STA9690 Final Project

Advanced Data Mining

Christopher Lang

December 12, 2017

The Spambase Dataset

A dataset of emails and several key features that describe the content written in the email body

Number of Emails (n)	2695
Number of Predictors (p)	57 – 3 = **54** (3 predictors are all zero)
Response	1512 (~56%) spam 1183 (~44%) non-spam

Sourced Spambase from:

UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/spambase

Predictor Description

Predictors are broken into three types of measures

Word Frequency
% of total words in an email

```
word_freq_make
word_freq_internet
word_freq_hp
…
…
```

45 predictors (83%)

Character Frequency
% of total characters in an email

```
char_freq_semicolon
char_freq_exclamation
char_freq_dollarsign
…
…
```

6 predictors (11%)

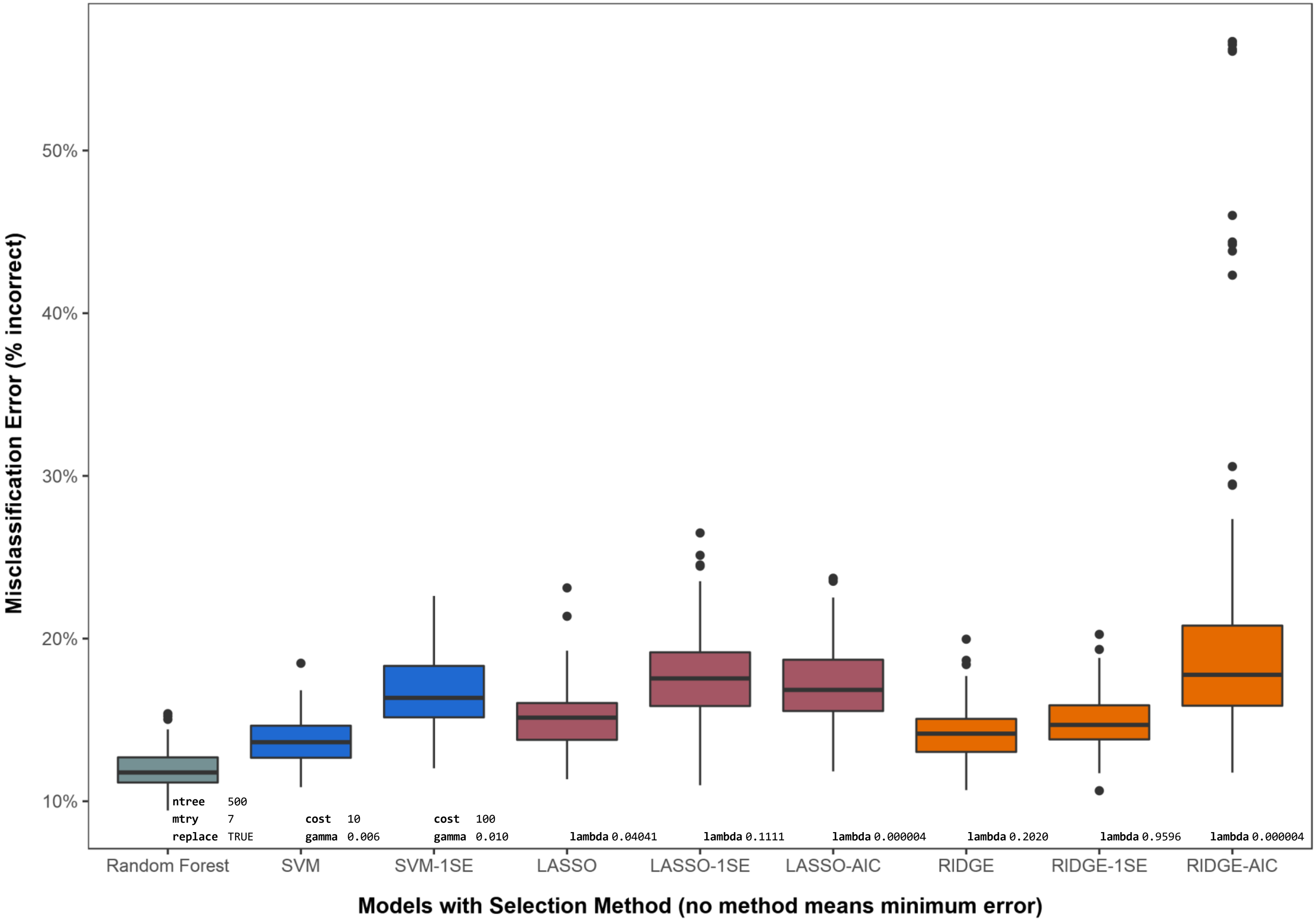
Capital Run Length
The length of a sequence of capital letters

```
capital_run_length_longest
capital_run_length_average
capital_run_length_total
```

3 predictors (6%)

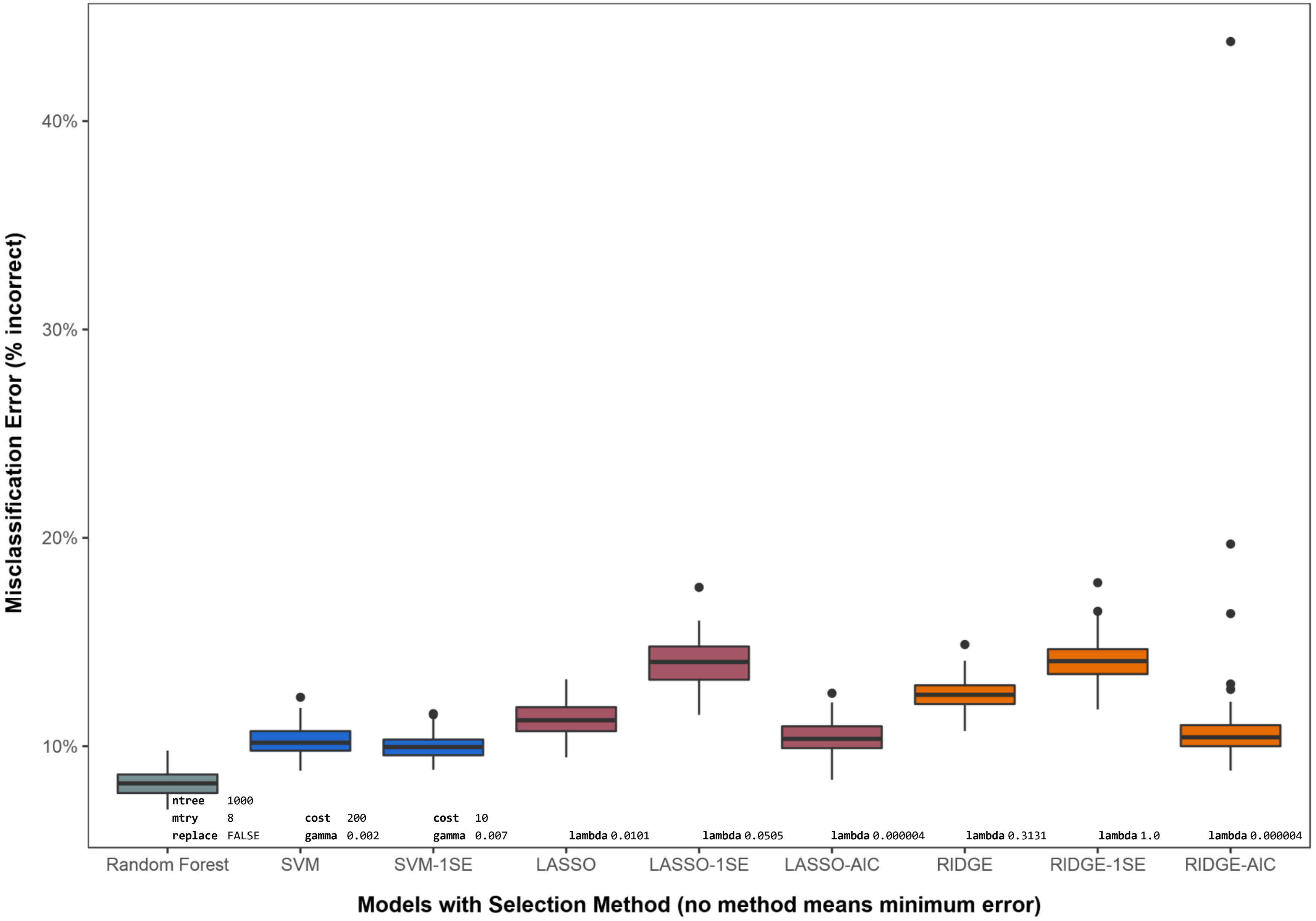
Misclassification Error for Learning Size 2p, nlearn=108

Parameters tuned through a single 10-fold Cross Validation run



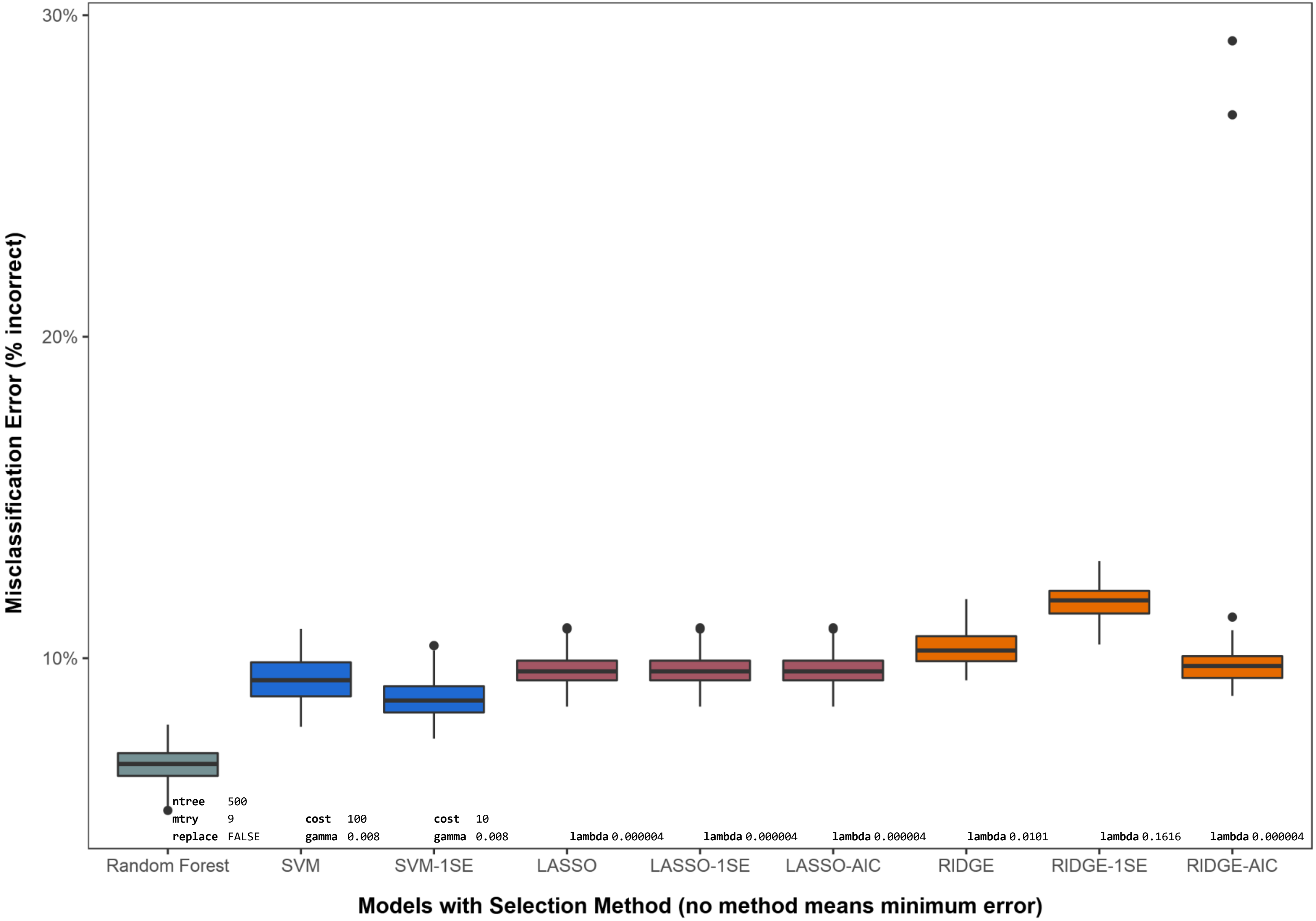
Misclassification Error for Learning Size 10p, nlearn=540

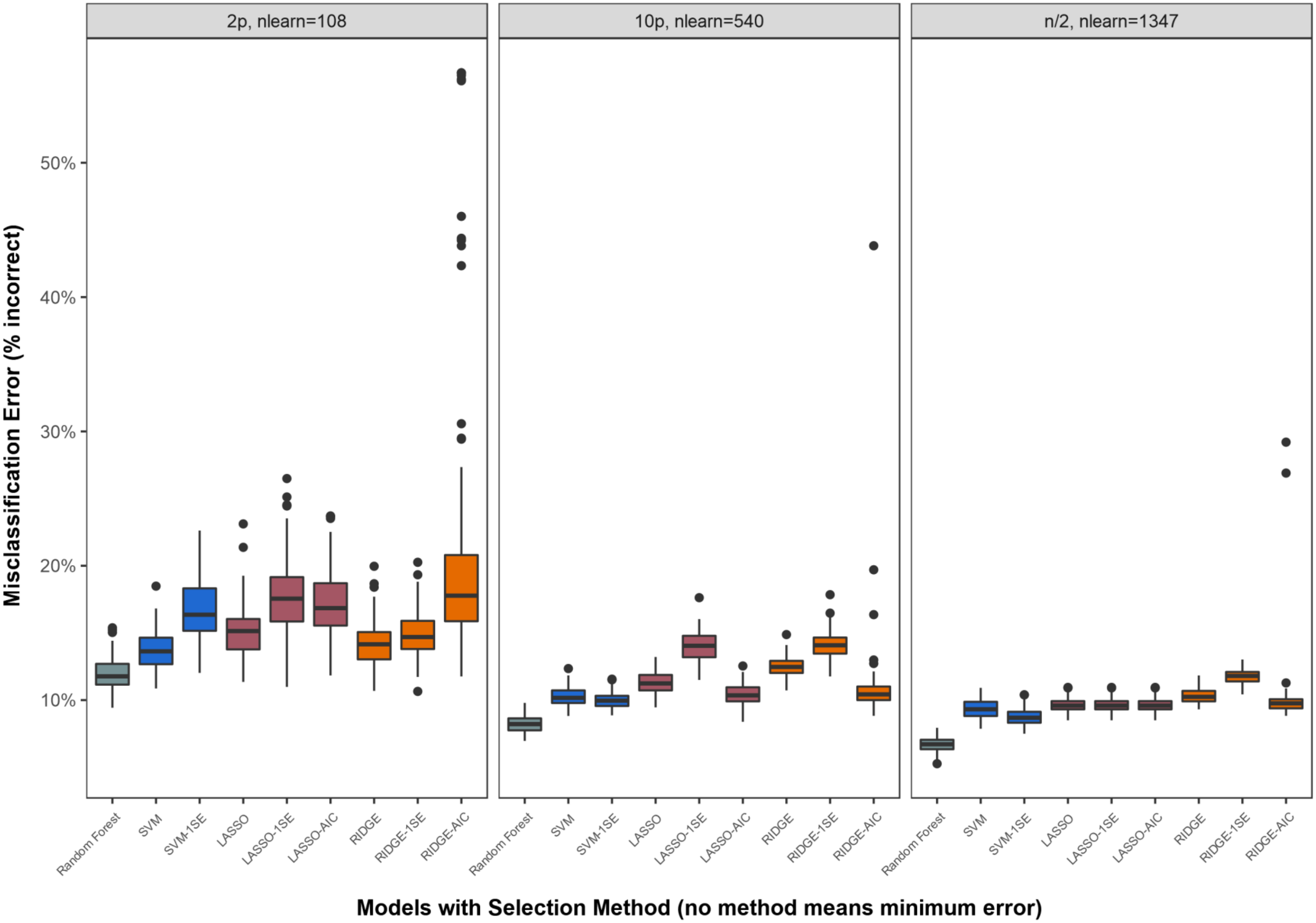
Parameters tuned through a single 10-fold Cross Validation run



Misclassification Error for Learning Size n/2, nlearn=1347

Parameters tuned through a single 10-fold Cross Validation run



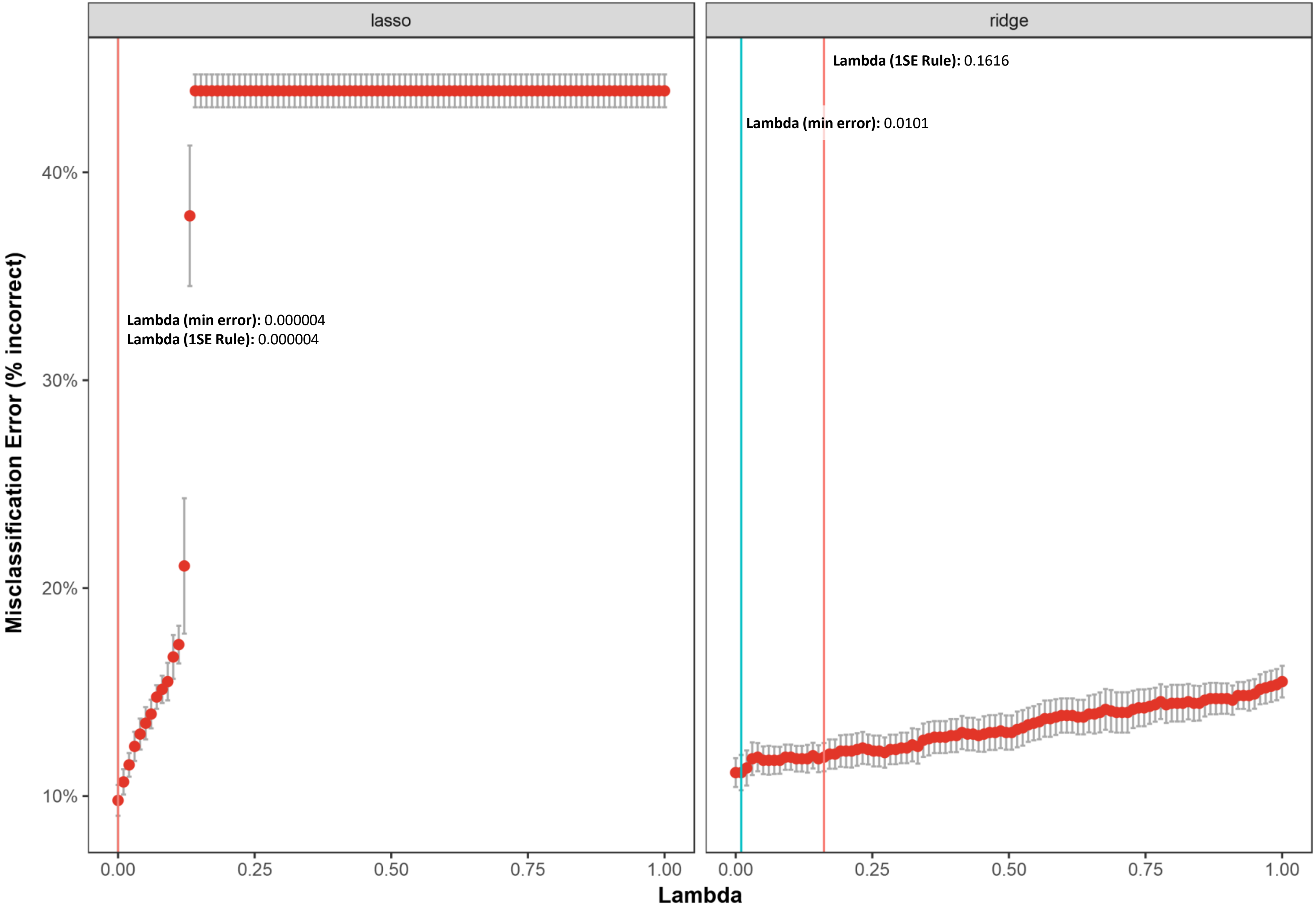


Misclassification Error for All Learning Data Sizes

Parameters tuned through a single 10-fold Cross Validation run

Ridge Regression 10-fold Cross Validation Curve

Curve for n/2, nlearn=1347 learning set



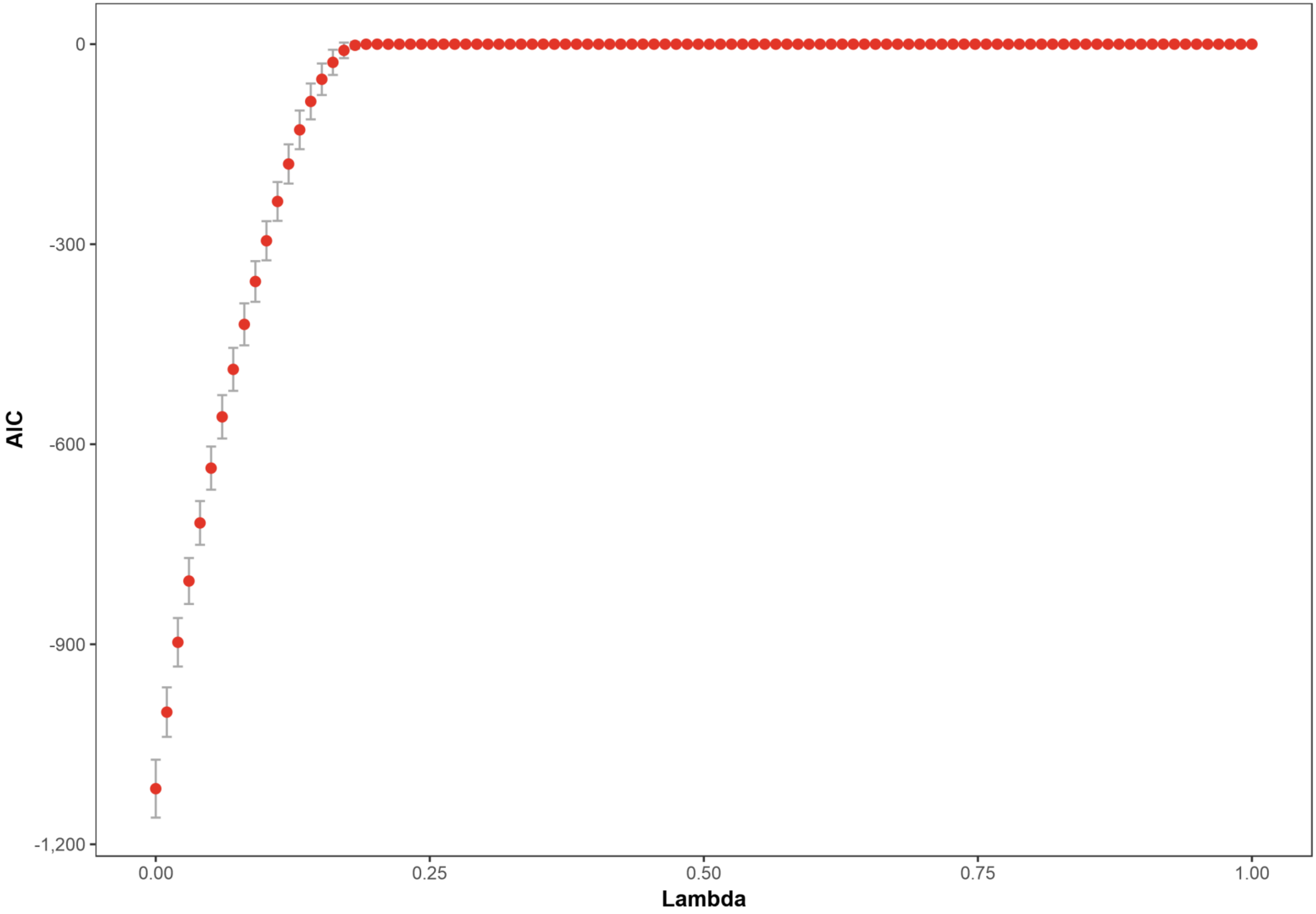
Lambda (min error): 0.000004
Lambda (1SE Rule): 0.000004

Lambda (1SE Rule): 0.1616

Lambda (min error): 0.0101

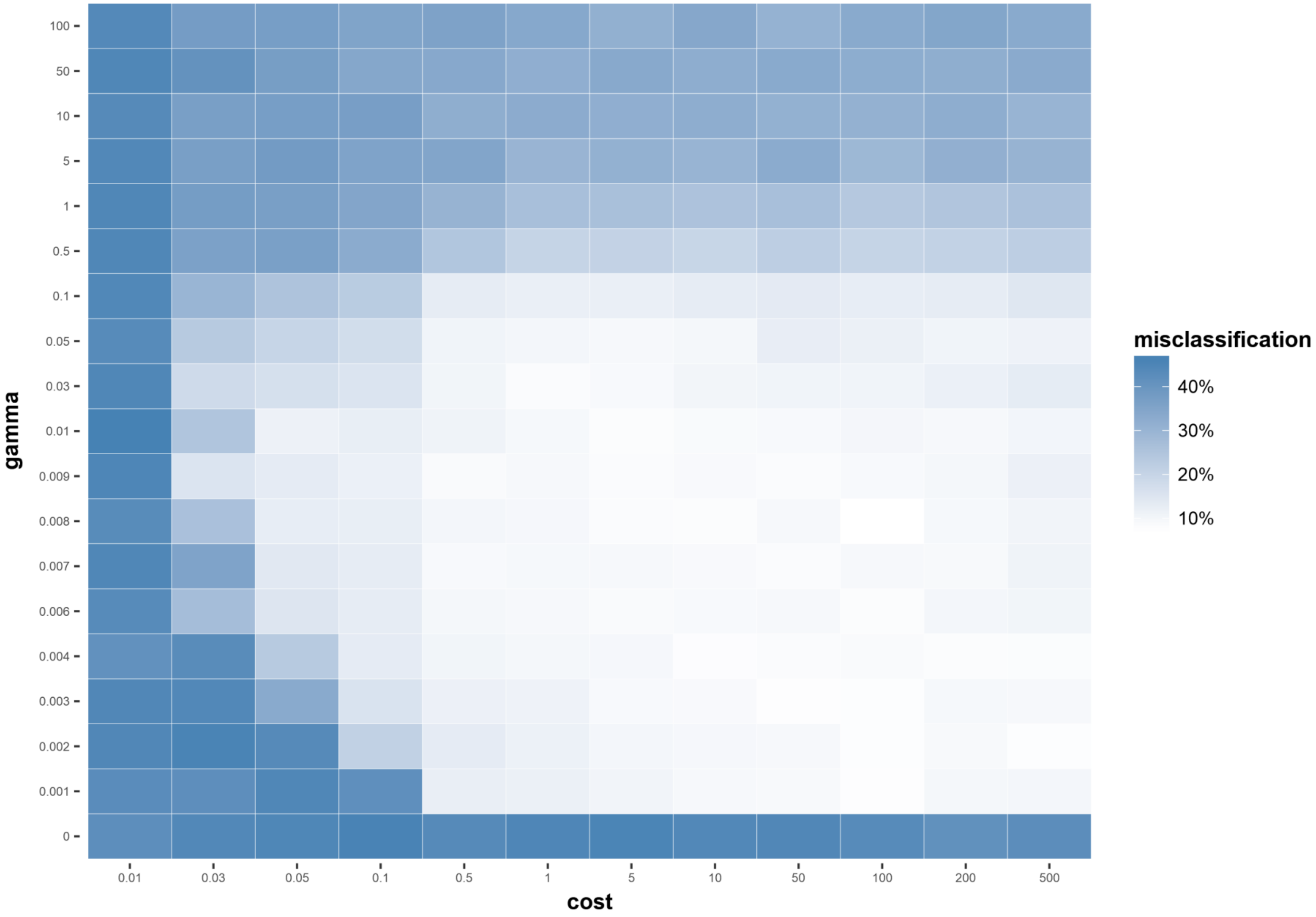
Lasso Regression AIC for 100 Repeated Sampling

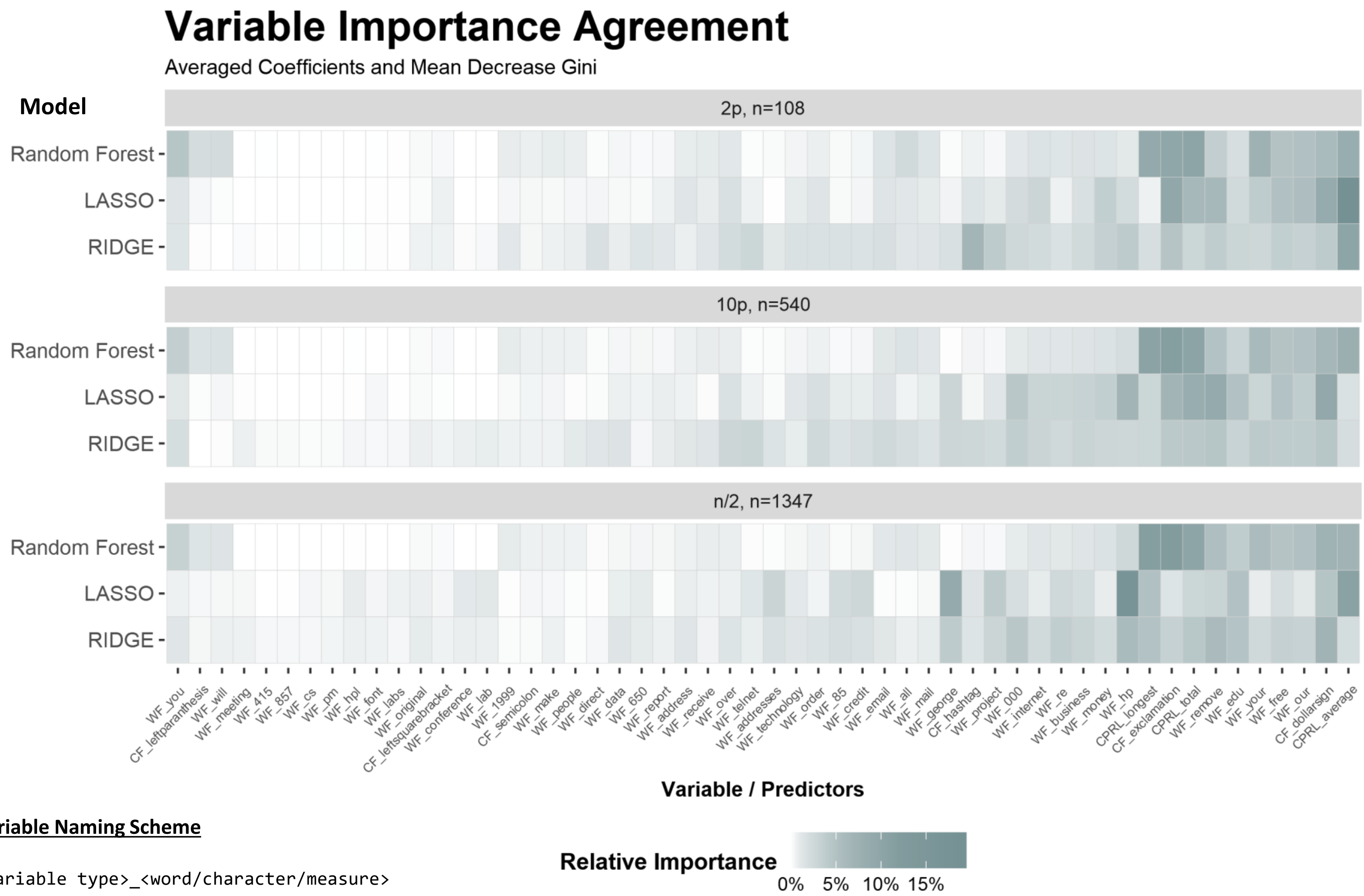
Curve for n/2, nlearn=1347 learning set



Support Vector Machine Parameter Performance

SVM heatmap for n/2, nlearn=1347 learning set





Variable Naming Scheme

```
<variable type>_<word/character/measure>
```

WF – **W**ord **F**requency (% of total words in an email)
CF – **C**haracter **F**requency (% of total characters in an email)
CPRL – **C**a**p**ital **R**un **L**ength (The length of a sequence of capital letters)