Epileptic Seizure Recognition

STA 9891, Machine Learning
Fall 2020

Patrick Parham
Tanay Mukherjee

Dataset Overview

Description:

Epilepsy is a serious brain illness that is an endemic neurological disorder all over the world. It is a clinical result that occurs with abnormal neurological electrical discharging of the brain. Epileptic seizures represent the most common positive signs and symptoms of brain disturbance, and epilepsy is one of the most common primary brain disorders.

For diagnosing epilepsy, research is needed for a better understanding of mechanisms causing epileptic disorders. The evaluation and treatment of neurophysiologic disorders are diagnosed with the EEG results. EEG is crucial for the accurate classification of different forms of epilepsy.

Data Dictionary

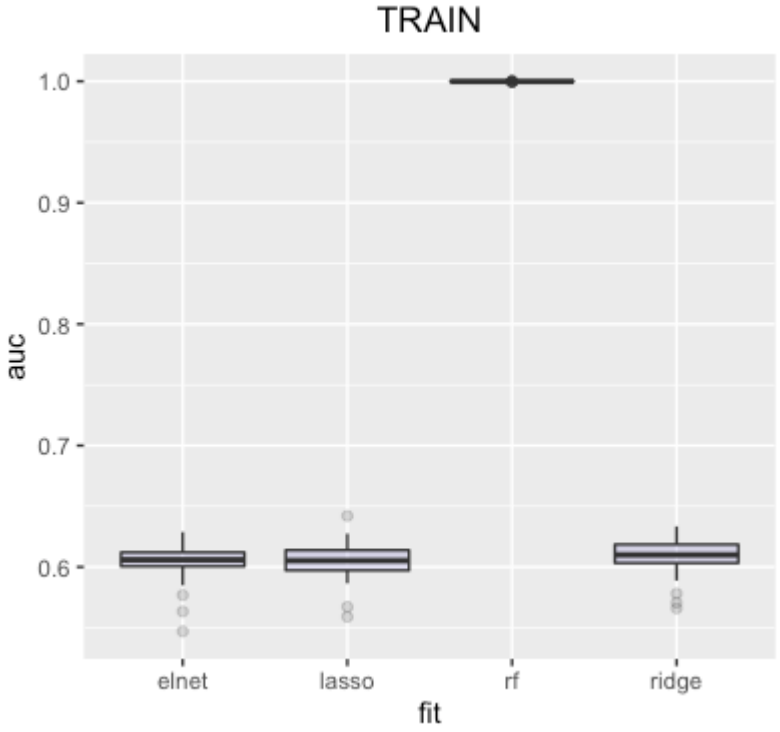
About the dataset:

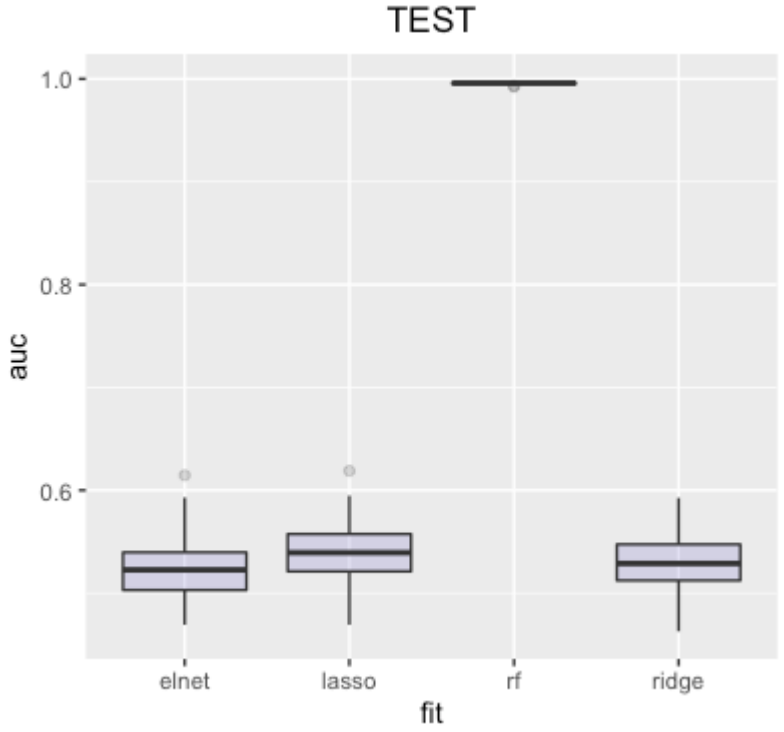
There are 178 features of which the Explanatory variables are X1, X2, ..., X178 (each feature represents the signals received by brain and its proximity to the point where the seizure was first recorded), with one response variable as y.
In this exercise, the aim of this study is to detect epileptic seizures using different feature extraction methods and comparison of the performance from various ML methods used for classification. Our aim will be to classify the brain activity and recognize whether there is a seizure identified or not.

Summary of the dataset:
p = 179
n = 11500
n+ = 2300
n- = 9200

The dataset and more information about how the data collected can be found here:
https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

AUC results comparison for train and test set





Model accuracy and run time

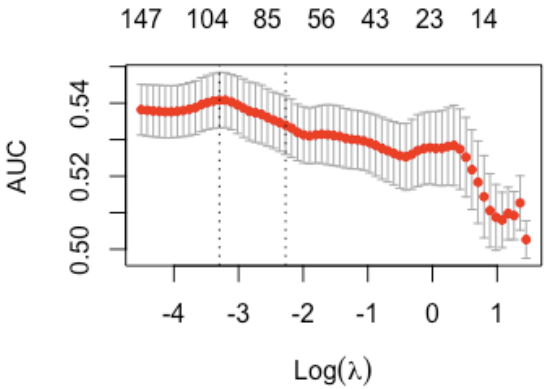
Fit	AUC*	Time**
Random Forest	0.992 - 0.997	9.52 mins (AUC: 0.99)
Elastic Net	0.48 – 0.57	2.71 mins (AUC: 0.60)
Lasso	0.50 – 0.58	2.57 mins (AUC: 0.59)
Ridge	0.49 – 0.58	2.99 mins (AUC: 0.62)

* 90% test AUC interval based on the 50 samples with 90% confidence interval

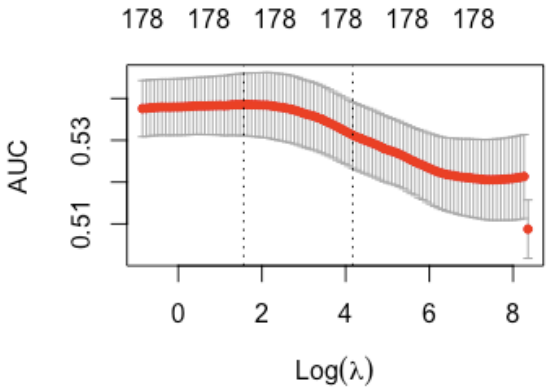
** The time it takes to fit the model on all the data

10-Fold CV Curves for each model

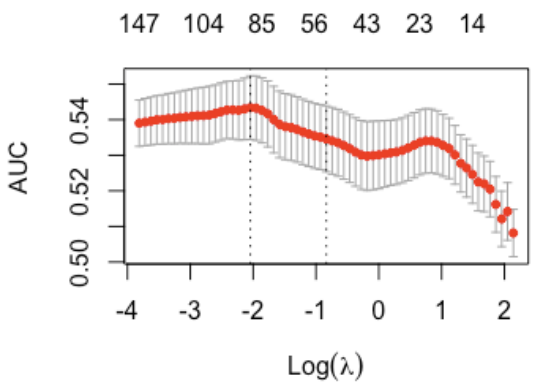
LASSO



RIDGE

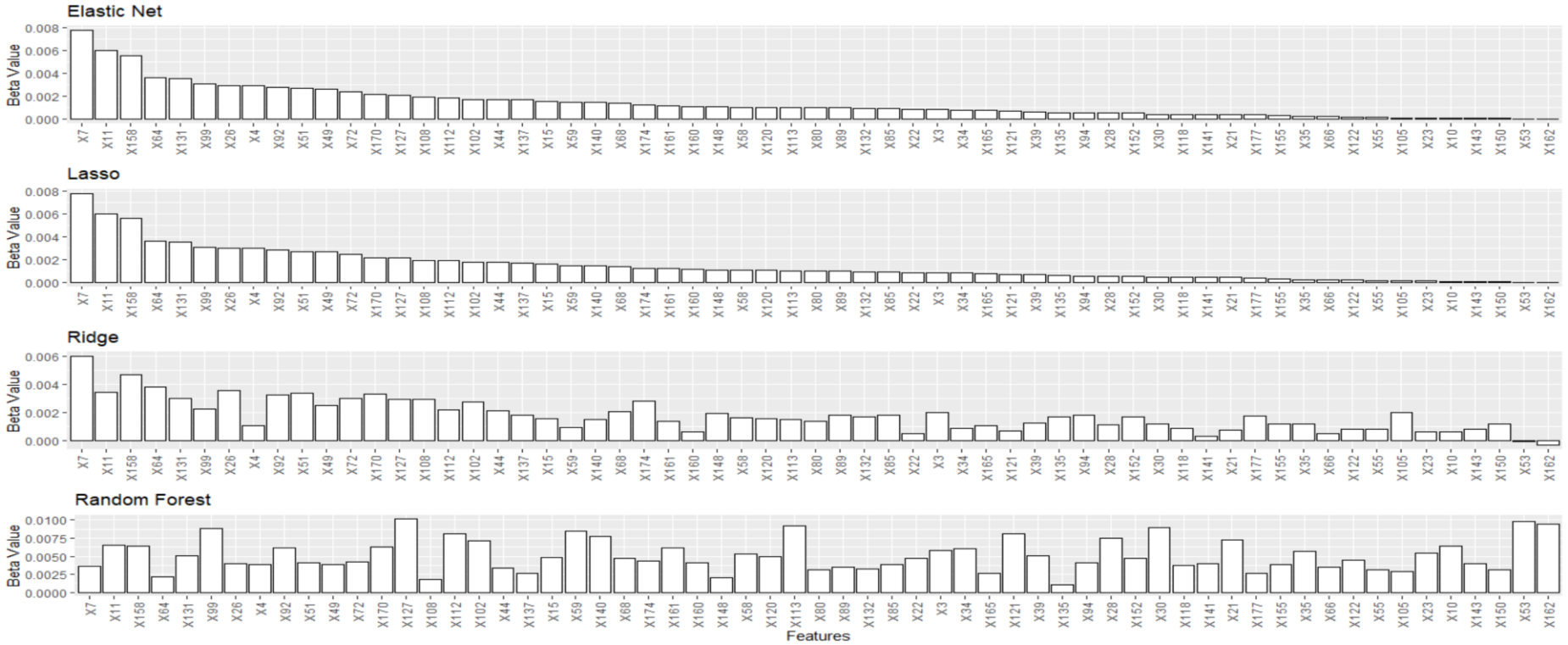


ELASTIC NET



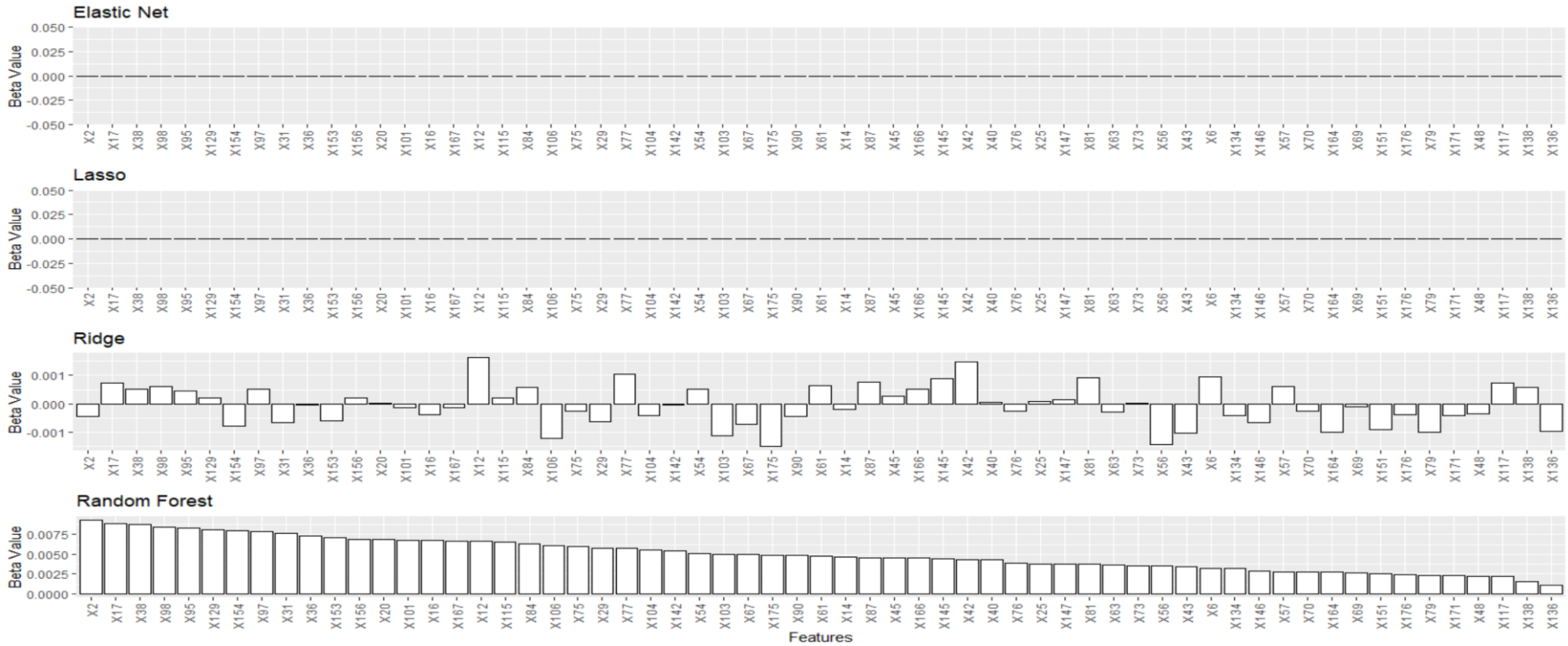
Time by Fit and Size	Lasso	Ridge	Elastic-Net
Individual Sample	2.35 mins	2.26 mins	2.33mins
50 Samples with Thresholds	3.25 hours	2.25 hours	2.10 hours

Estimated Coefficients – I



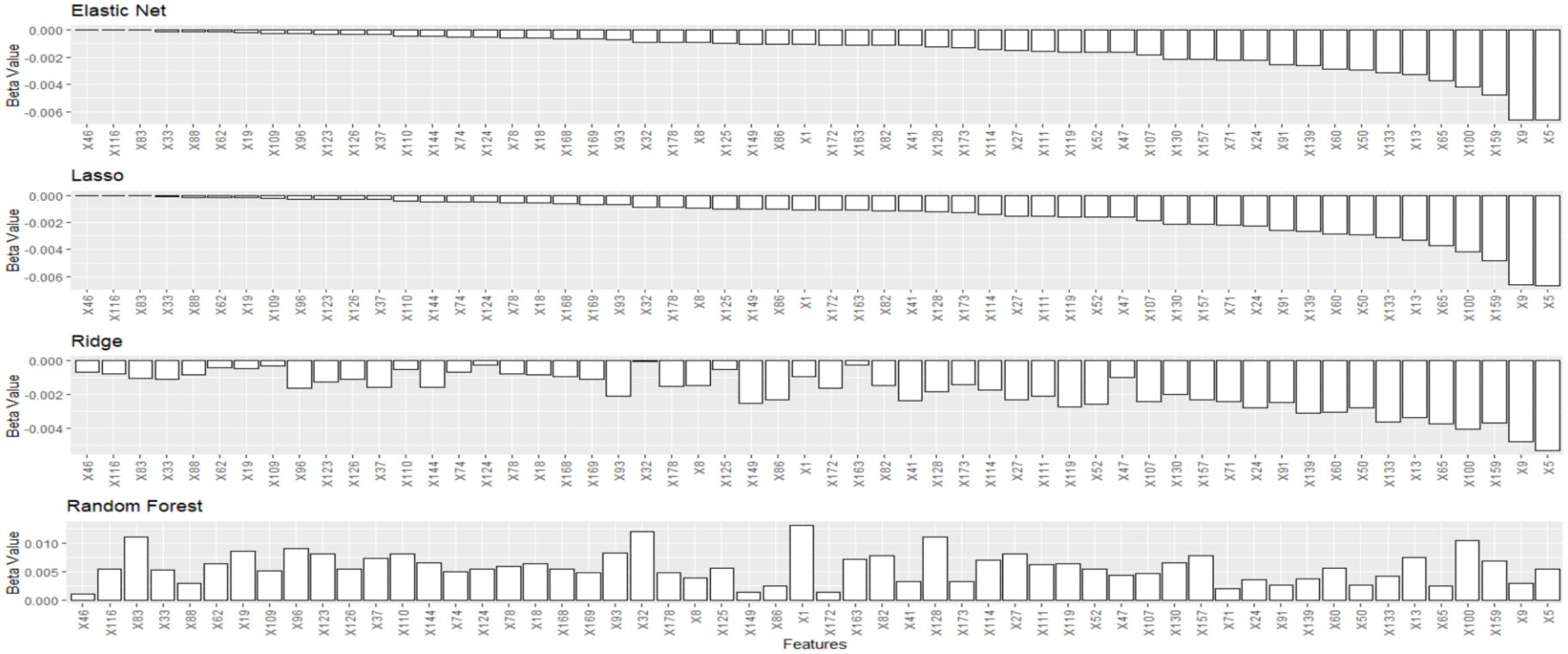
Top 60 features in descending order of beta values from Elastic Net

Estimated Coefficients – II



Middle 60 features (61 - 120) in descending order of beta values from Elastic Net

Estimated Coefficients – III



Bottom 58 features (121-178) in descending order of beta values from Elastic Net

Conclusion

1. We observed the AUC calculation for each of the 4 models. The results for lasso, ridge and elastic-net logistic regression were similar. However, the results for Random Forest was in the range of 99.2% to 99.7%.
    a. *The median value of AUC for Lasso, Ridge and Elastic net for 90% of the AUC interval lied between 45% to 62%.*
    b. *Random Forest is clearly the best model in classifying the seizure occurrences.*

2. The order in terms of time it took for each model to run is as follows:
    a. *Random Forest > Ridge > Elastic-net > Lasso*

3. For 10-fold cross validation, each model took close to two and a half minute to produce result for randomized samples.
    a. *Elastic-net uses 88 features, where as Lasso uses 99 features. While Ridge as we would expect uses all 178 features for tuning.*
    b. *When the same experiment was run for 50 such samples, the order of the time it took to run each model is as follows:*
    Lasso > Ridge > Elastic-net

4. We looked at the important features for each models and arranged it in the descending order of beta values from Random Forest.
    a. *Random Forest is not great for classification as decision trees overfit the model greatly. Therefore, some of the features which will appear important for RF might not appear significant for other models. However, feature X7, X11, X158, X92 appears to be important for all the models.*

THANK YOU!