ds2_final_report

Weijia Xiong, Yimeng Shang, Xue Jin

5/15/2020

Introduction

Heart disease is one of the leading cause of death for both men and women all over the world. Heart disease detection based on typical clinical features is being increasingly critical to population health. It is a desirable way to help people enhance self protection ability and prevent faults. In this report, we built several predictive models to predict the heart disease.

There are 303 observations, 14 predictions in the data set and 2 missing values.

- age: age in years **continuous variable**
- sex: (1 = male; 0 = female) **categorical variable**
- cp: chest pain type **categorical variable**
- trestbps: resting blood pressure (in mm Hg on admission to the hospital) **continuous variable**
- chol: serum cholestoral in mg/dl **continuous variable**
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) **categorical variable**
- restecg: resting electrocardiographic results **categorical variable**
- thalach: maximum heart rate achieved **continuous variable**
- exang: exercise induced angina (1 = yes; 0 = no) **categorical variable**
- oldpeak: ST depression induced by exercise relative to rest **continuous variable**
- slope: the slope of the peak exercise ST segment **categorical variable**
- ca: 3 = normal; 6 = fixed defect; 7 = reversable defect **categorical variable**
- thal: 3 = normal; 6 = fixed defect; 7 = reversable defect **categorical variable**
- target: 1 = disease; 0 = not disease **response**

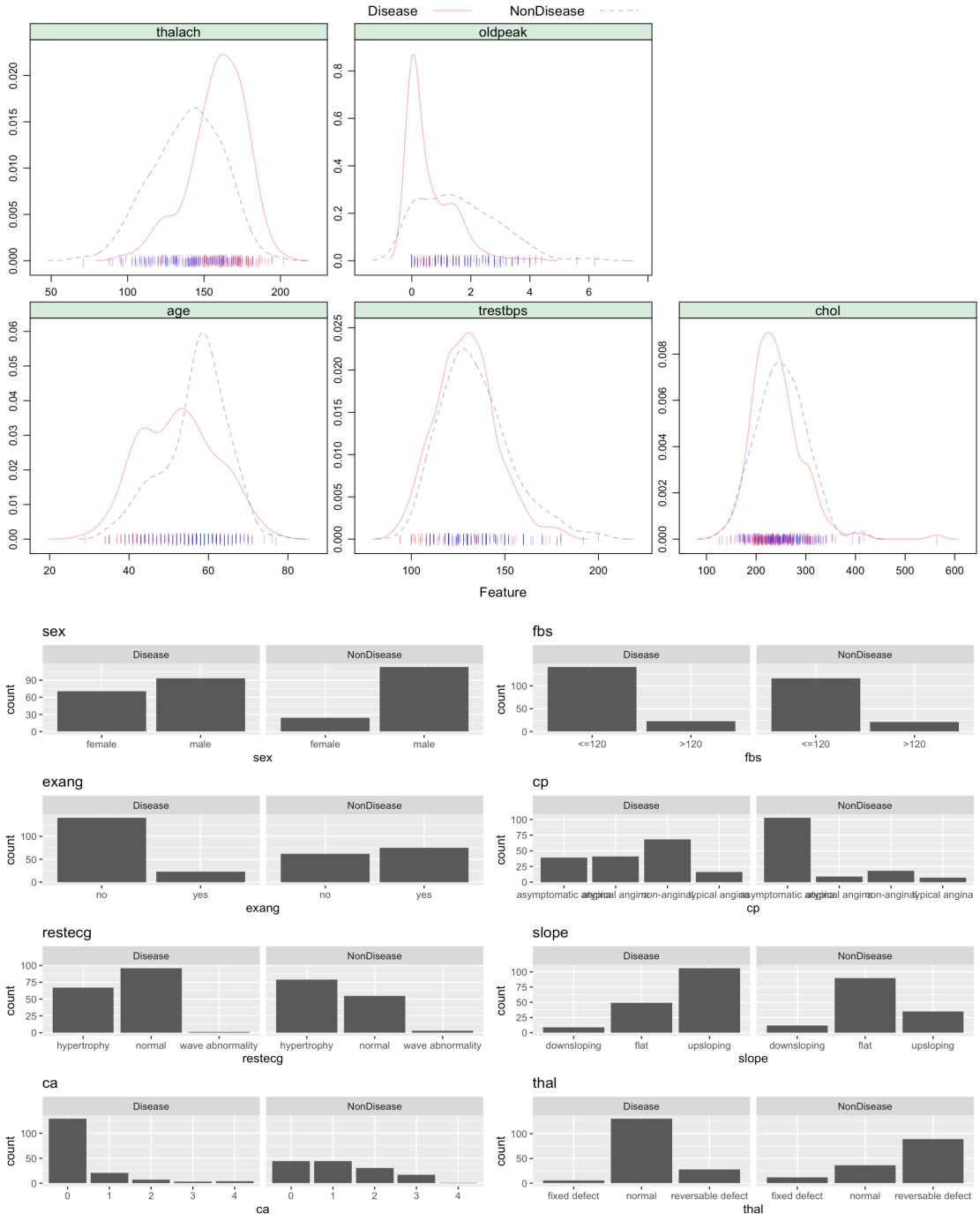
Among the predictors, there are 8 categorical variables and 5 continuous variables. We first recoded all the categorical variables with its true meaning for further usage. Then, we removed missing values and changed all character variables into factor.

In this report, we aim to built a model to predict whether a person is likely to get heart disease or not. We use cross validation to compare the prediction performance of these models.

Exploratory analysis/visualization

Among the continuous variables, thalach/oldpeak/age have very different distribution among those disease and non-disease population. Disease group tend to have higher *thalach* and smaller *oldpeak*.

Among the categorical variables, the distribution of fbs is similar among disease and non-disease groups while others are not similarly distributed.



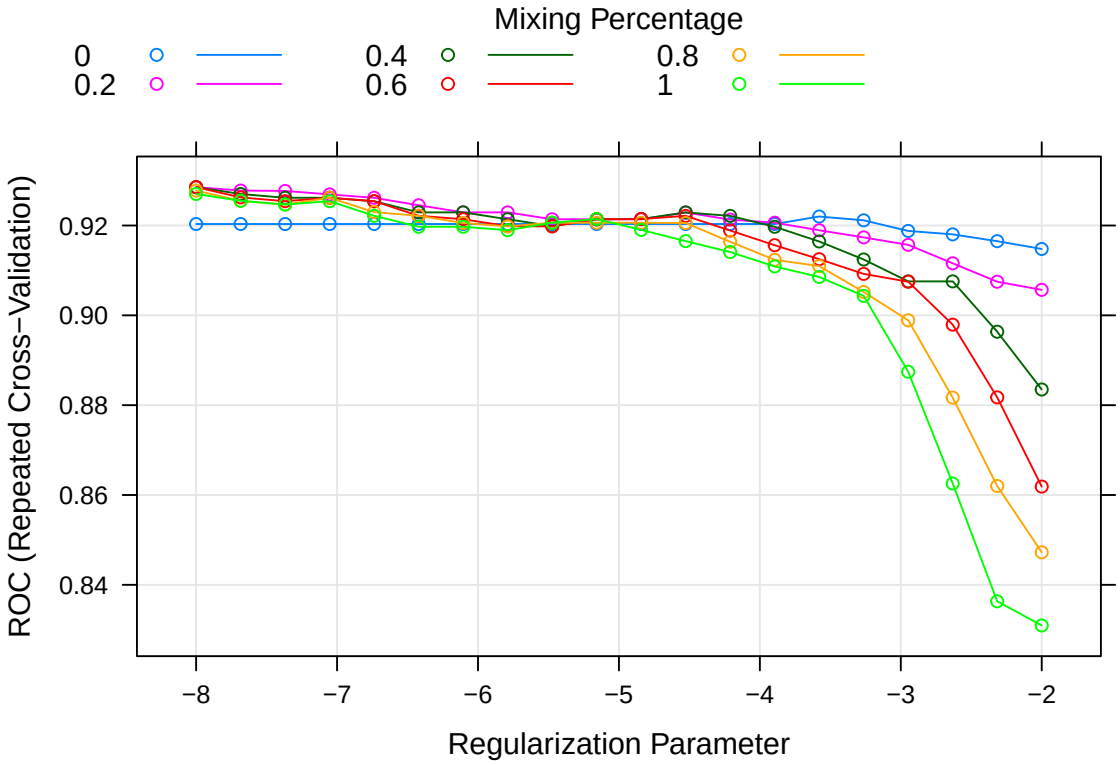
Models

This is a binary classification problem. In this report, we mainly focus on the following models: GLM (logistic regression with penalty term), KNN model, Tree model, Random Forest model, Boosting and Support vector machine. We use all predictor into the models. And we set ROC as the metric.

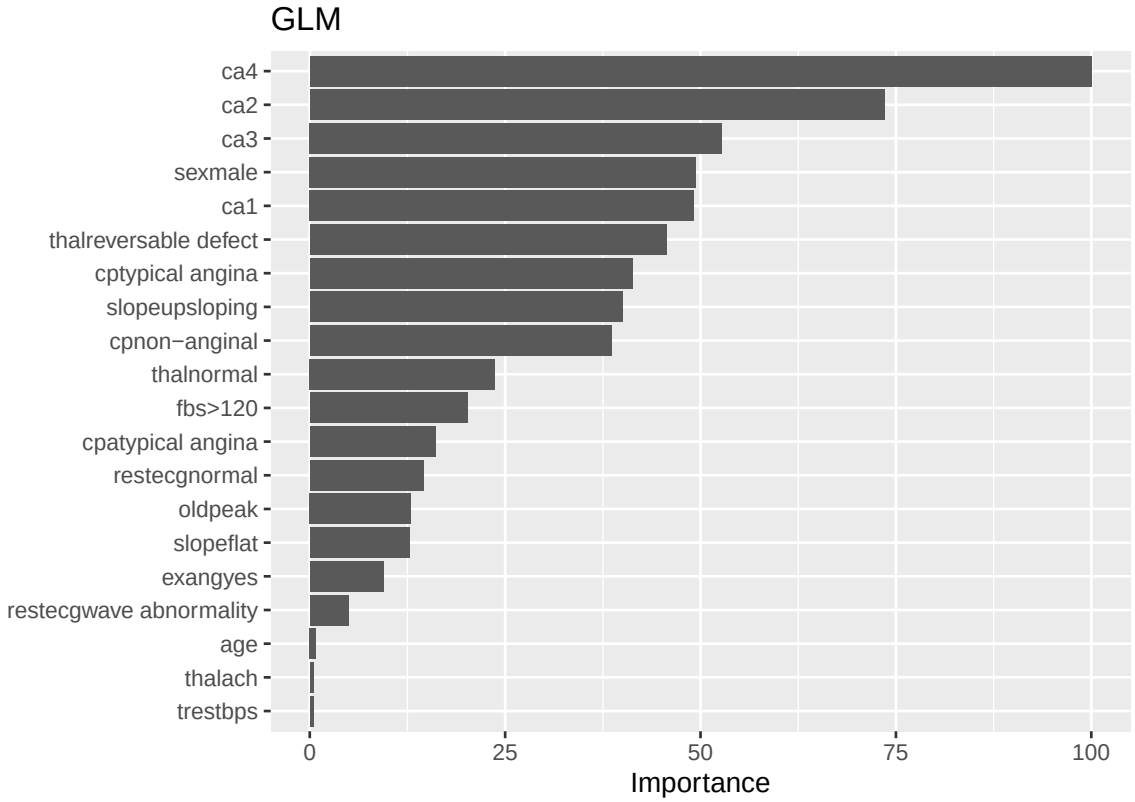
We randomly split the data into 75% train set and 25% test set. We built our model based on the train set and did prediction on the test set. We used repeated CV for models comparison.

GLM

Firstly, we use elsetic model, which is a combination with lasso regression and ridge regression. There are two tuning parameters in the model: $\alpha \in [0, 1]$: $Ridge : \alpha = 0$; $Lasso : \alpha = 1$ and λ: the penalty coefficient.



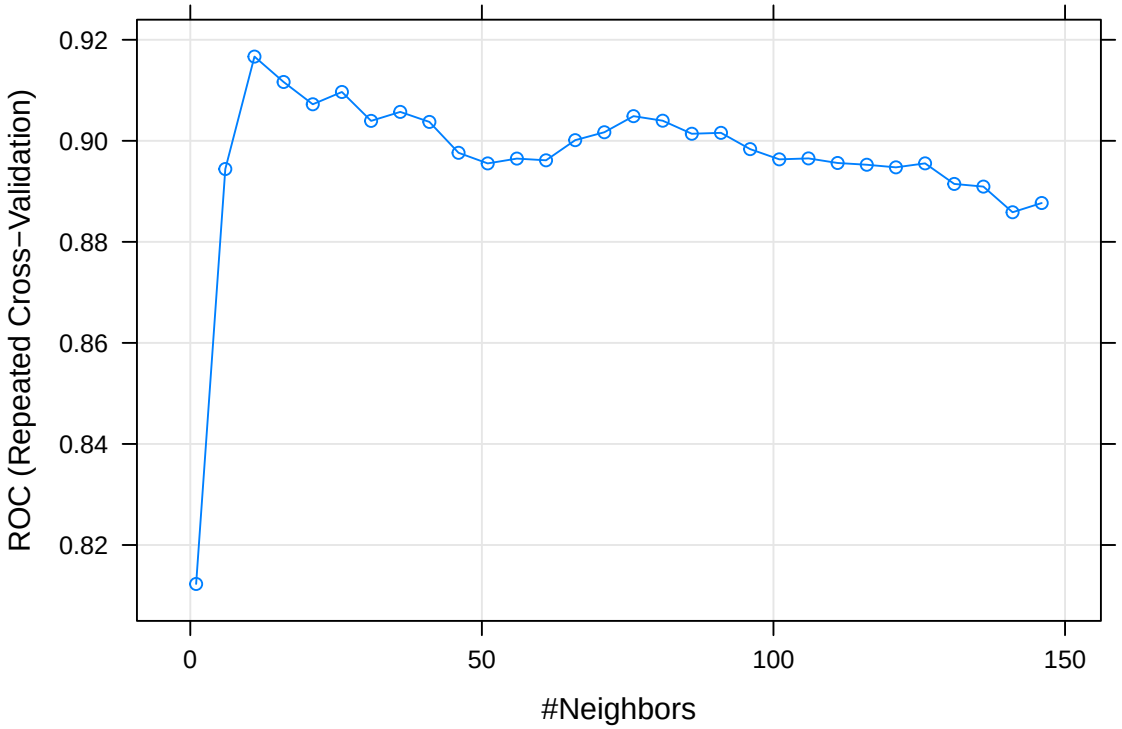
From the plot, we can see that when $\alpha = 0.2, \lambda = 0.011$, we get the best model with largest ROC. The Accuracy of the best model from GLM is 0.7733. Sensitivity is 0.7805 and Specificity is 0.7647.



From the importance plot, we can see the level of ca is the most important feature in prediction the disease status.

KNN

We also used unsupervised learning method KNN to model this problem. When using KNN method, it's important to do preProcess: center and scale the data. Then the tuning parameter in KNN method is the number of neighborhood. We choose to tune from 1 to 150. From the plot, when the number of neighborhood is 11, we get the largest ROC. The accuracy of KNN is 0.8533, sensitivity is 0.9024 and specificity : 0.7941

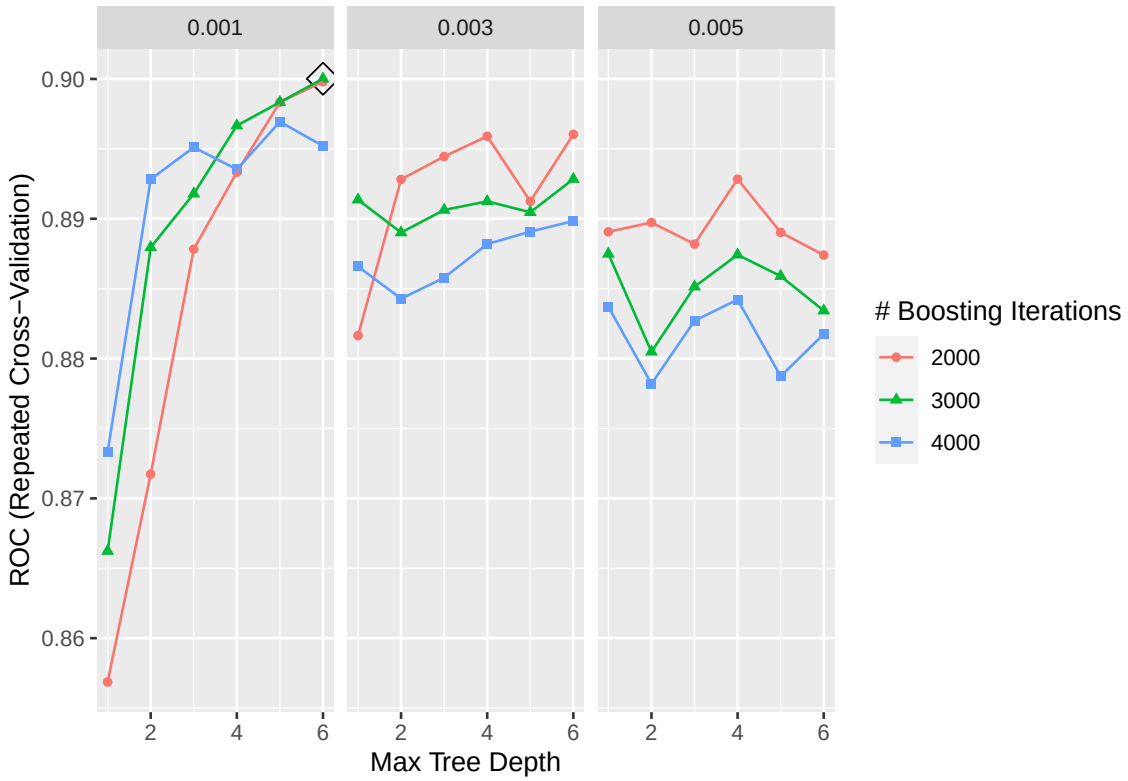


Tree

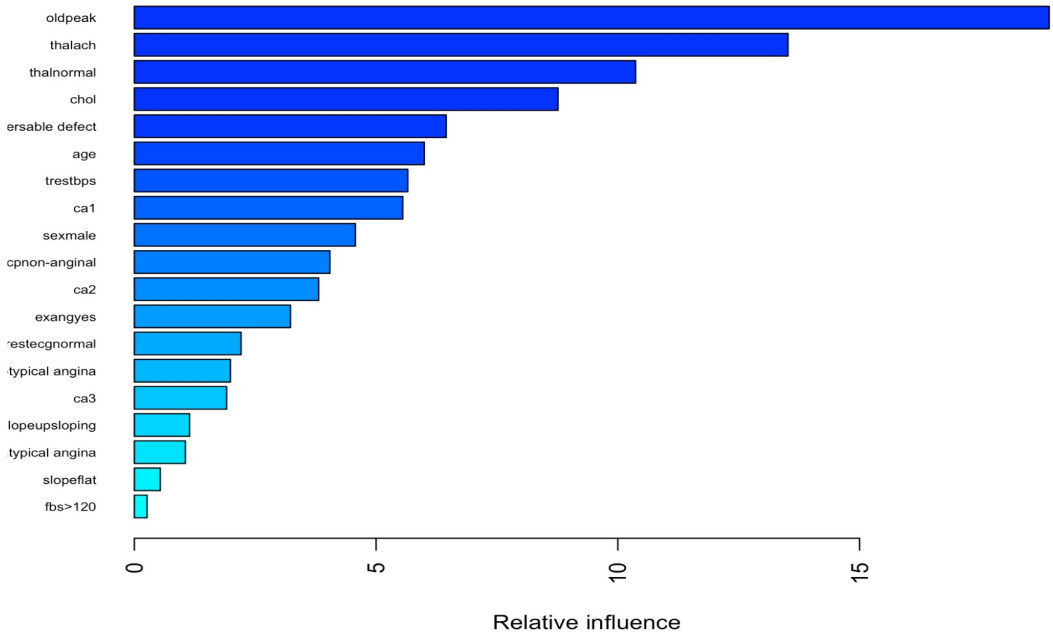
Random Forest

Boosting

In addition, we use boosting model for Heart data. Then we tune some parameters. We choose the number of trees from $(2000, 3000, 4000)$, the interaction depth from 1 to 6, and the shrinkage parameters λ from $(0.001, 0.003, 0.005)$. And we fix the minimum number of observations in the terminal nodes of the trees 1. From the plot we find that when the number of trees is 3000, the interaction depth is 6, the shrinkage equals 0.001, we get the largest ROC. The Accuracy of the best model from Boosting is 0.7867, Sensitivity is 0.8049 and Specificity is 0.7647. The test error rate is 0.2133.



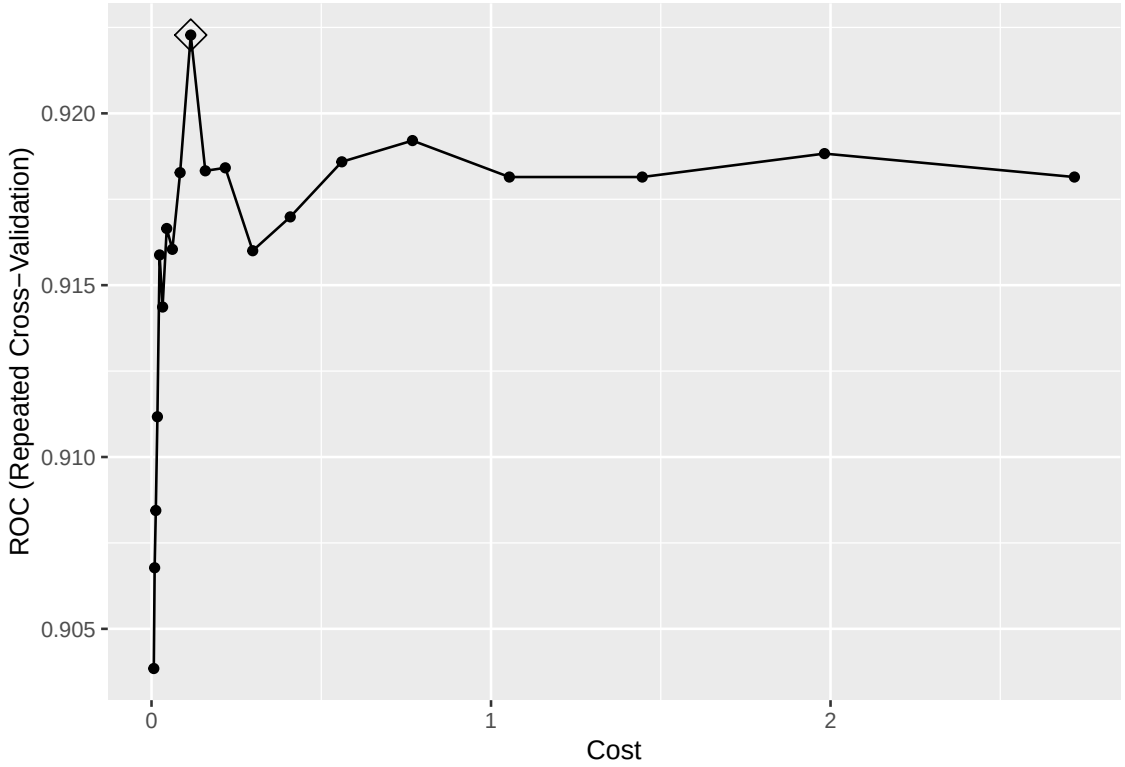
The following plot shows the variance importance. Here we can find that oldpeak is the most important variable while the fasting blood sugar is the least important variable.



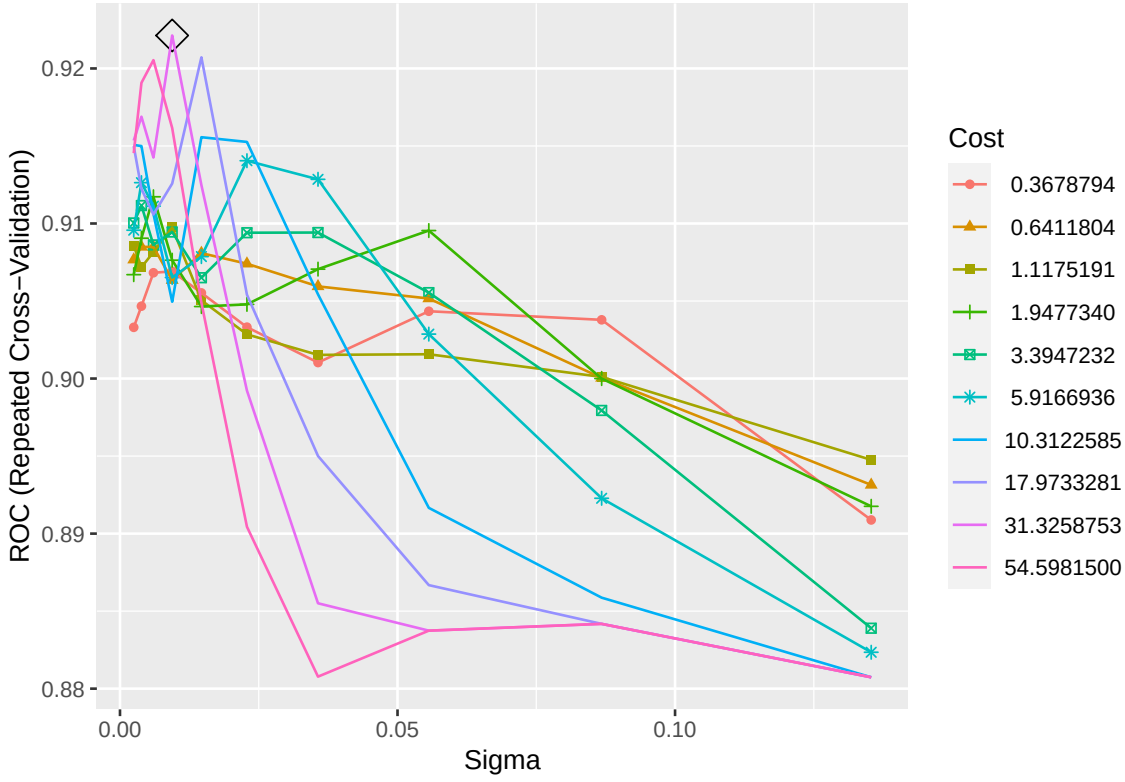
SVM

Finally, we use support vector machine (both linear kernel and radical kernel model) to train the data.

For linear kernel model, the tuning parameter cost C is chose from e^{-5} to e. When $C = 0.115568$, we gain the largest ROC.



For radical kernel model, cost C is chose from e^{-1} to e^{4}, γ(sigma in the plot) is chose from e^{-6} to e^{-2}. When $C = 31.32588, \gamma = 0.0094$, we gain the largest ROC.



The following table shows some results for two SVM models.

	Accuracy	Kappa	Sensitivity	Specificity	test_error_rate
svmr	0.760	0.518	0.765	0.756	0.240
svml	0.813	0.623	0.794	0.829	0.187

Discuss the training/test performance if you have a test data set. Which variables play important roles in predicting the response? What are the limitations of the models you used (if there are any)? Are the models flexible enough to capture the underlying truth?

Conclusions

Model comparison

Explain/visualize the final model you select.

What were your findings? Are they what you expect? What insights into the data can you make?