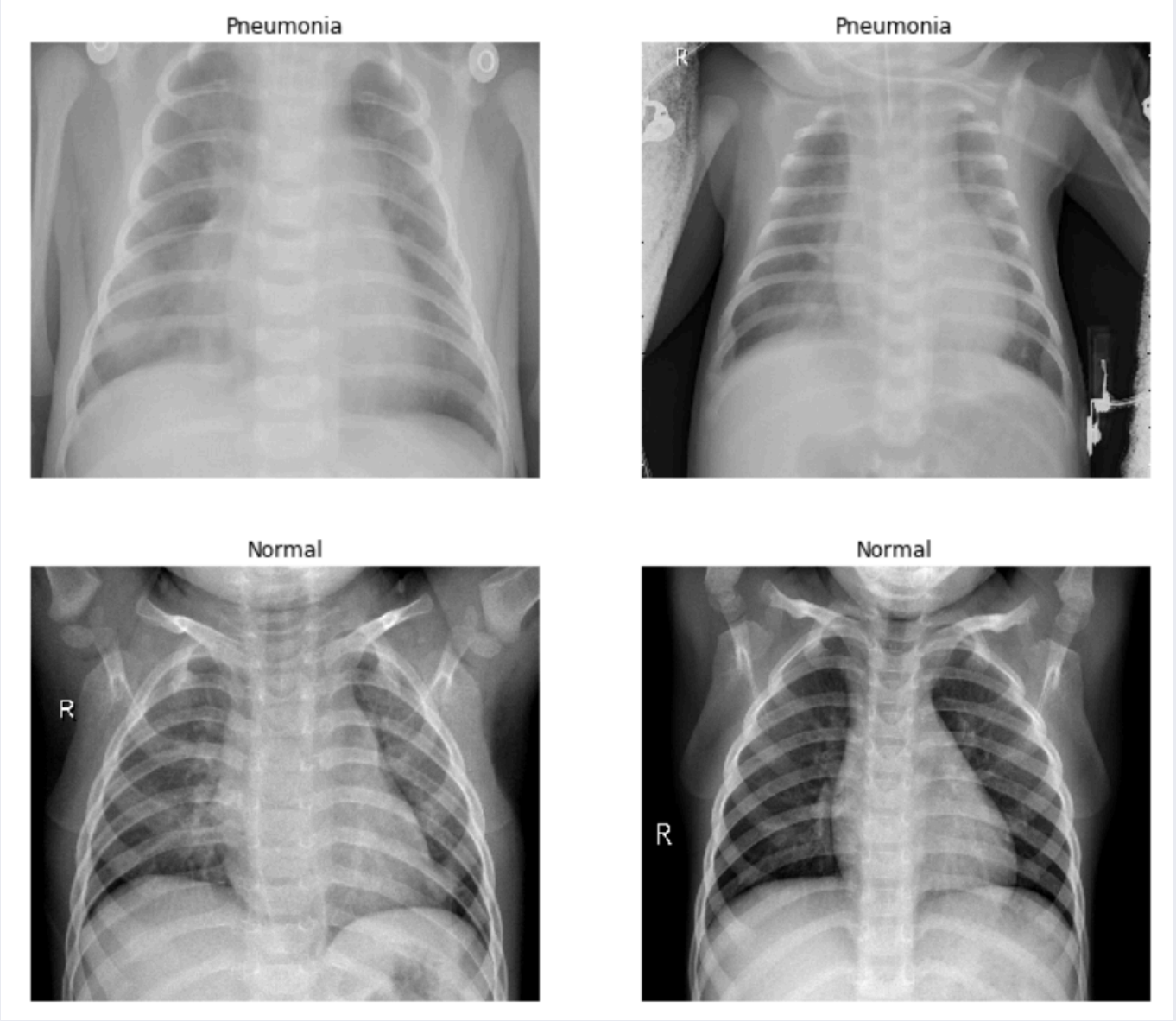
→ The dataset is from Kaggle

→ The target variable is binary class, it is pneumonia vs. normal.

→ p = 784. The predictors variables are 28 x 28-pixel images

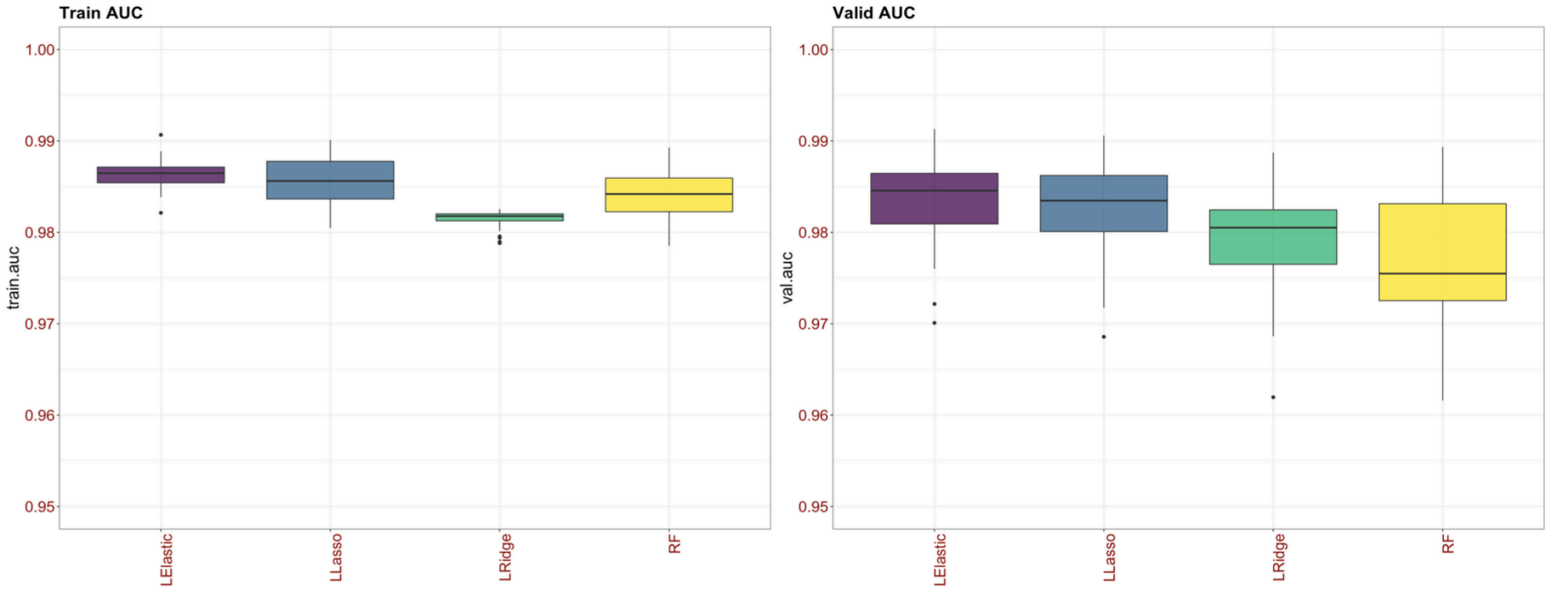
→ n = 5231, mixed of pneumonia and normal cases. Pneumonia cases are 3883 (0.74) and normal cases are 1348 (0.26)

→ We will be able to predict if a chest x ray image of patient is normal or infected with pneumonia.



AUC

- → Elastic Net has the best train and valid AUC score.
- → Ridge is the lowest on train set and RF is lowest on valid set.
- → Valid AUC scores are lower and more spread on all models
- → Logistic models have some outliers on train and valid sets



CV-Curve

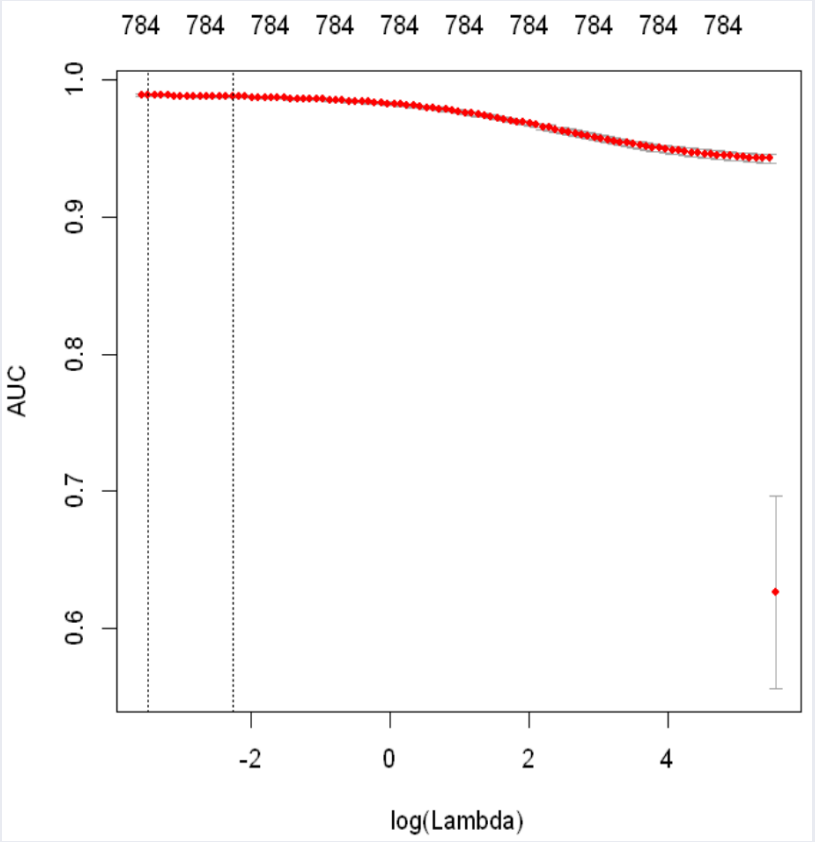
→ Average CV time for LRidge, Llasso, LElasticNet are **57s, 120s** and **98s**, respectively.

→ Optimized LLasso has around **121** non-zero coefficients

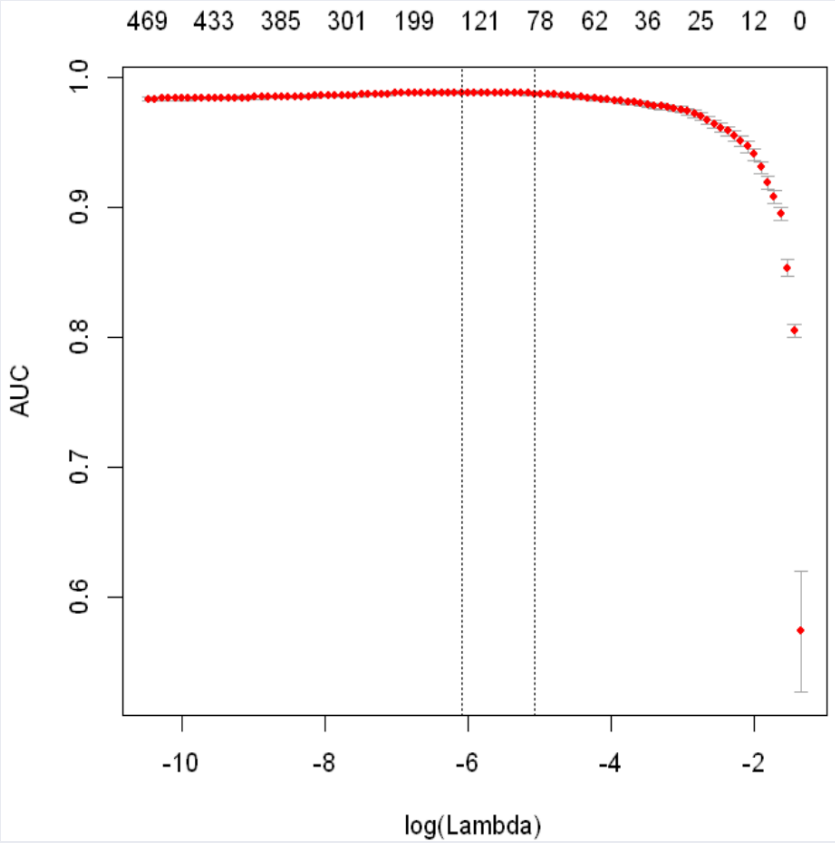
→ Optimized LElastic Net has around **184** non-zero coefficients

→ As expected, Optimized LRidge has maintained **all** predictors coefficients

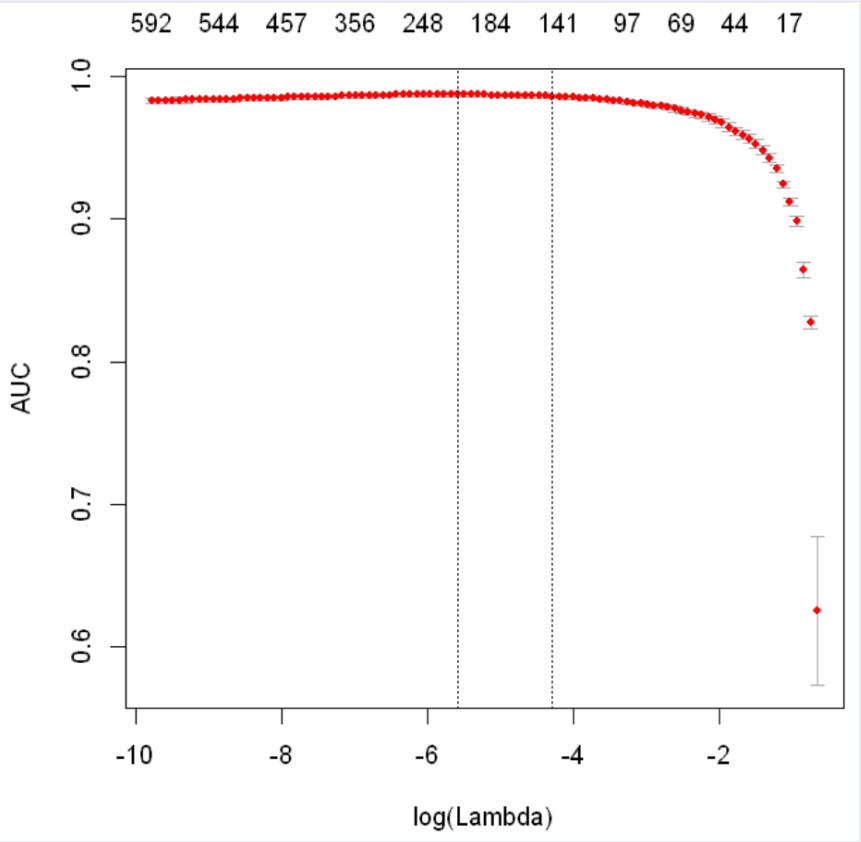
LRidge



LLasso



LElastic



Performance

Model	Valid AUC (Median)	Time (s)
LRIDGE	0.9805	**60.31**
LLASSO	0.9835	147.52
LELASTICNET	**0.9846**	114.37
RANDOMFOREST	0.9755	268.5

→ There is no trade off between time and model performance comparing between 3 constrained logistic models and RF.

→ Compared to LElastic, LLasso takes longer time to run but does not perform much better than LElastic which has the best valid score.

→ LRidge is fastest, but its score is less comparable.

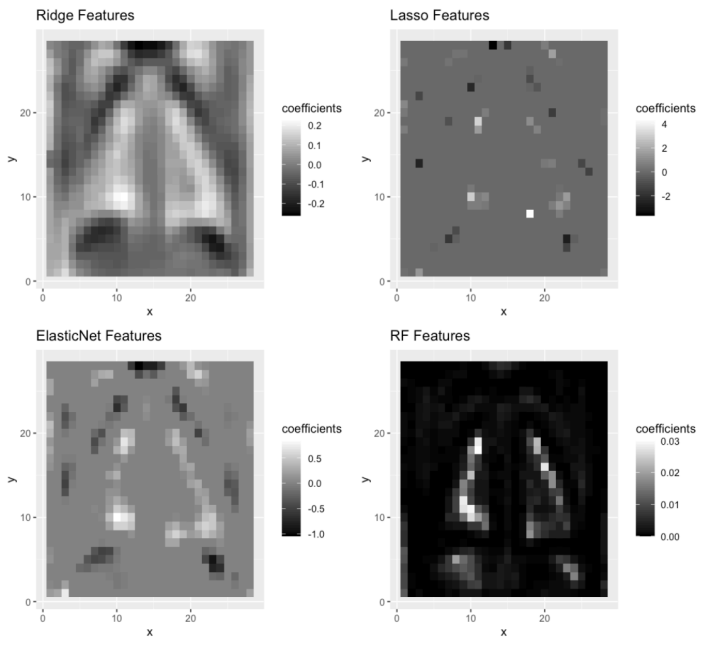
→ RF is slowest with lowest AUC score.

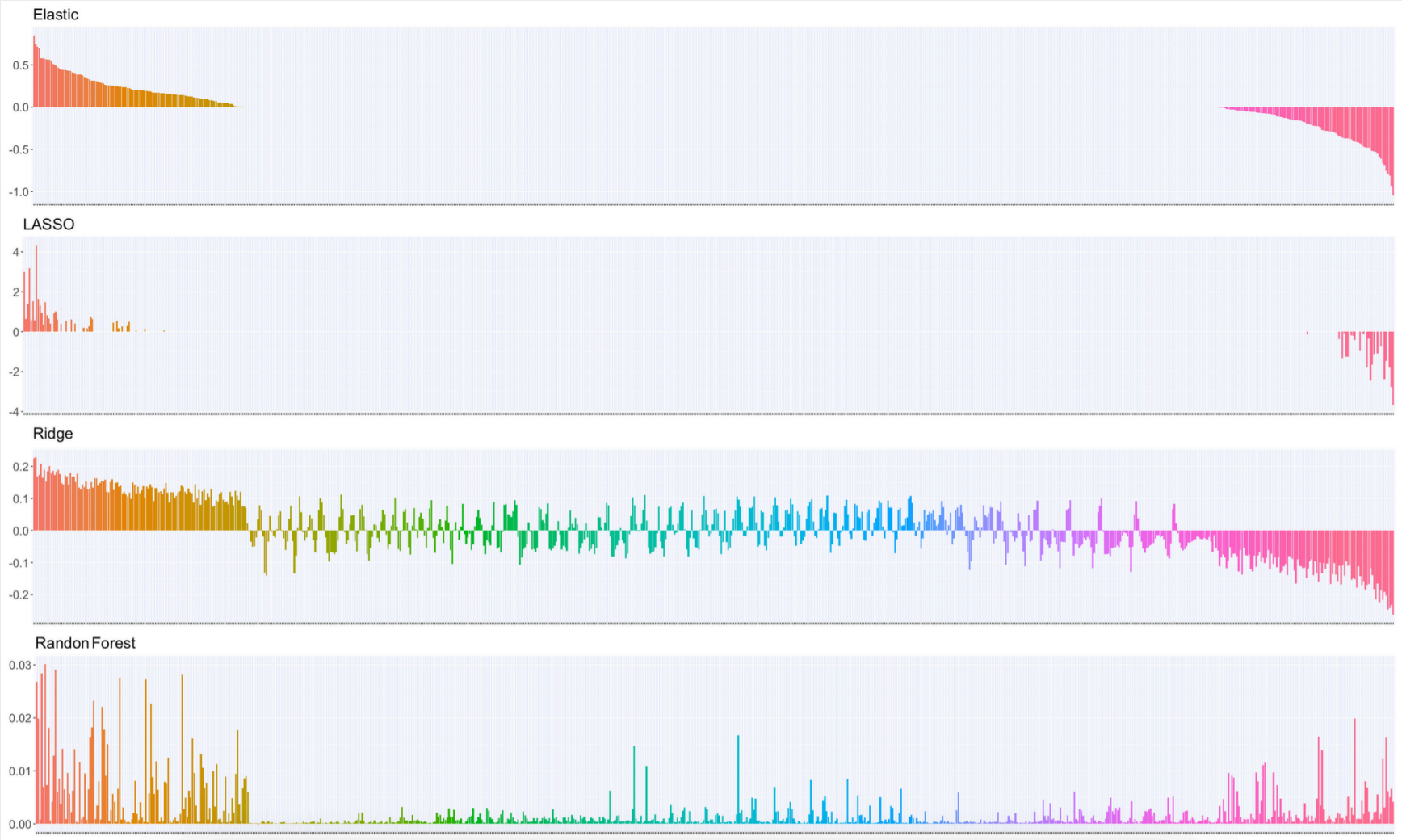
Coefficients

→ Like LRidge, RF have maintained almost predictors.

→ LLasso model has biggest, and RF has smallest coefficient values

→ LLasso and LElastic Net capture features of the area surrounding between the edge and the central of the images.





Conclusion

→ There is no big differences between train and valid AUC scores which proves that all our models are not overfitted.

→ RF takes longest time to run but has the lowest AUC score.

→ Elastic Net is the best method with highest AUC score.

→ The features importance located around between central with the edges of images which is also where our lungs should be on the images