MACHINE LEARNING METHODS CLASSIFICATION ON HEDGE FUND X: FINANCIAL MODELING CHALLENGE DATA

Lily (Lizheng) Zhou

INTRODUCTION

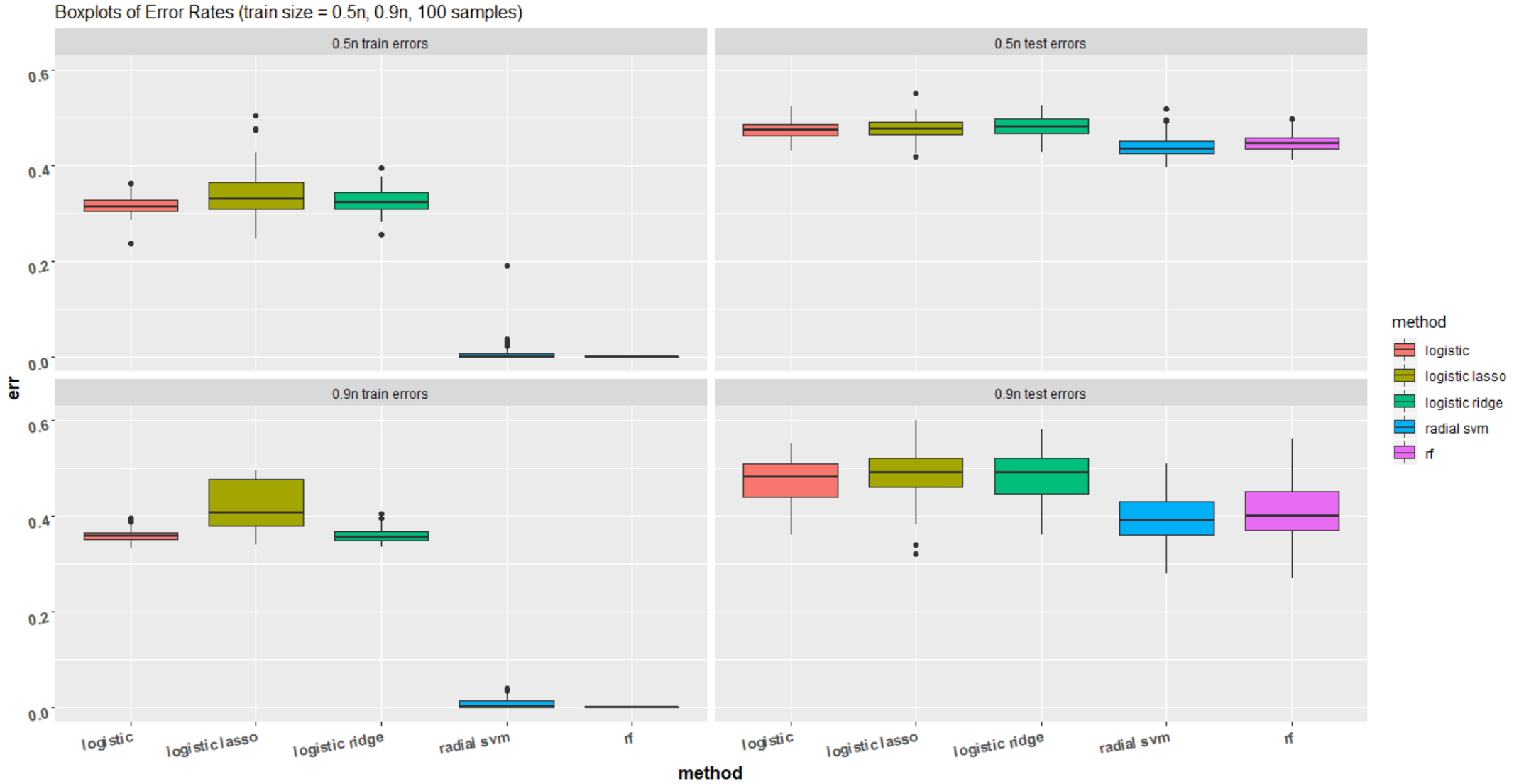
Project Description

- This project is aimed to classify and predict a binary financial target based on numeric variables with several methods: Random Forrest, Radial SVM, Logistics regression, Logistic Lasso Regression, Logistic Ridge Regression.

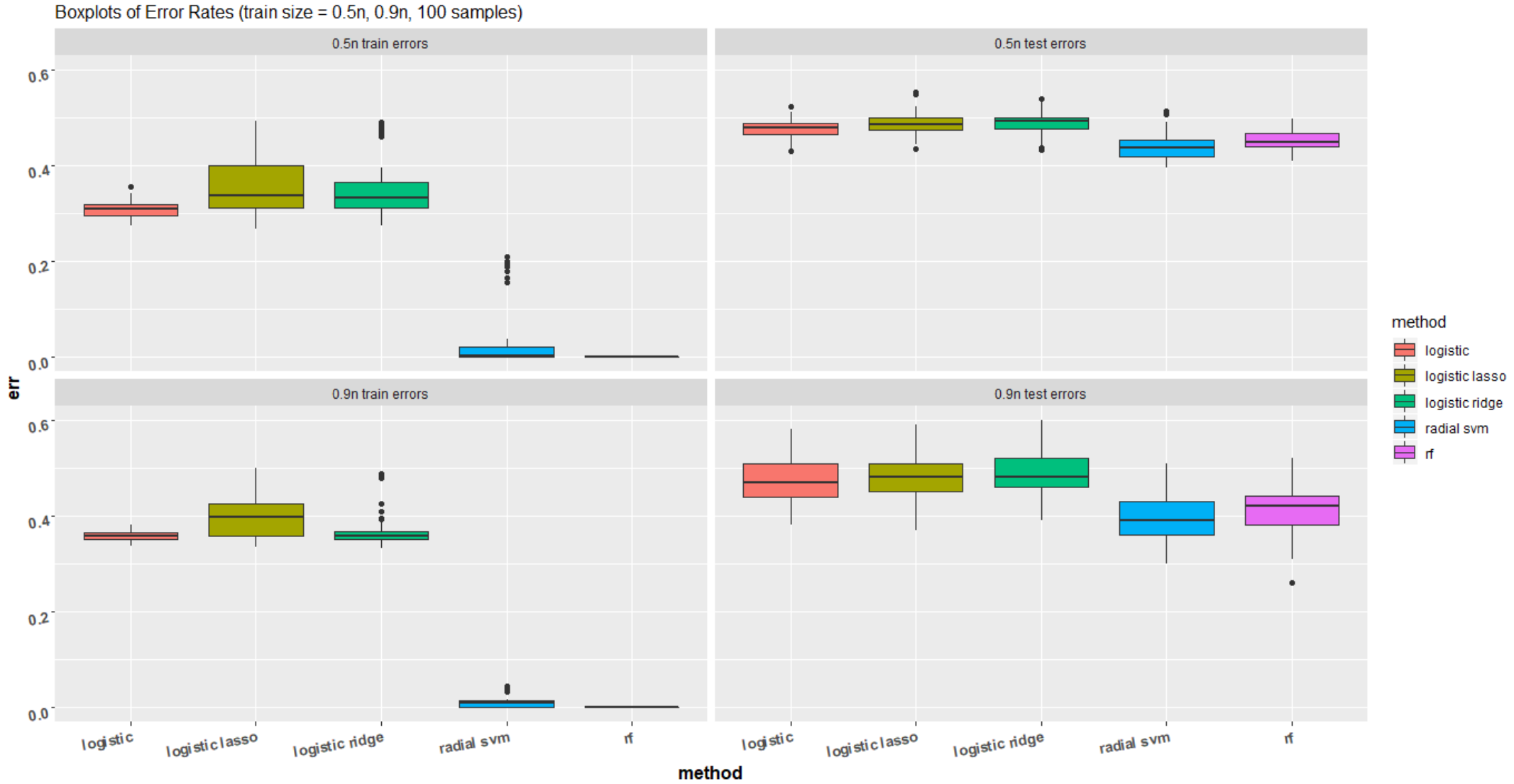
Data Description

- This dataset is a sample of the training dataset used in the DeepAnalytics competition, Hedge Fund X: Financial Modeling Challenge (https://deepanalytics.jp/compe/53).
- Data Set Structure: (n = 10000, p = 88)
  - data_id: ID
  - period: A preiod where observation data belong (train9)
  - c1 - c88: Explanatory variables (full numeric)
  - target: Binary class to predict (0 or 1)
    - Balanced (Number of 0: 5006, Number of 1: 4994)
- Subset of Data Set:
  - n = 1000, p = 88
  - **Still balanced** (Number of 0: 512, Number of 1: 488)

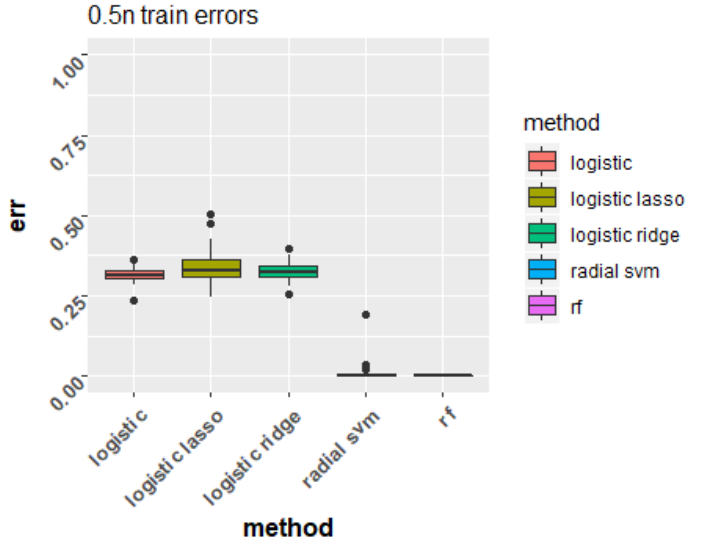
ERROR RATES (0.5N, 0.9N) – IMBALANCE=TRUE

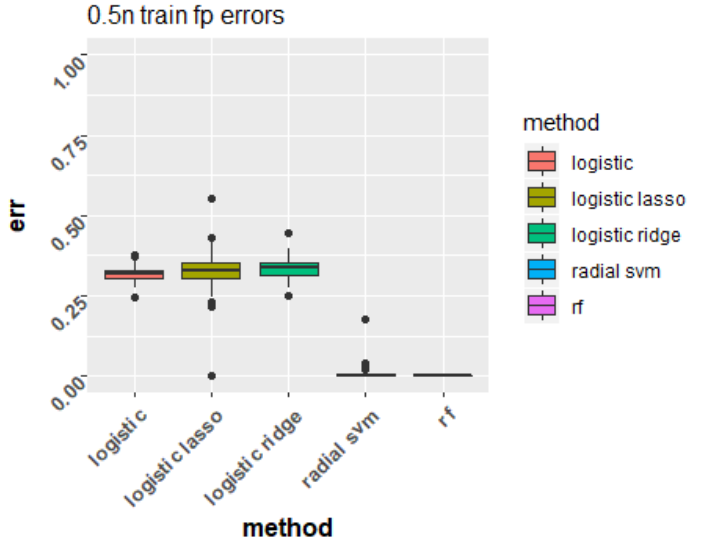


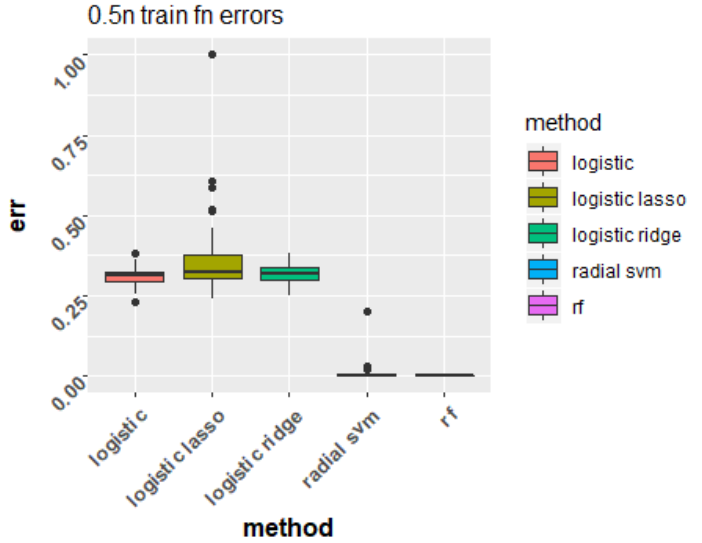
ERROR RATES (0.5N, 0.9N) – IMBALANCE=FALSE

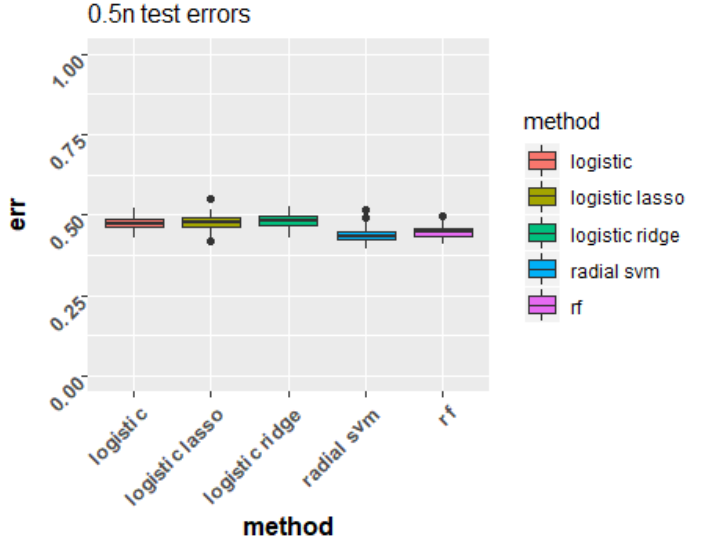


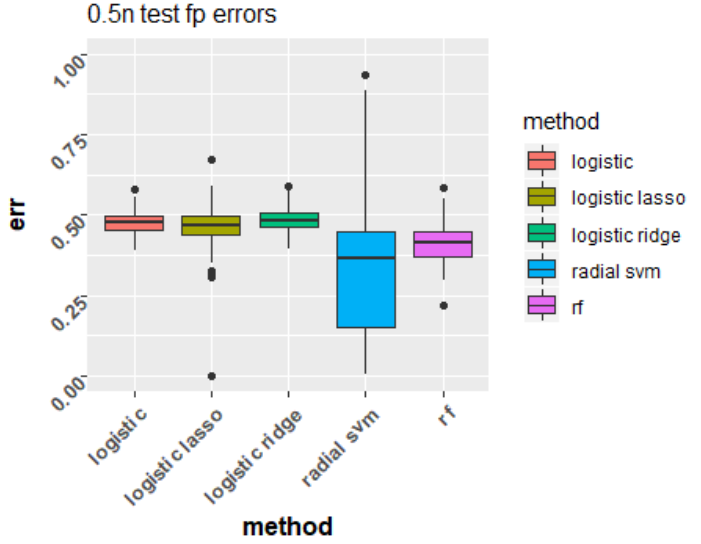
FP/FN PLOTS (0.5N) – IMBALANCE=TRUE

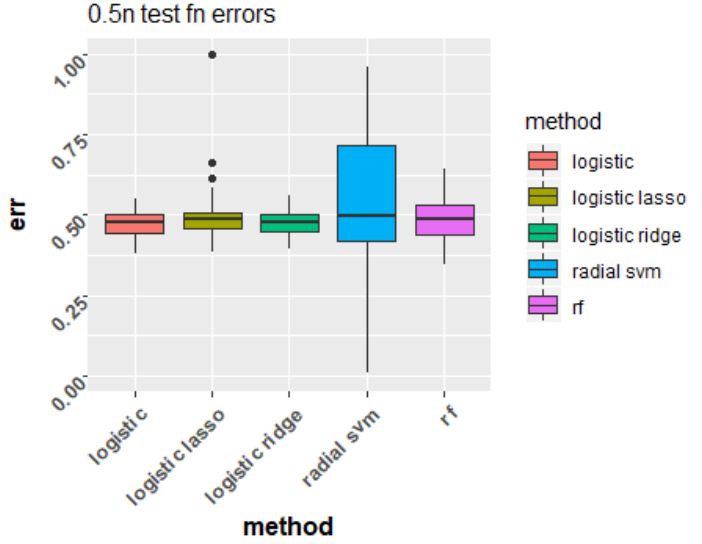




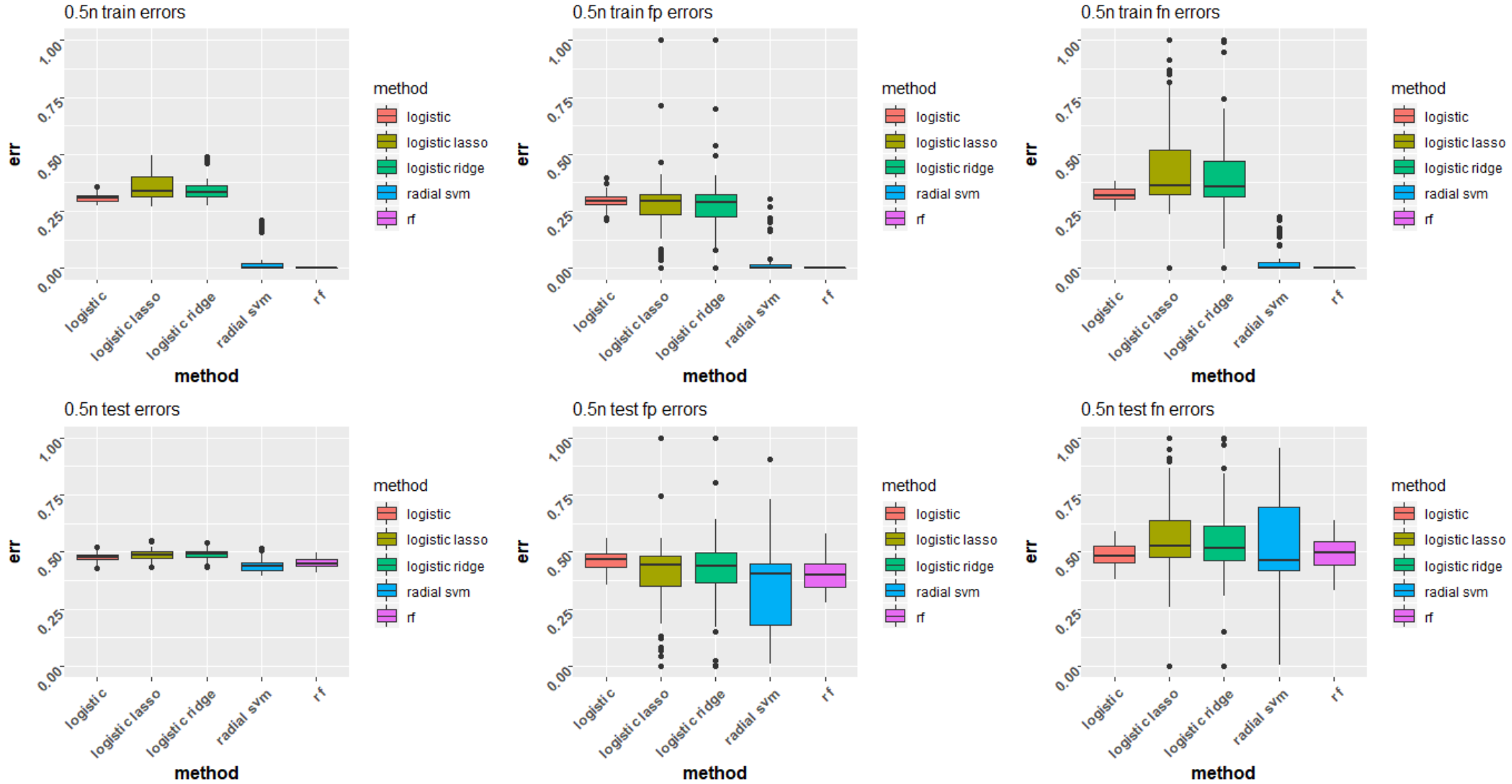




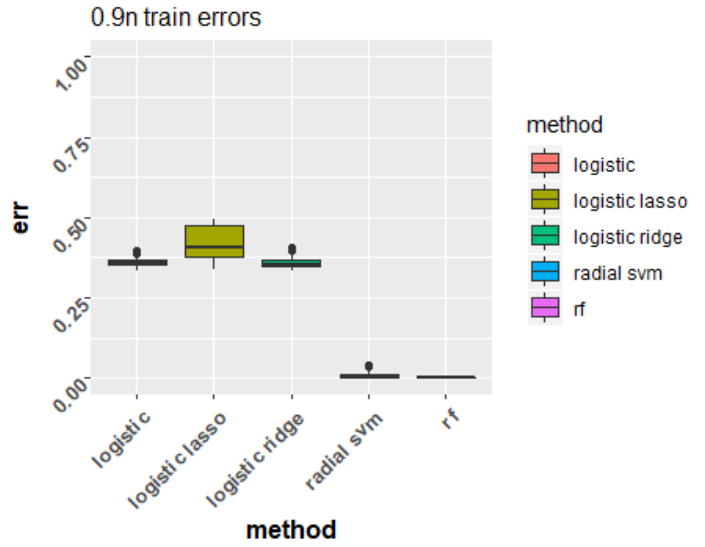


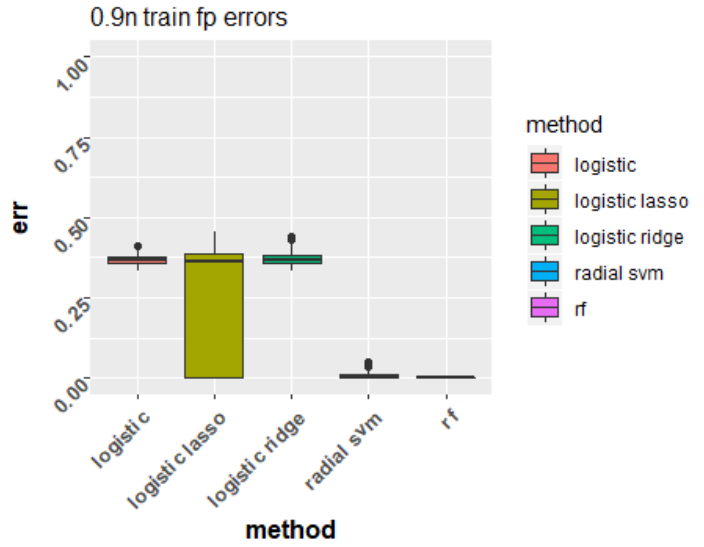


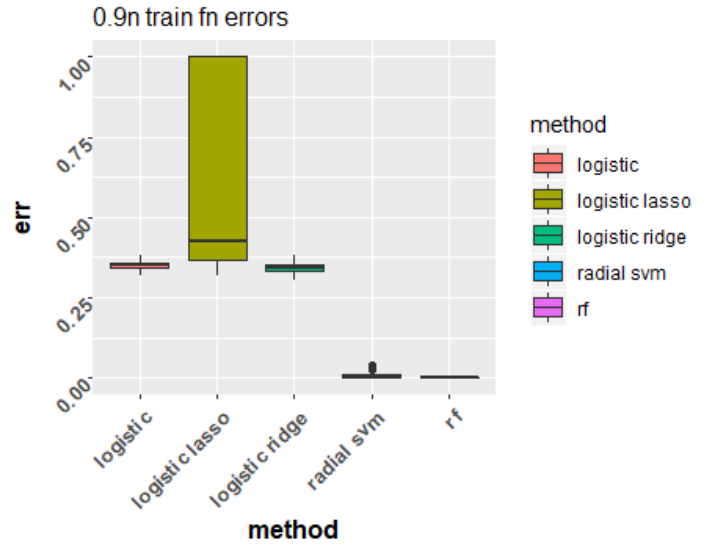
FP/FN PLOTS (0.5N) – IMBALANCE=FALSE

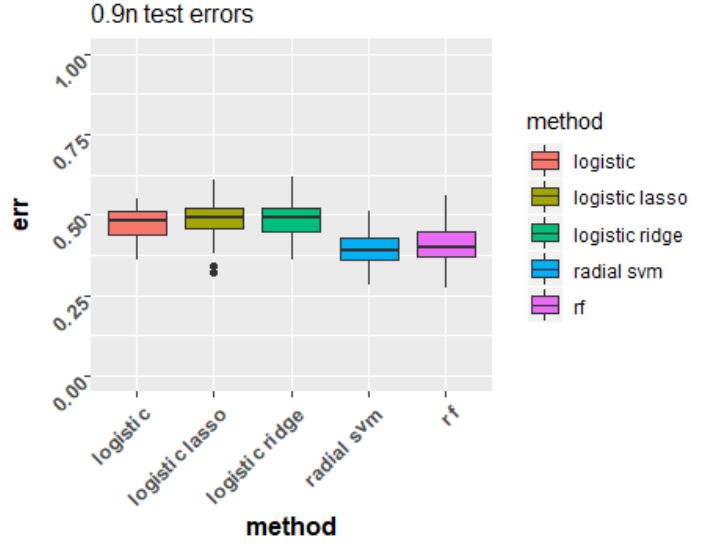


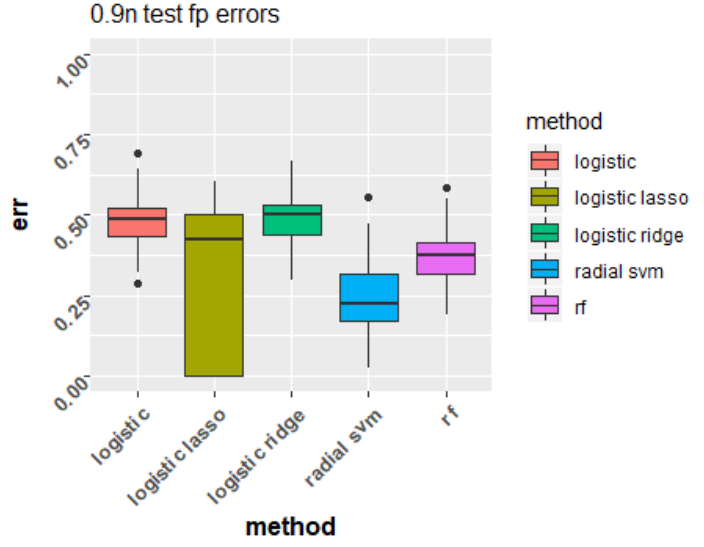
FP/FN PLOTS (0.9N) – IMBALANCE=TRUE

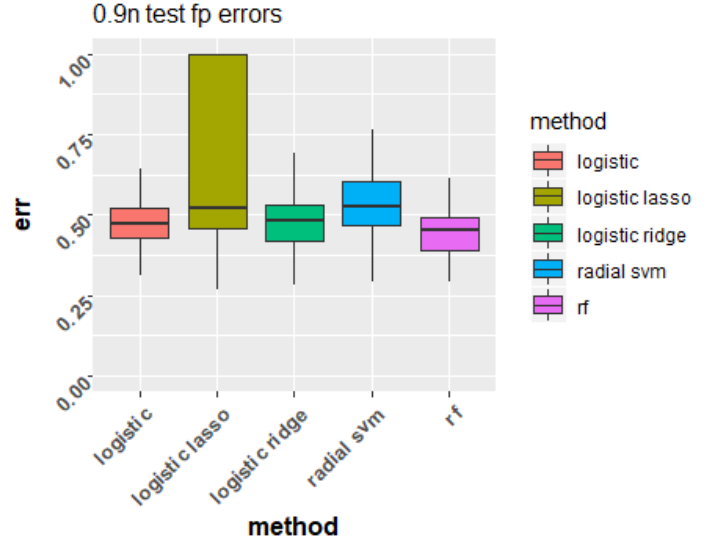




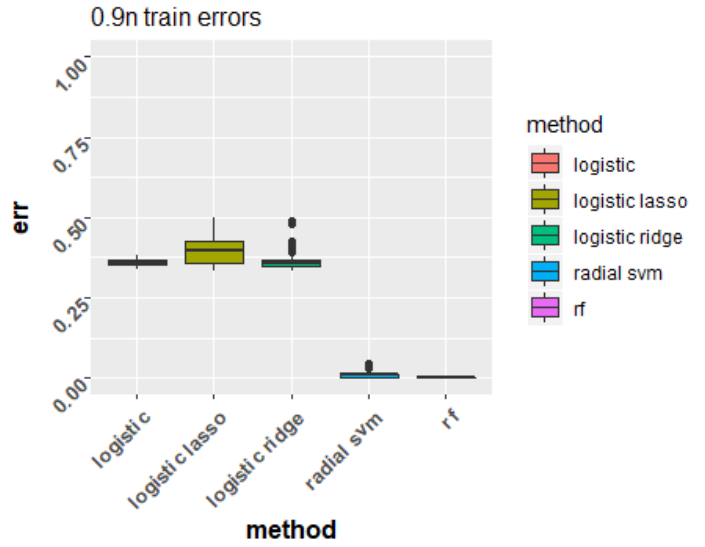


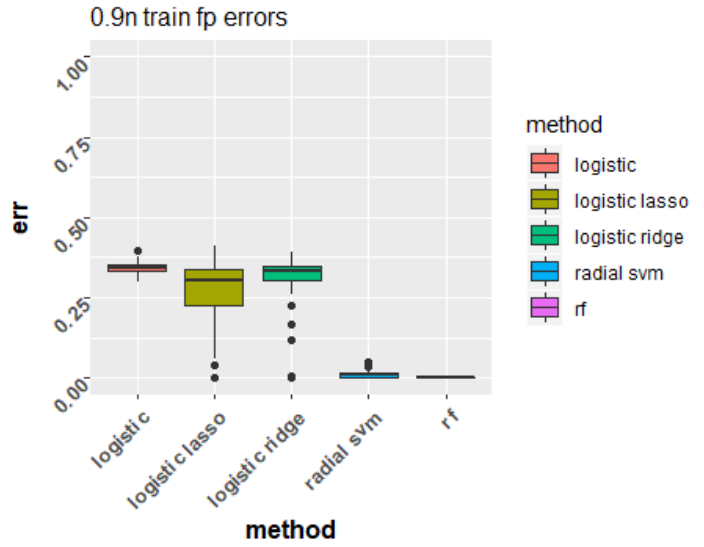


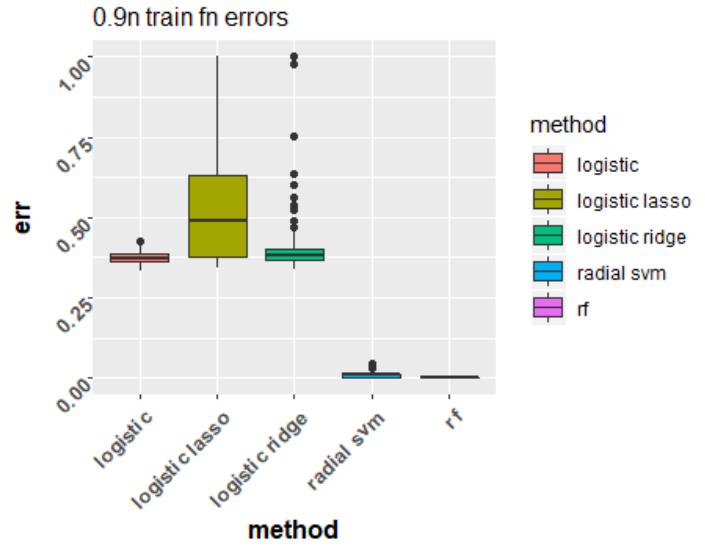


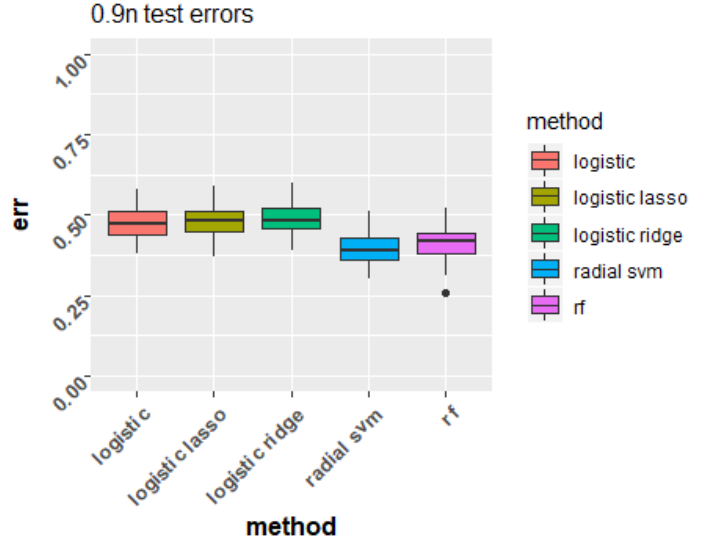


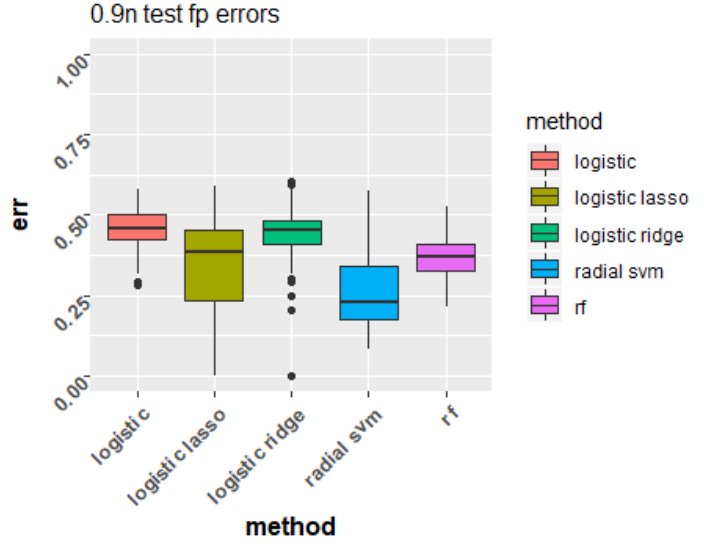
FP/FN PLOTS (0.9N) – IMBALANCE=FALSE

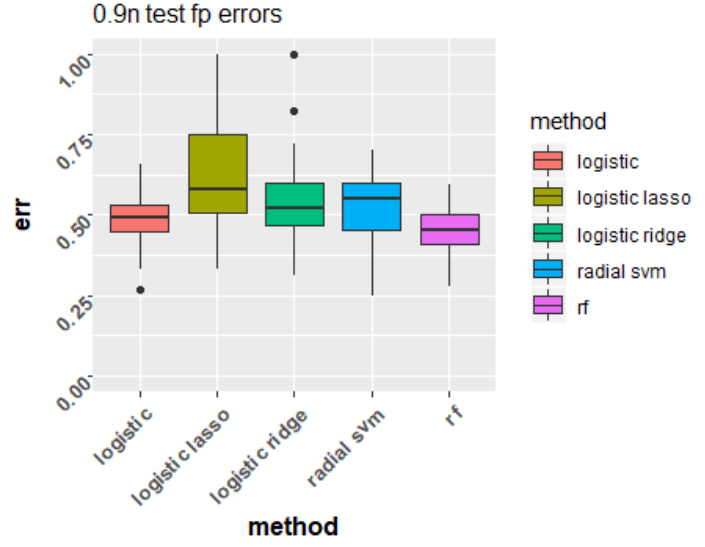




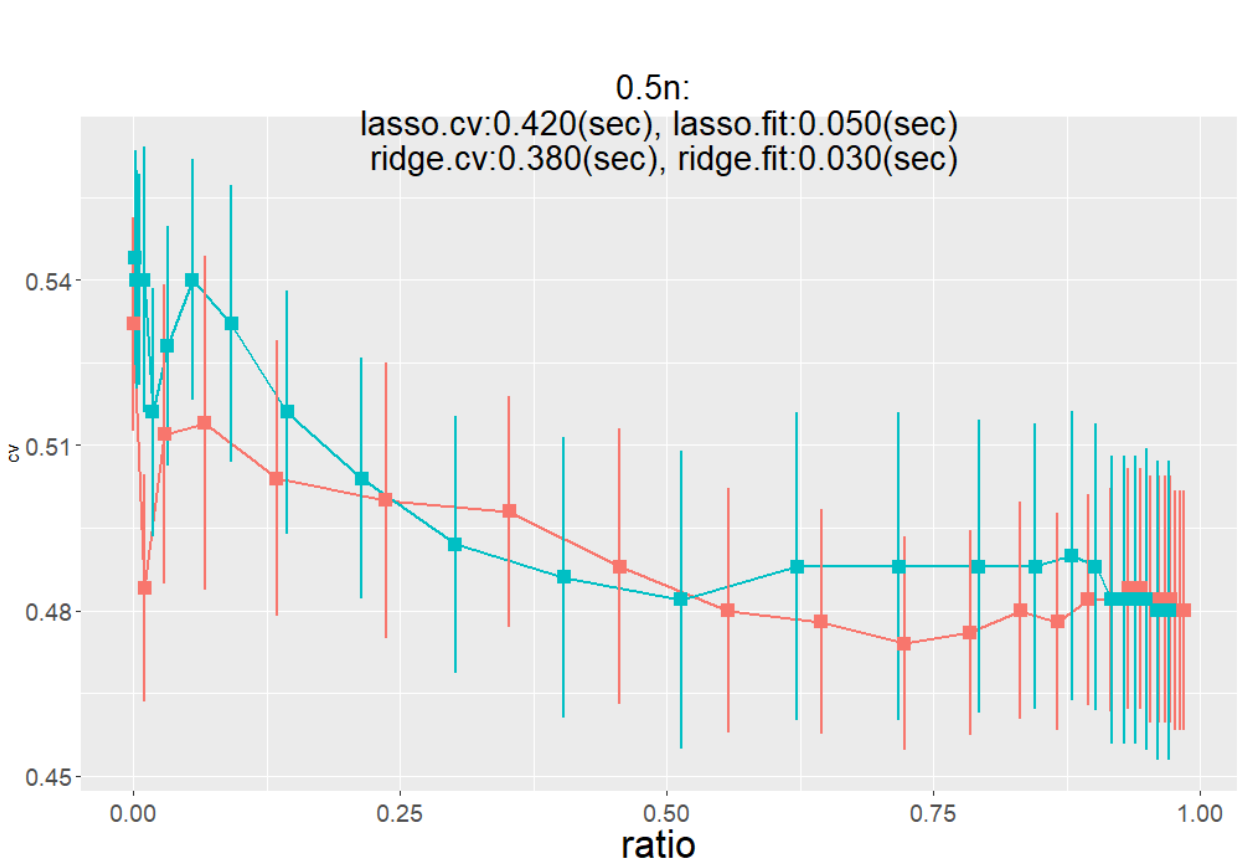


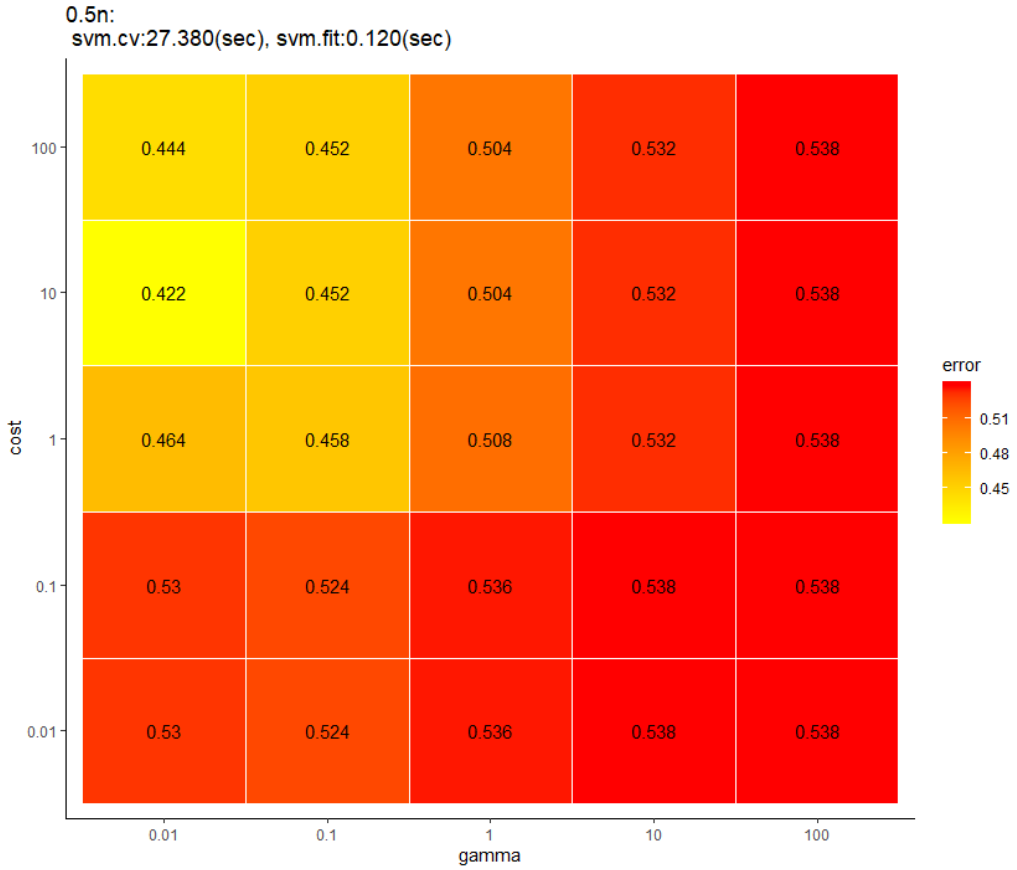




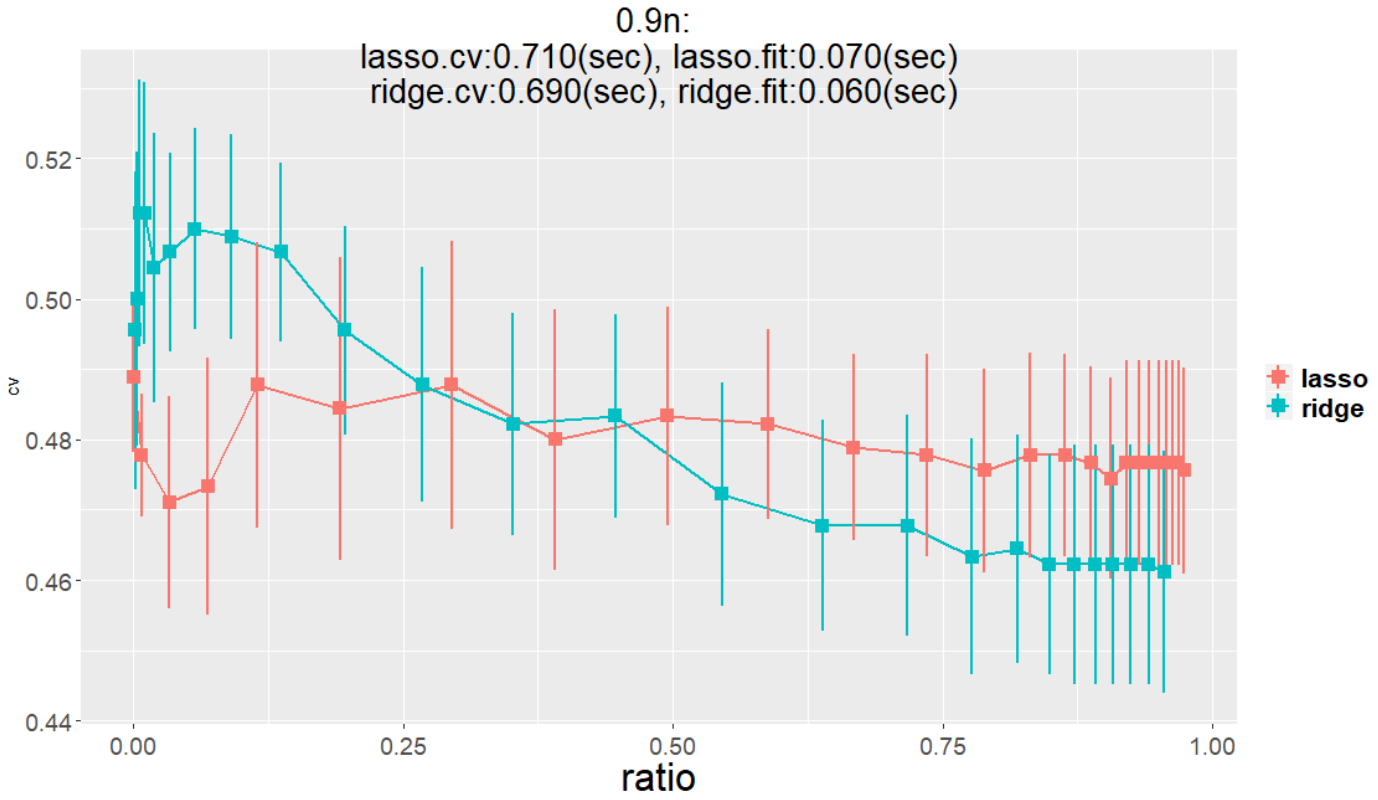


10-FOLD CV ERROR CURVES/HEATMAP (0.5N)



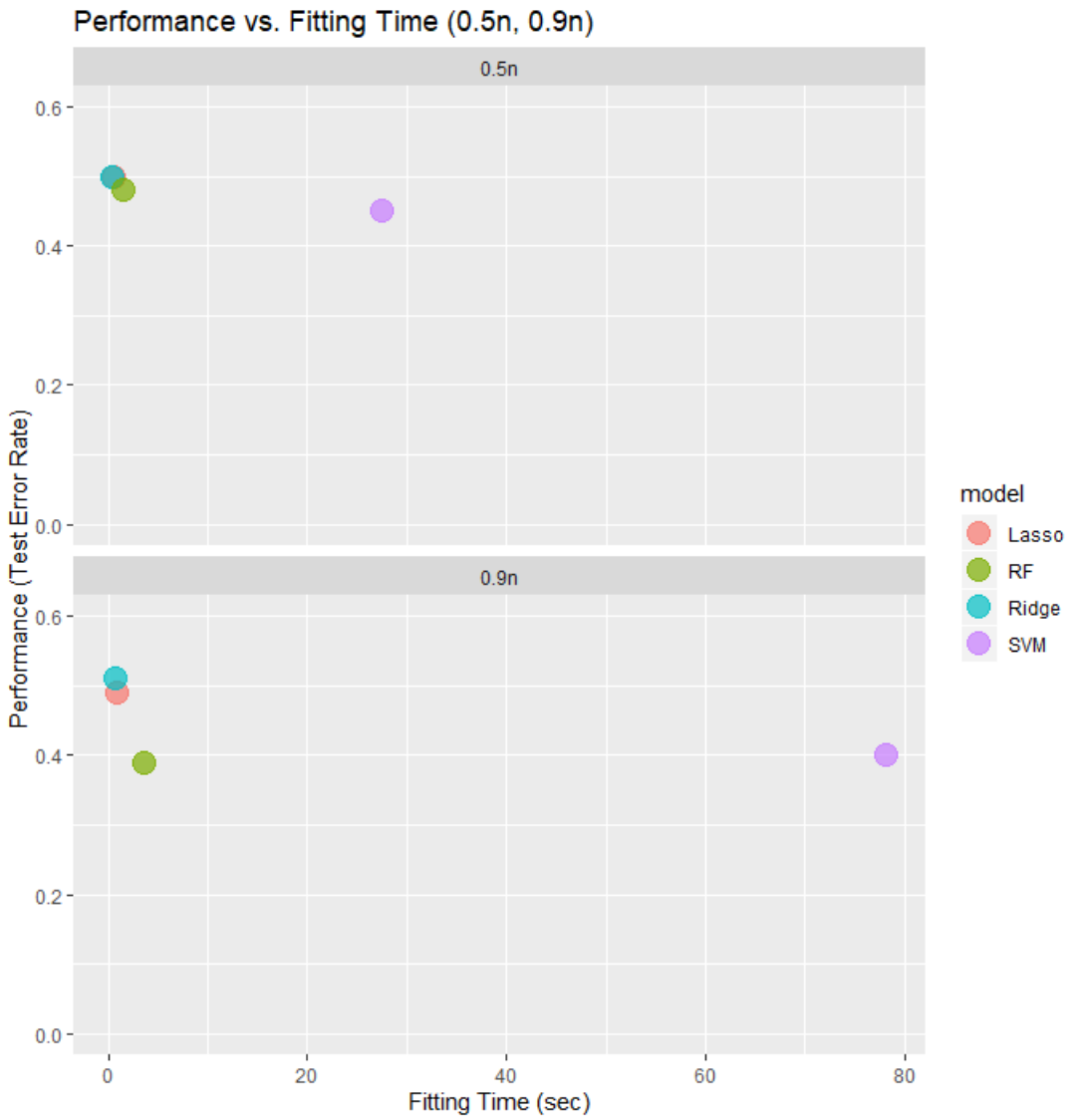


10-FOLD CV ERROR CURVES/HEATMAP (0.9N)

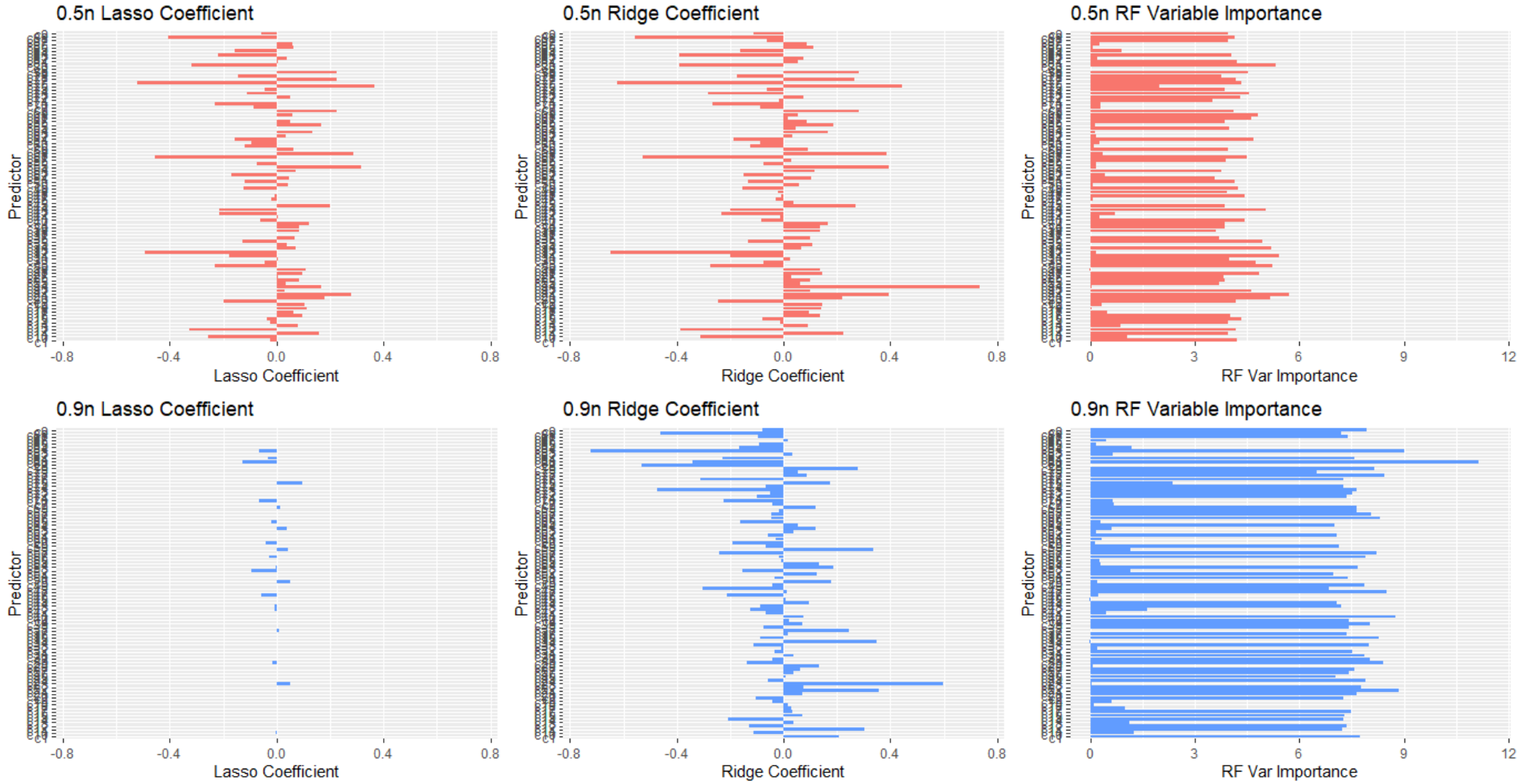


PERFORMANCE VS. TIME

Model	Model Fitting Time (sec)		Model Performance (Error rates)	
	0.5n	0.9n	0.5n	0.9n
Lasso	0.470	0.780	0.50	0.49
Ridge	0.410	0.750	0.50	0.51
RF	1.440	3.590	0.48	0.39
SVM	27.500	78.180	0.45	0.40

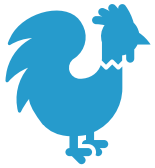


VARIABLE IMPORTANCE (FULL VARIABLES)



COMMENTS

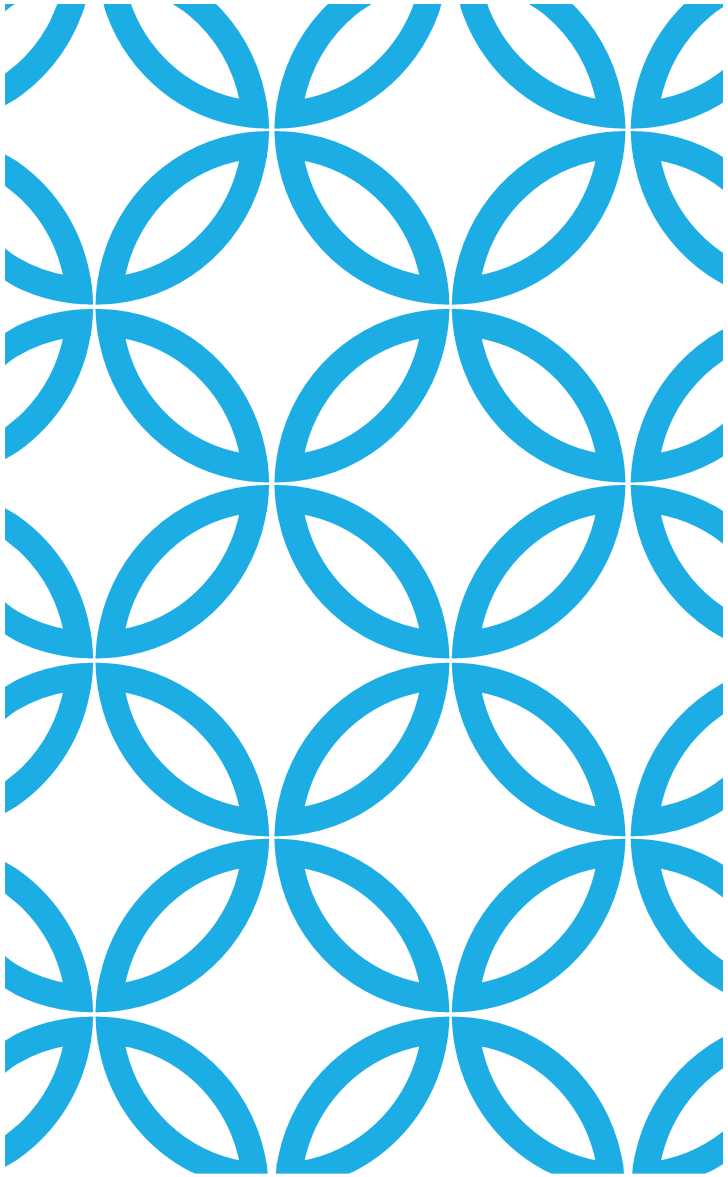
SVM performs decent at the cost of high time complexity



Random Forest performs decent before parameter tuning, may need to tune it

No major differences shown between different training data size

Lack of real meaning of target (y) and 88 predictors



THANK YOU