Stay Alert!
The Ford Challenge

—2011 kaggle competition

Jumao Yuan and Ruoxi Chen

EXST 7152 Midterm Project
March 31, 2015

Agenda

Background

Data Desciption

Data Preprocessing

Evaluation

Models

Results

Summary

Agenda

- Background
- Data Desciption
- Data Preprocessing
- Evaluation
- Models
- Results
- Summary

https://images.google.com/

Agenda

Background

Data Desciption

Data Preprocessing

Evaluation

Models

Results

Summary

30 independent variables
604329 observations

- ✓**Physiological (8)** **P1, P2, ……, P8**
- ✓**Environmental (11)** **E1, E2, ……, E11**
- ✓**Vehicular (11)** **V1, V2, ……, V11**

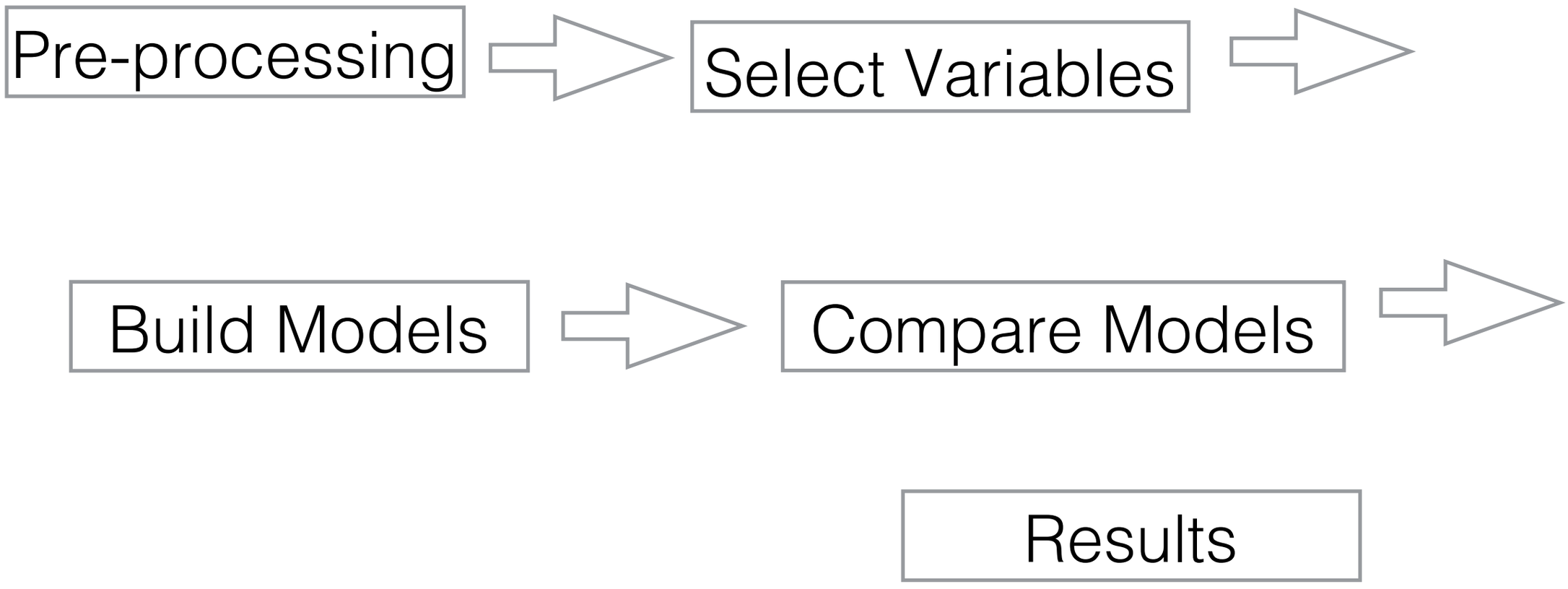
Goal:

Predict response variable “IsAlert”

- IsAlert = 1 if the driver is alert
- IsAlert = 0 if the driver is not alert

https://www.kaggle.com/c/stayalert/data.

Workflow


Pre-processing
Select Variables
Build Models
Compare Models
Results

Agenda

Background

Data Desciption

Data Preprocessing

Evaluation

Models

Results

Summary

Types of bad data:

1. missing values: NA, Unknow, NULL
2. Typos: 0 (numeric), negative values, possibility>1

Methods:

Data Deletion: easy to implement and fast

Data Imputation: complicated but more accurate

Here, **<0.1%** missing values, use data deletion

Agenda

Background

Data Desciption

Data Preprocessing

Evaluation

Models

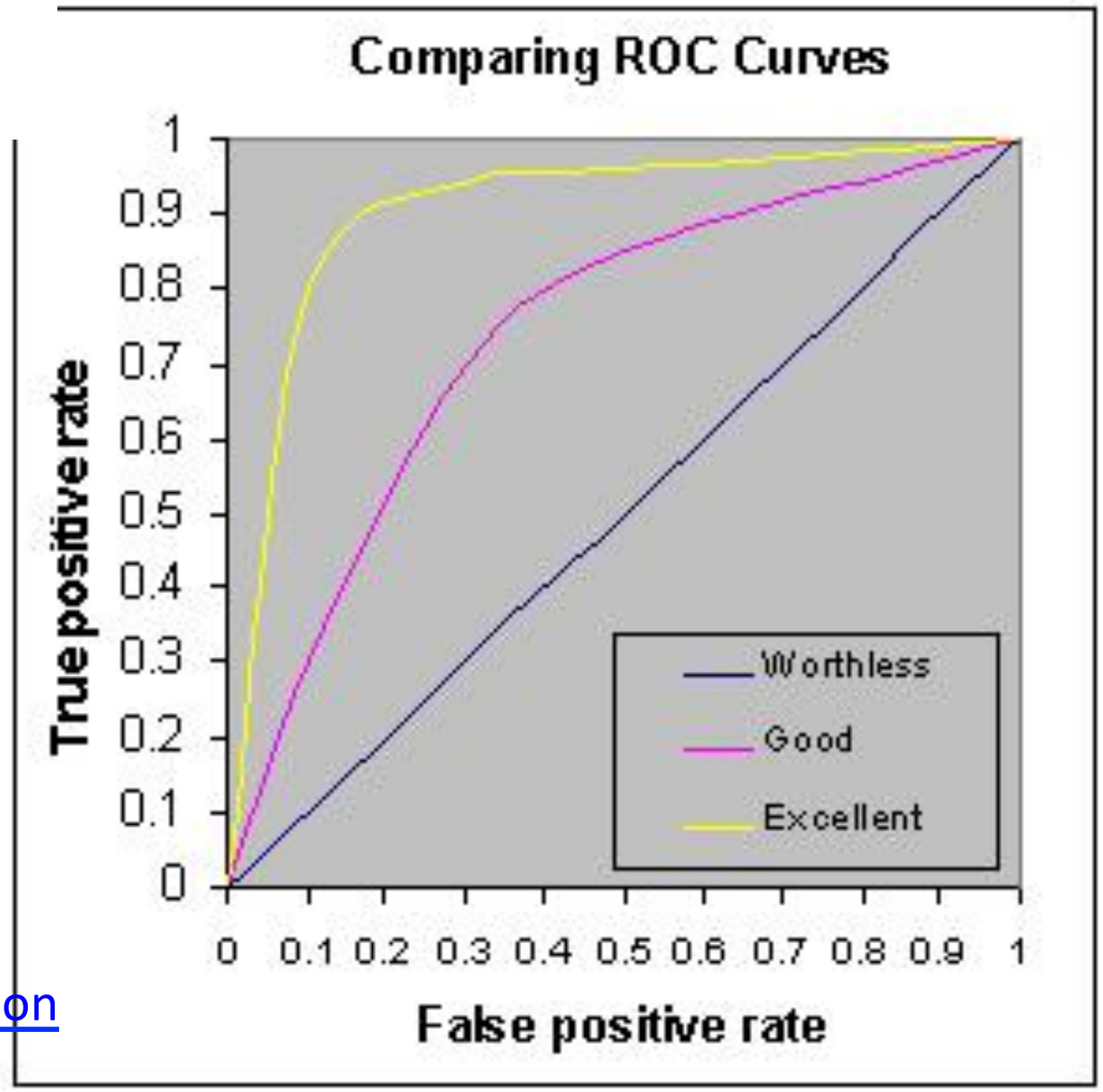
Results

Summary

	p' (Predicted)	n' (Predicted)
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

AUC Score

ROC Curve



https://www.kaggle.com/c/stayalert/details/Evaluation

http://gim.unmc.edu/dxtests/roc3.htm

Agenda

Background

Data Desciption

Data Preprocessing

Evaluation

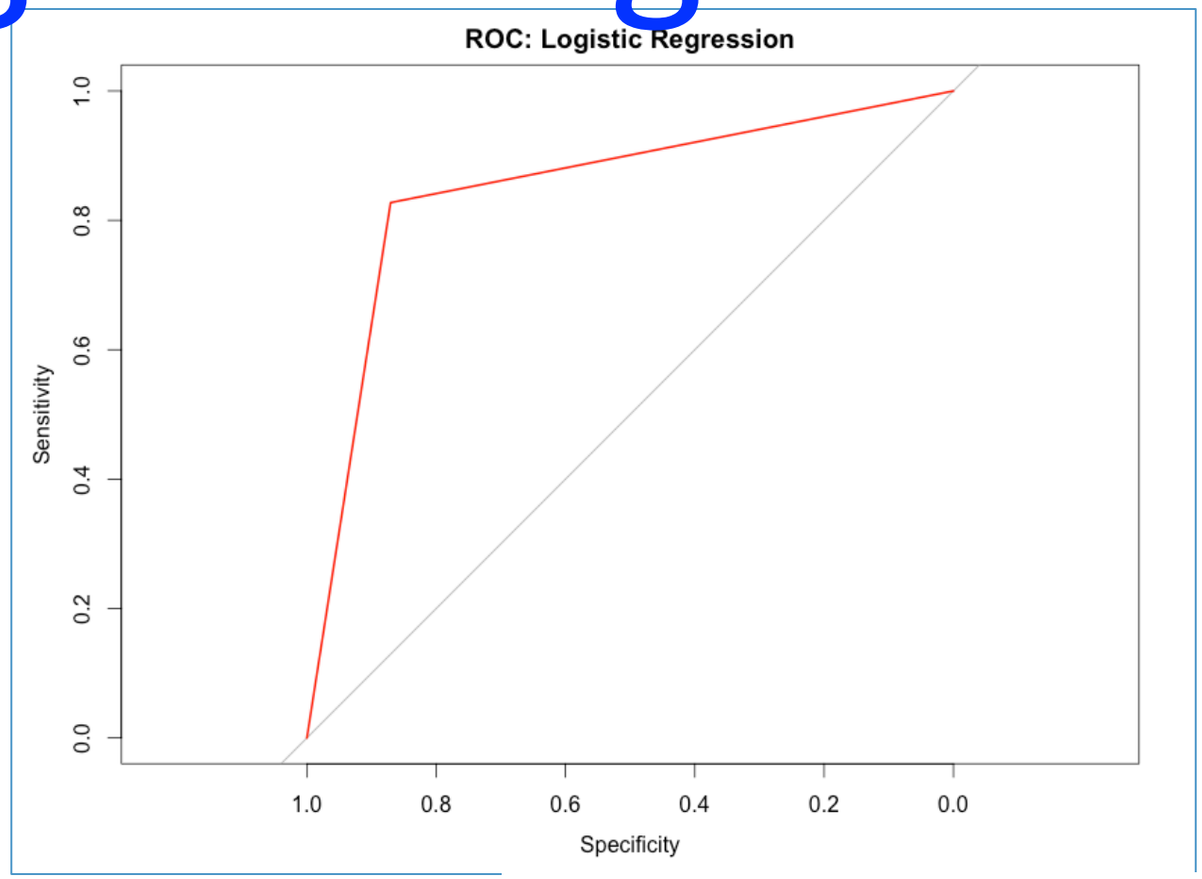
Models

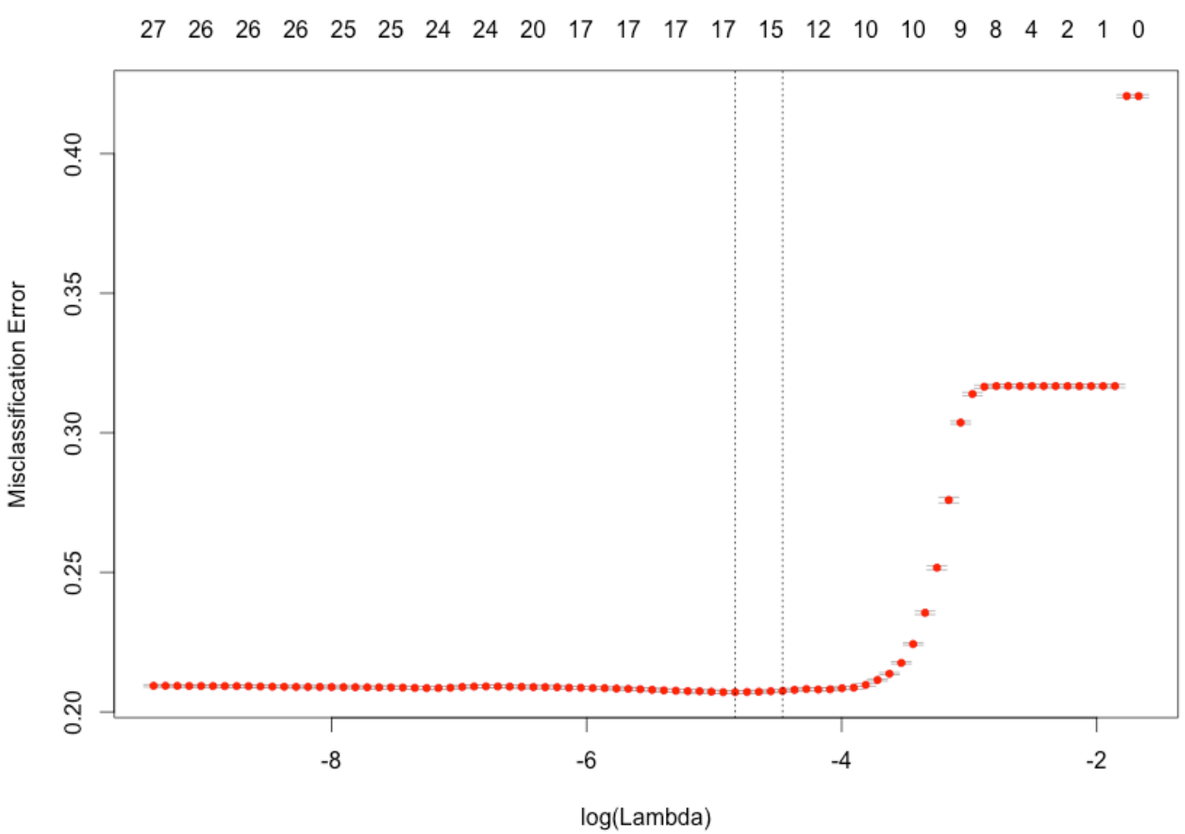
Results

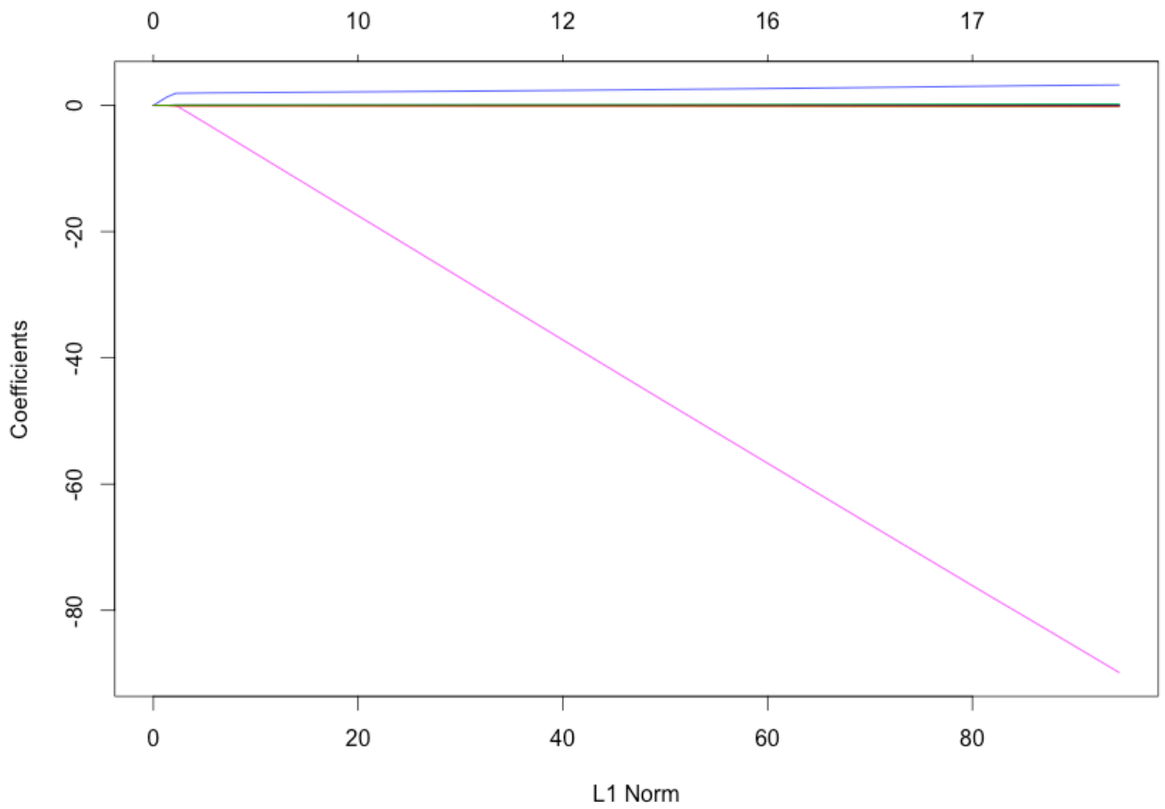
Summary

Logistic Regression

Library(glmnet)

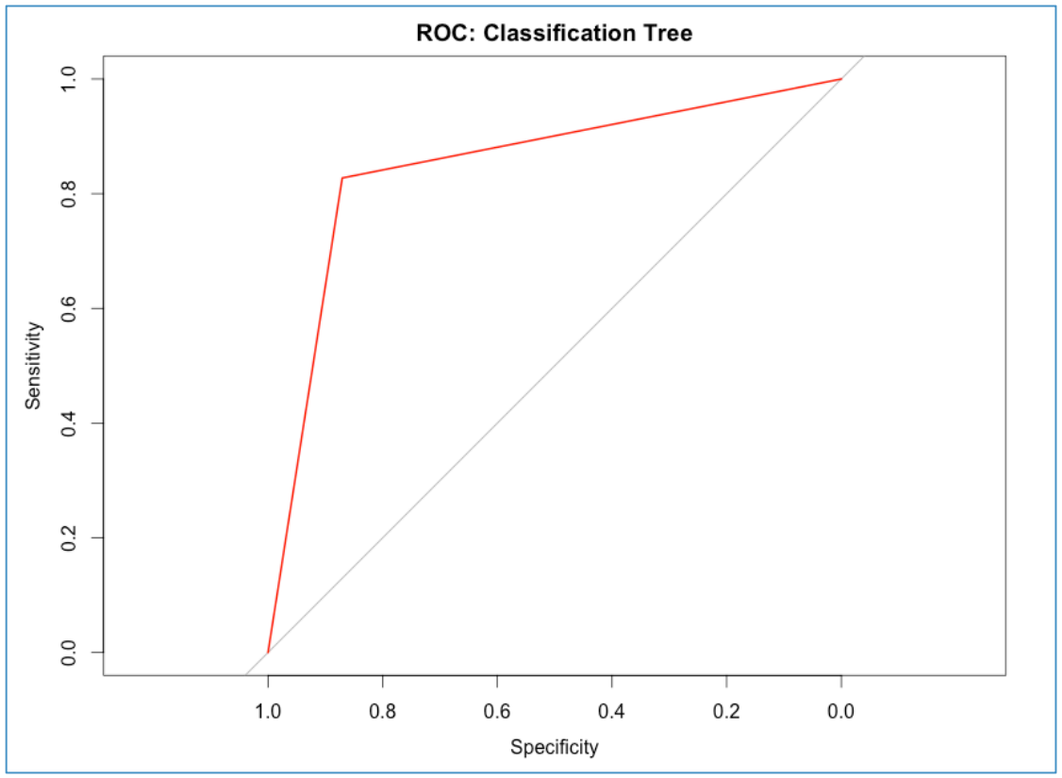


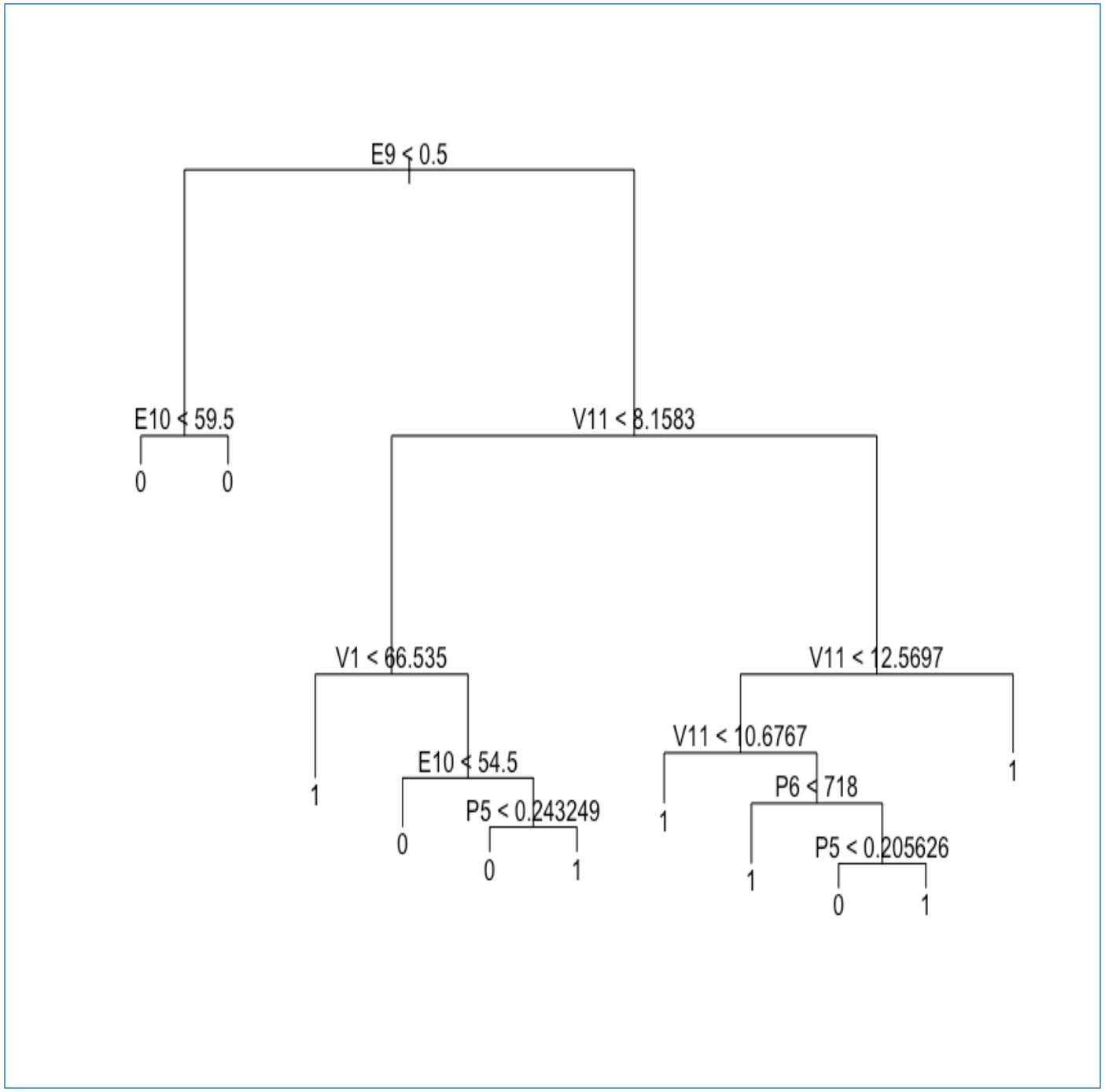




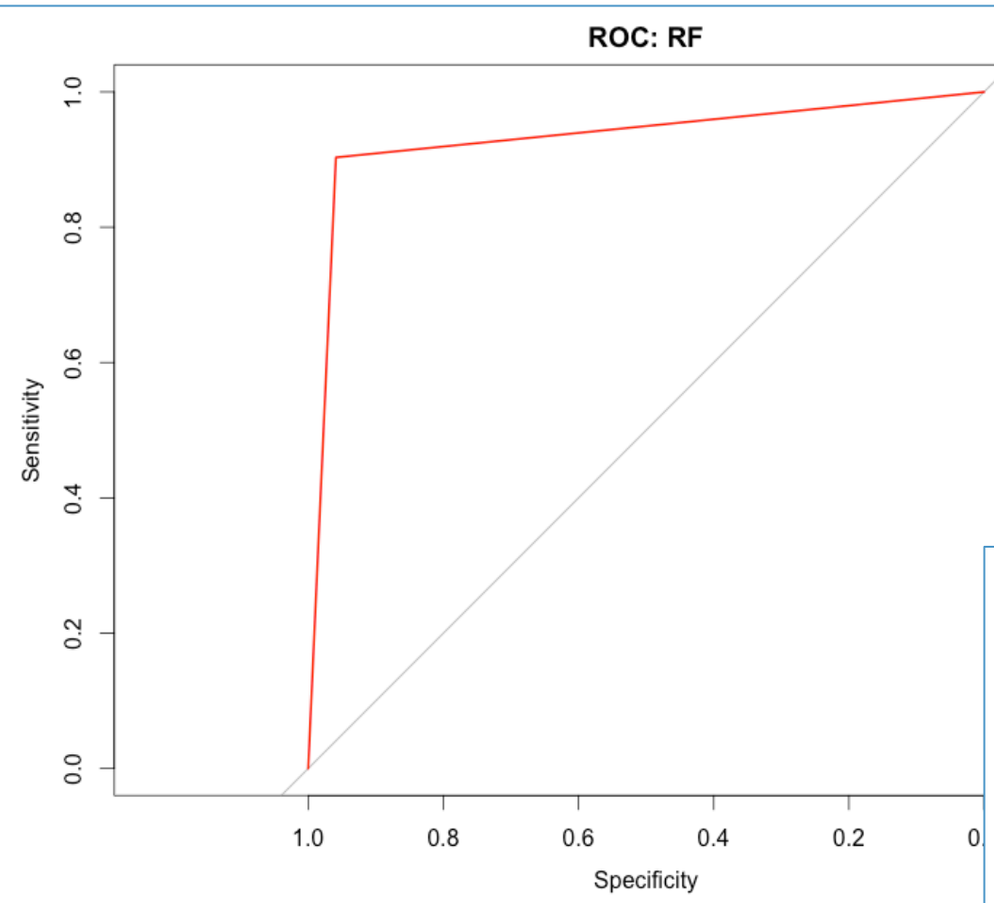
Classification Tree

Library(tree)



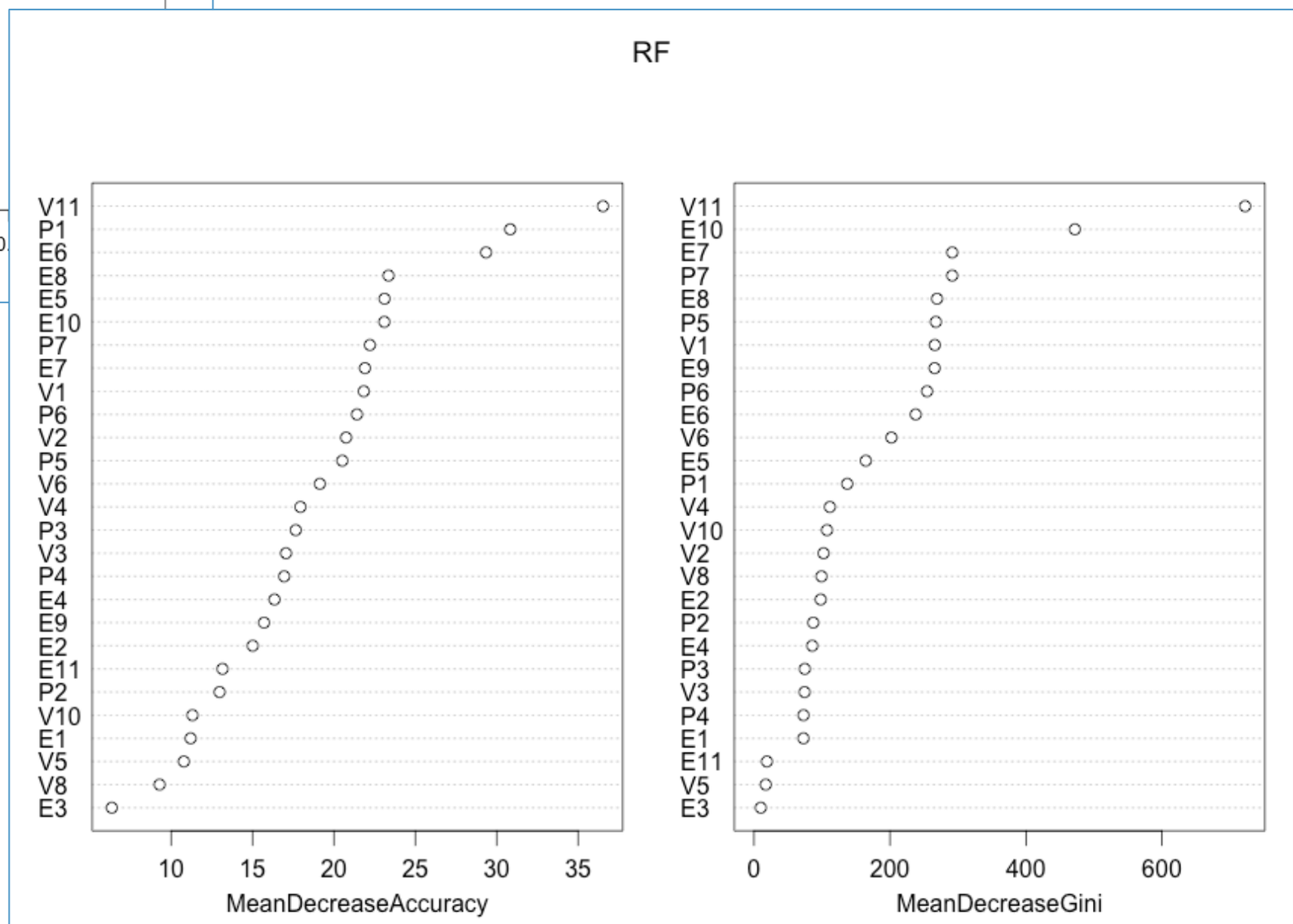


Random Forest



Library(randomForest)

RF <- randomForest
(training[,-c(1,9,27,29)],
factor(trainig$IsAlert),
sampsize=10000,
do.trace=TRUE,
importance=TRUE,
ntree=**100**,
forest=TRUE)



Other models

- ✓Naïve Bayes
- ✓SVM
- ✓GLM
- ✓Neural Network (NN)
- ✓CART – regression tree

Agenda

Background

Data Desciption

Data Preprocessing

Evaluation

Models

Results

Summary

Method	AUC Score	Variables Selected	Computation Time (s)
Logistic	0.78	23 vars	10.722
Random Forest	0.93	V11 E10 E7	393.314
Decistion Tree	0.83	V11 E10 P5 P6 V1	10.722
Naive Bayes	0.76	–	–
NN(two layer)	0.77	–	–
SVM	0.73	–	–

Reference: “First place in the ’Stay Alert!’ competiion”, inference, March 2011.

Agenda

Background

Data Desciption

Data Preprocessing

Evaluation

Models

Results

Summary

- ✓Random Forest works good
- ✓Only V11, E10, etc. variables are important
- ✓R is slow for large data computation

--- Python, Perl, R on HPC?

Thanks ☺