

# 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

<b>Number of Emails (n)</b>	2695
<b>Number of Predictors (p)</b>	$57 - 3 = 54$ (3 predictors are all zero)
<b>Response</b>	1512 (~56%) spam 1183 (~44%) non-spam

Sourced Spambase from:

## UCI Machine Learning Repository

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

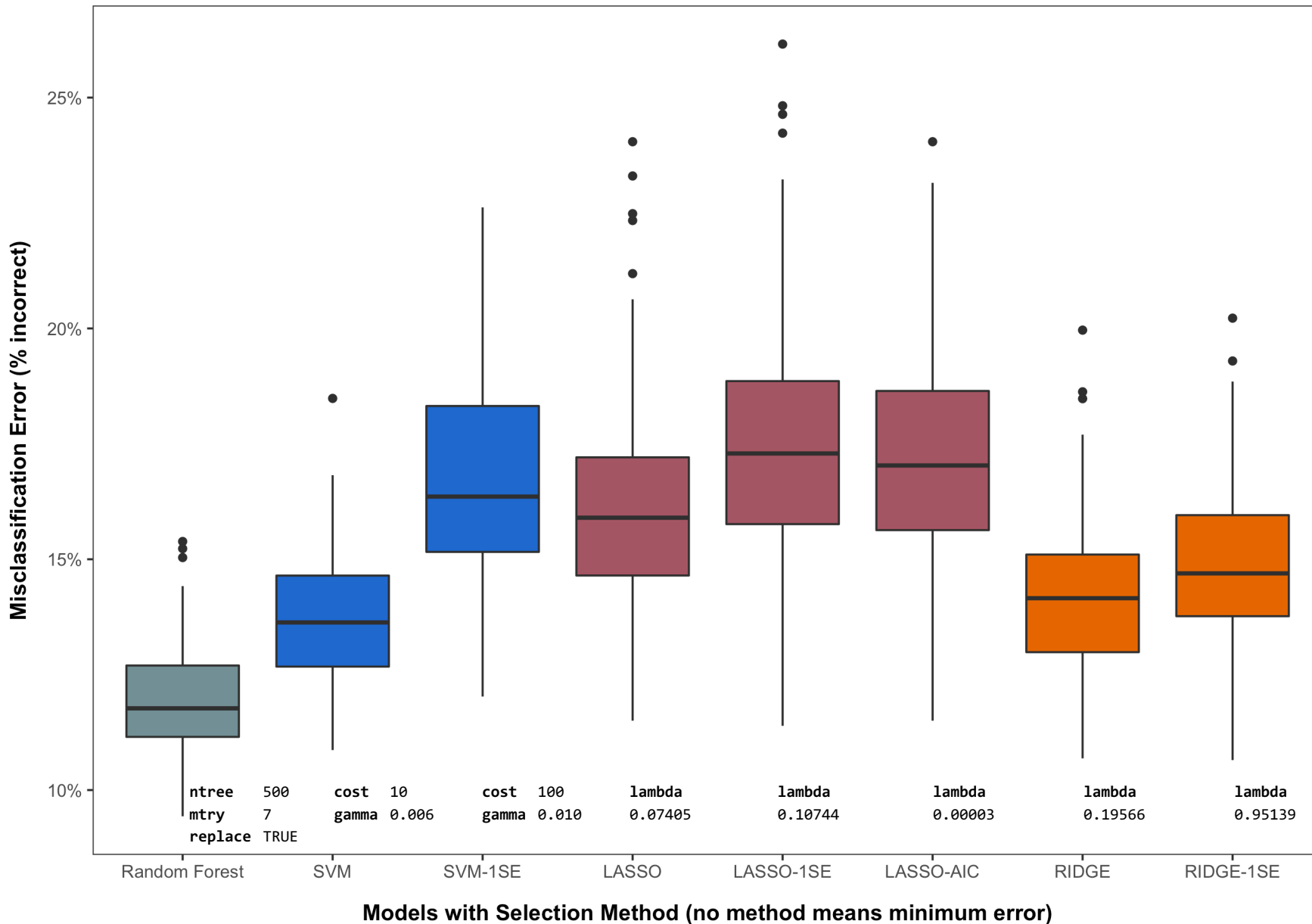
## Predictor Description

Predictors are broken into three types of measures

<b>Word Frequency</b> % of total words in an email	word_freq_make word_freq_internet word_freq_hp ... ...	<b>Character Frequency</b> % of total characters in an email	char_freq_semicolon char_freq_exclamation char_freq_dollarsign ... ...	<b>Capital Run Length</b> The length of a sequence of capital letters	capital_run_length_longest capital_run_length_average capital_run_length_total
45 predictors (83%)		6 predictors (11%)		3 predictors (6%)	

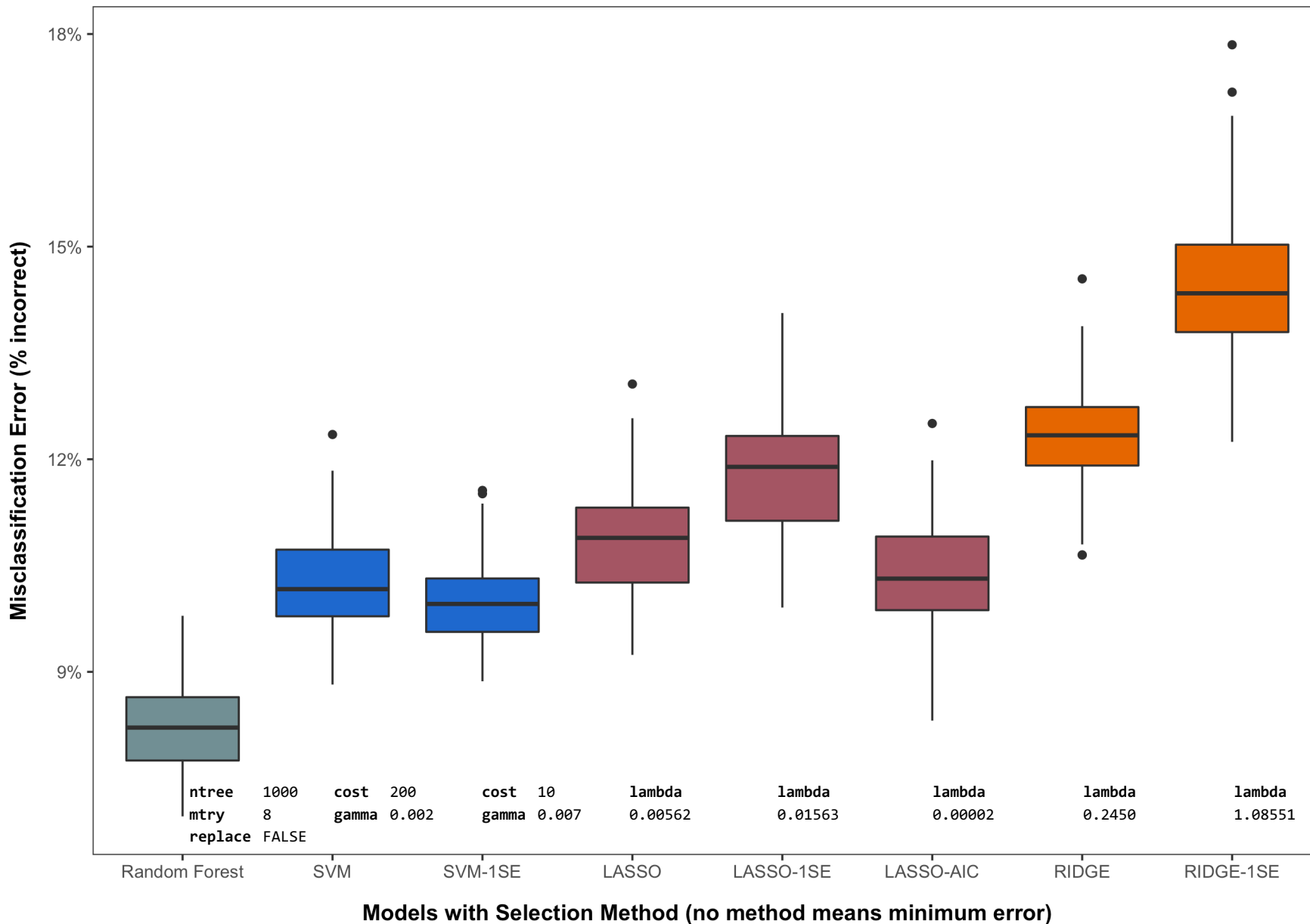
# Misclassification Error for Learning Size 2p, nlearn=108

Parameters tuned through a single 10-fold Cross Validation run



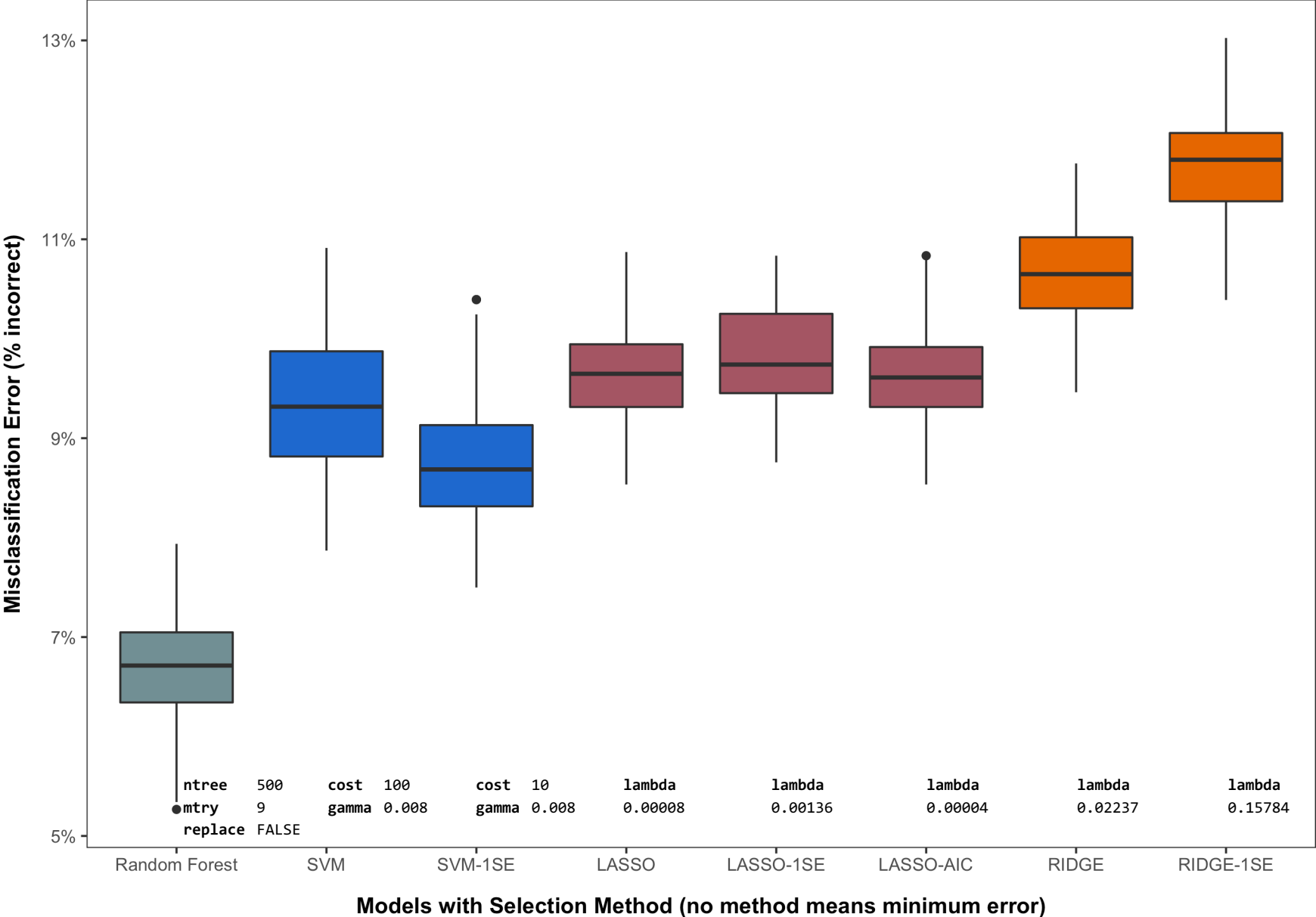
# Misclassification Error for Learning Size 10p, nlearn=540

Parameters tuned through a single 10-fold Cross Validation run



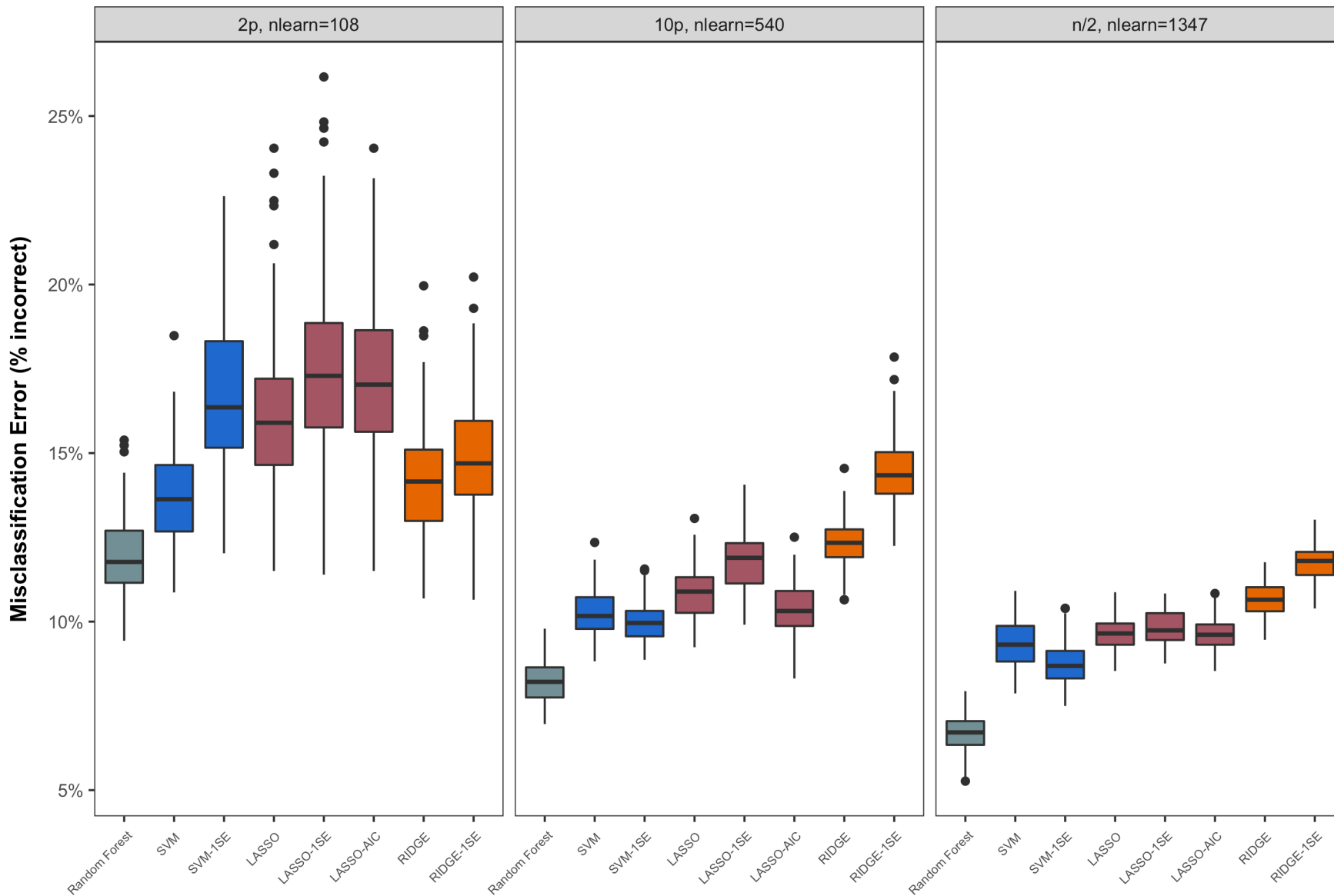
# Misclassification Error for Learning Size $n/2$ , $n_{\text{learn}}=1347$

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



Models with Selection Method (no method means minimum error)

# Cross Validation Curve

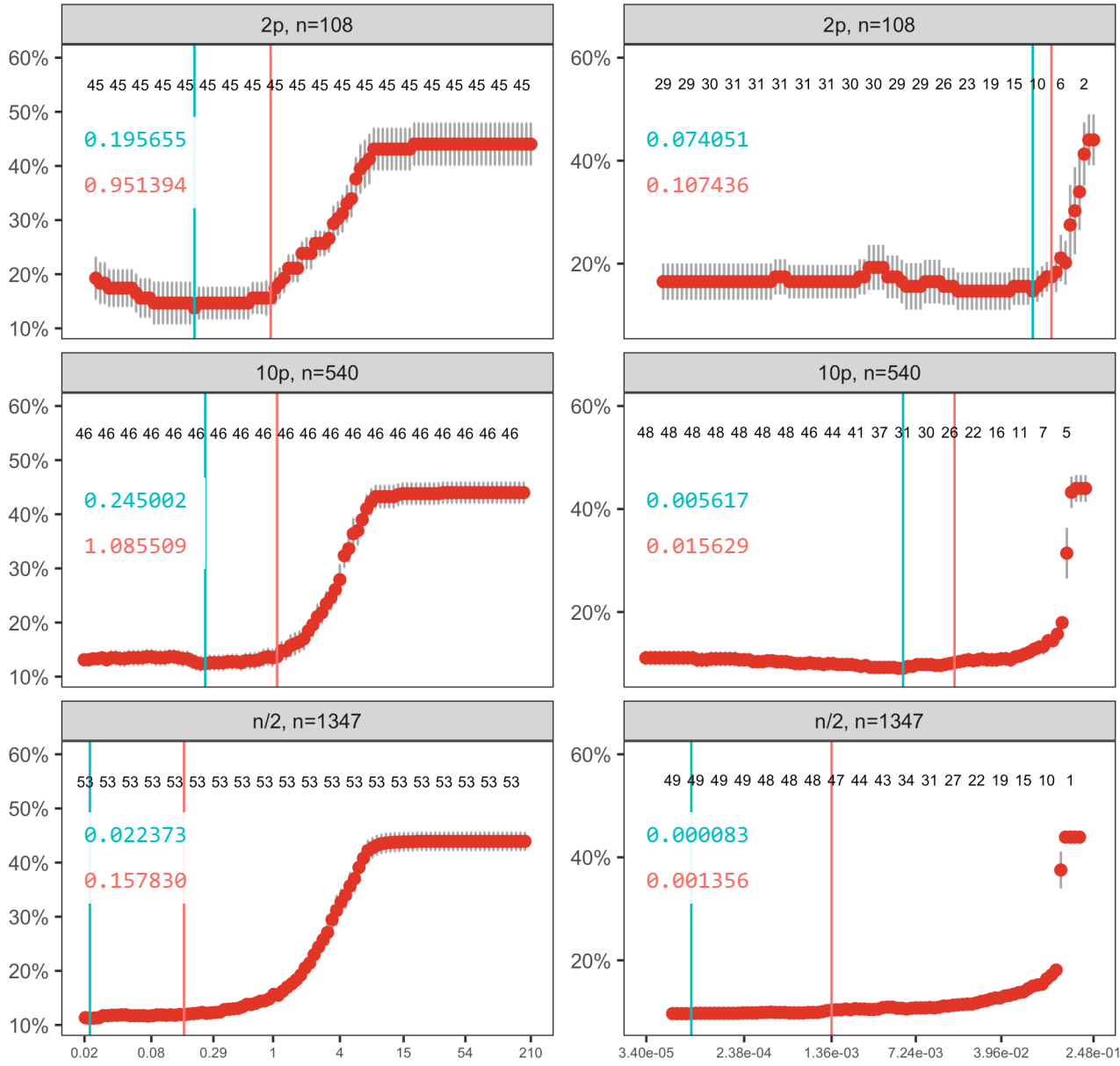
Ridge

Minimum Error

1 Standard Error

Lasso

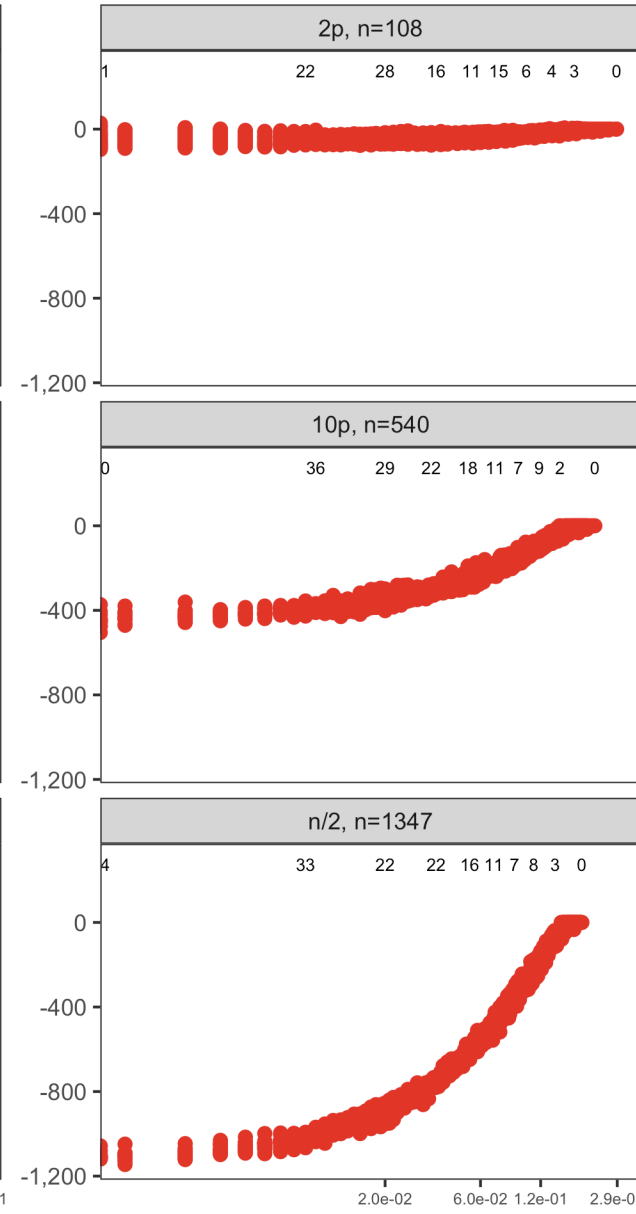
*X-axis is Misclassification Error Y-axis is Lambda Values*



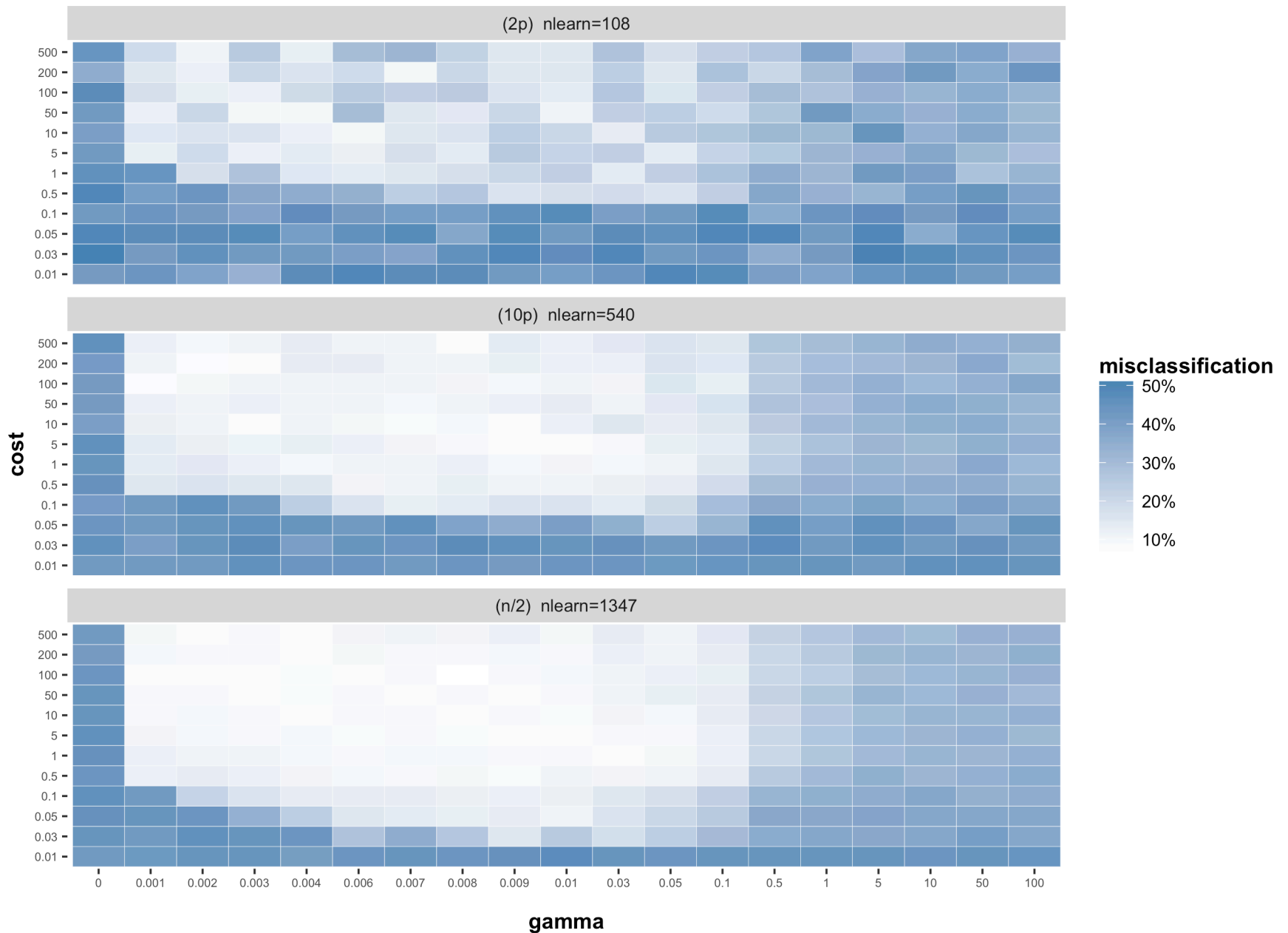
# Lasso AICC Curve

Lasso

*X-axis is AICC Y-axis is Lambda Values*



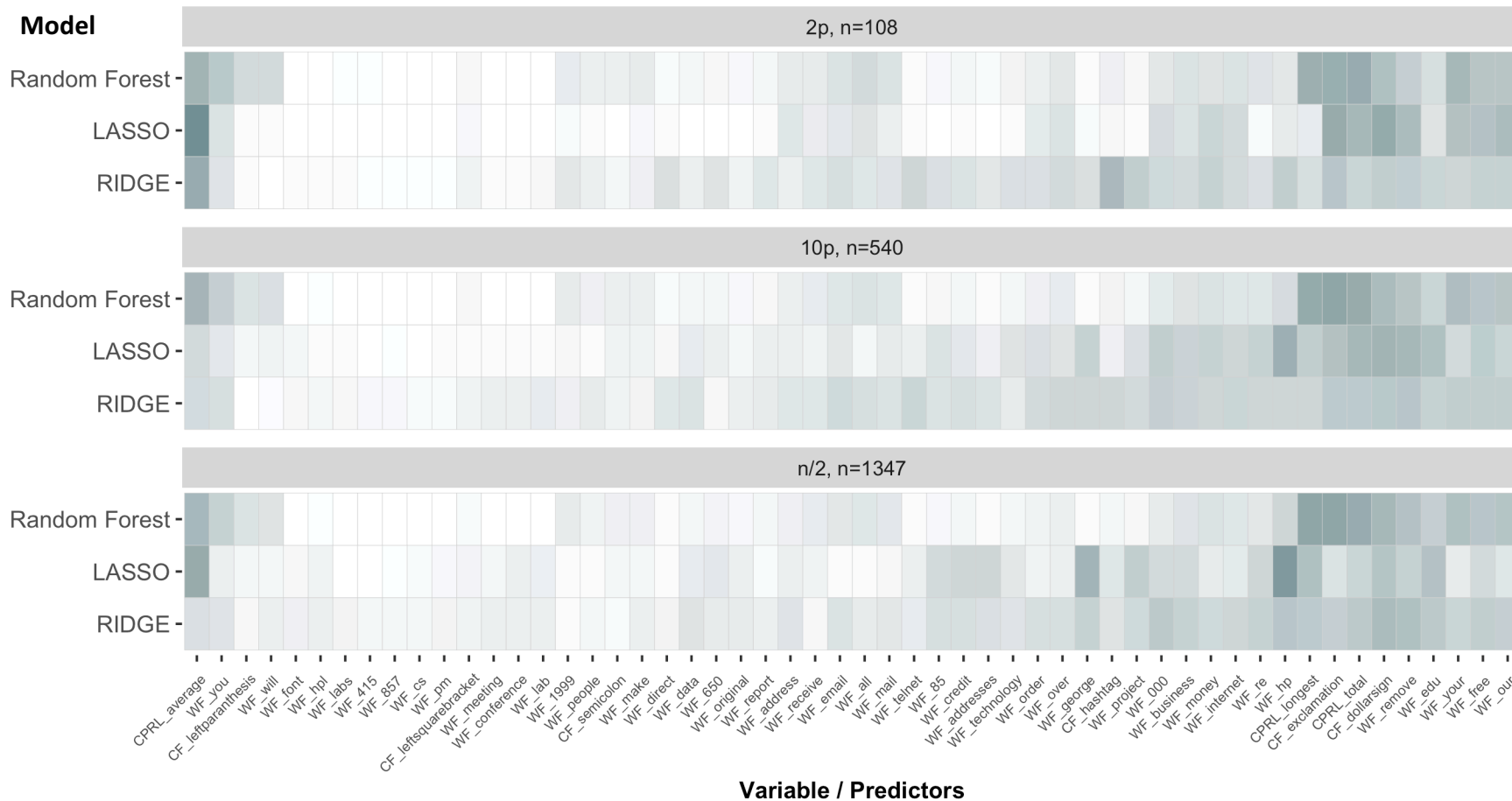
# Support Vector Machine Parameter Performance





# Variable Importance Agreement

Averaged Coefficients and Mean Decrease Gini



## Variable Naming Scheme

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

**WF** – Word Frequency (% of total words in an email)

**CF** – Character Frequency (% of total characters in an email)

**CPRL** – Capital Run Length (The length of a sequence of capital letters)

Relative Importance

