Statistics and Machine Learning
Homework 3

中研院統計所 林松江

Data discription：

Title：Pima Indians Diabetes Dataset

Number of Instances: 768 (positive：268 negative：500)

For Each Attribute: (all numeric-valued)：

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Methods：

- logistic regression model
- Fisher linear discriminant
- linear regression

Testing data set：Sampling 1/10 of negative examples and 1/10 from positive examples。

Training data set：All the rest(9/10) will be use for “training/modeling。

- **Logistic Regression（LR）**

Model：

glm(formula = Label ~ ., family = binomial(link = "logit"), data = trn.data)

Deviance Residuals：

Min	1Q	Median	3Q	Max
-2.5100	-0.7253	-0.4193	0.7288	2.8332

Coefficients：

	Estimate	Std. Error	z value	Pr(>|z|)	Signif. codes
(Intercept)	-8.4706	0.7477	-11.3280	< 2e-16	***
pregnant	0.1301	0.0336	3.8680	0.0001	***
glucose	0.0334	0.0038	8.7580	< 2e-16	***
pressure	-0.0112	0.0055	-2.0550	0.0399	*
triceps	0.0009	0.0074	0.1280	0.8980	
insulin	-0.0011	0.0009	-1.1620	0.2451	
mass	0.0933	0.0161	5.8020	0.0000	***
pedigree	0.8490	0.3096	2.7420	0.0061	**
age	0.0155	0.0097	1.5940	0.1109	

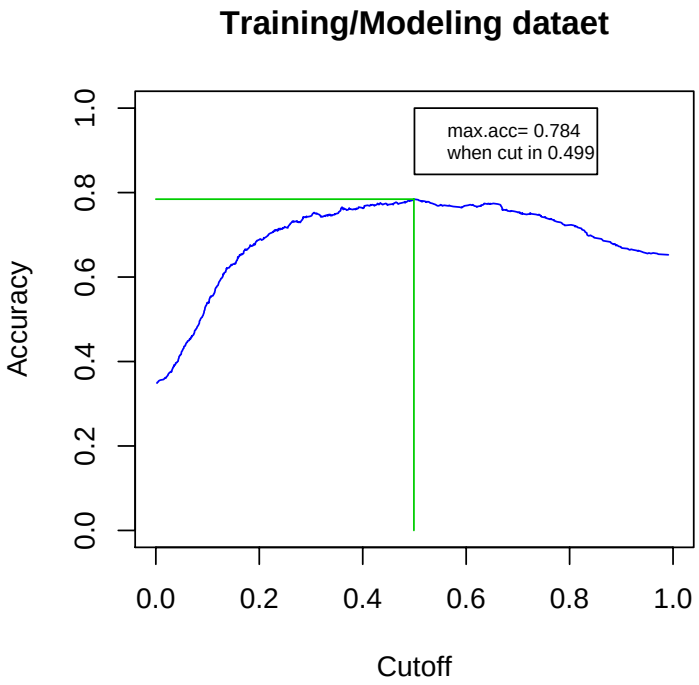
. ：0.5 ＜ P value ≦ 0.1

*：0.01 ＜ P value ≦ 0.05

**：0.001＜ P value ≦ 0.01

***：P value ≦ 0.001

Choosing Cutting point from Training/Modeling dataset：choose **0.499** as cutting point



By choosing 0.499 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

True Label \ Prediction	Negative	Positive
Negative	401	49
Positive	100	141

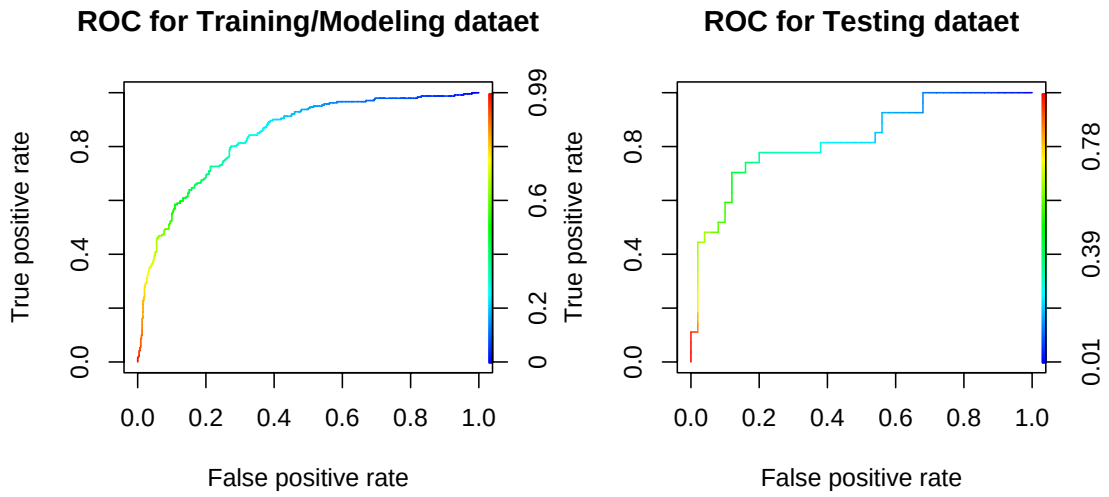
Confusion matrix for testing data set：

True Label \ Prediction	Negative	Positive
Negative	44	6
Positive	11	16

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.784	0.109	0.415
Testing	0.779	0.120	0.407

ROC Curve：



We apply variable selection method to construct a model again. (selected variables: pregnant, glucose, pressure, mass, pedigree).

✧ LR_Variable Selection-1：

Choosing 5 variables – pregnant、glucose、pressure、mass、pedigree

Model：

glm(formula = Label ~ ., family = binomial(link = "logit")
, data = trn.data[,c(1,2,3,4,7,8)])

Deviance Residuals：

Min	1Q	Median	3Q	Max
-2.7043	-0.7357	-0.4209	0.7183	2.8671

Coefficients：

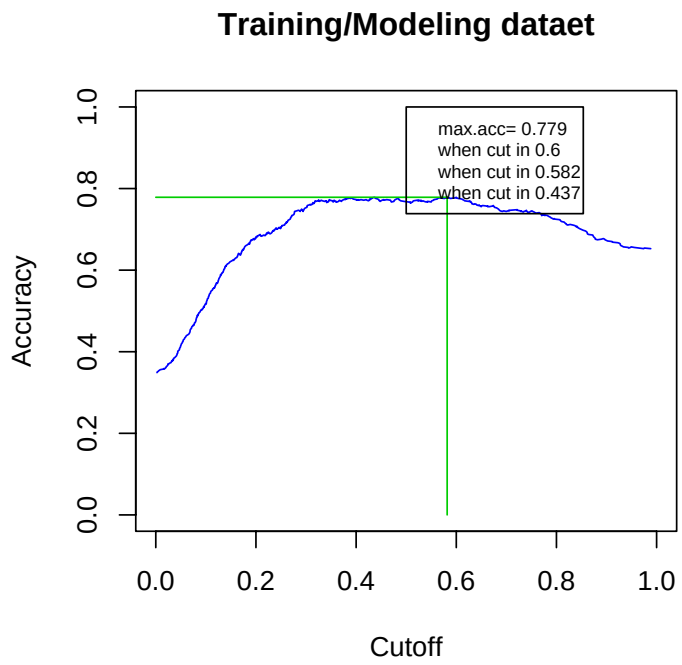
	Estimate	Std. Error	z value	Pr(>|z|)	Signif. codes
(Intercept)	-8.0350	0.7065	-11.3720	< 2e-16	***
pregnant	0.1607	0.0294	5.4750	0.0000	***
glucose	0.0331	0.0035	9.4550	< 2e-16	***
pressure	-0.0098	0.0052	-1.8730	0.0610	.
mass	0.0887	0.0150	5.9020	0.0000	***
pedigree	0.8317	0.3038	2.7370	0.0062	**

. ：0.5 ＜ P value ≦ 0.1

* ：0.01 ＜ P value ≦ 0.05

** ：0.001＜ P value ≦ 0.01

*** ：P value ≦ 0.001



How to Choose Cutting point from Training/Modeling dataset：

1. We will obtain training/modeling max. accuracy=0.779 when choosing one of 0.437、0.582、0.6 as cutting point. If we choose median of them, **0.582**, as cutting point. We can obtain results as follows:
   By choosing 0.582 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

Prediction / True Label	Negative	Positive
Negative	415	35
Positive	118	123

Confusion matrix for testing data set：

Prediction / True Label	Negative	Positive
Negative	47	3
Positive	13	14

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.779	0.078	0.490
Testing	0.792	0.060	0.481

2. From previous Training/Modeling cutoff figure, we can find the cutting interval (0.3, 0.6) will obtain similar Training/Modeling accuracy. Thus, considering the prior prob.: $P(Y_1) = 0.3488$, $P(Y_0) = 0.6512$. We set the cutting point = 0.3488 for posterior prob.

   By setting 0.3488 as cutting point, we can obtain the results as follows：more balance for False Positive Rate and False Negative Rate.

Confusion matrix for training/modeling data set：

True Label \ Prediction	Negative	Positive
Negative	399	95
Positive	64	177

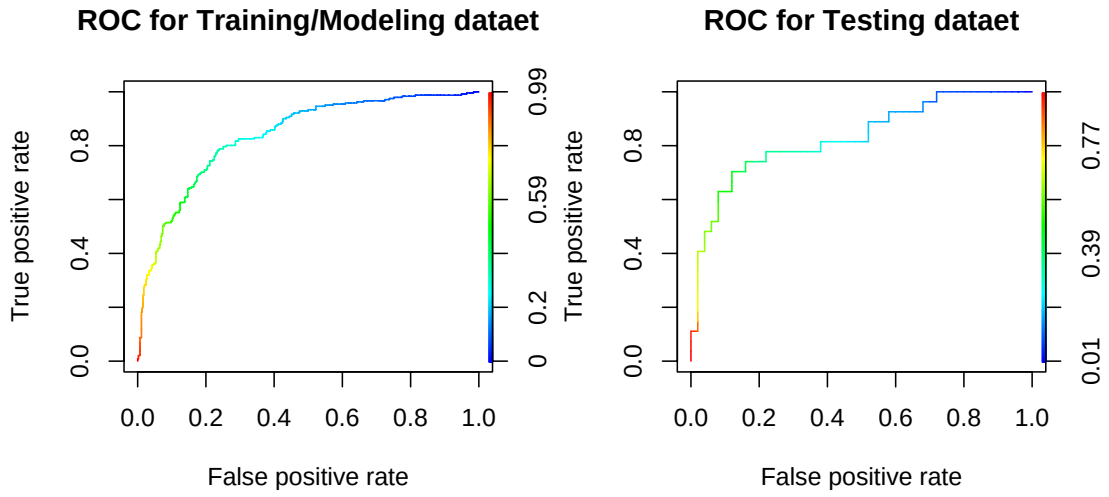
Confusion matrix for testing data set：

True Label \ Prediction	Negative	Positive
Negative	39	11
Positive	7	20

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.770	0.211	0.266
Testing	0.766	0.220	0.259

ROC Curve：



✧ LR_Variable Selection-2：

Choosing 3 variables – pregnant、glucose、mass

Model：

```
glm(formula = Label ~ ., family = binomial(link = "logit")
, data = trn.data[,c(1,2,3,7)])
```

Deviance Residuals：

Min	1Q	Median	3Q	Max
-2.1741	-0.7307	-0.4262	0.7524	2.8089

Coefficients：

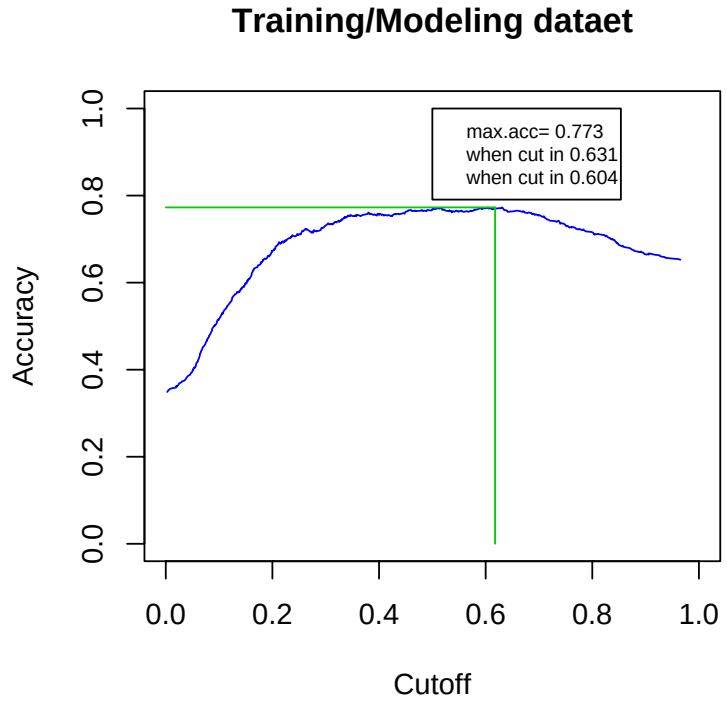
	Estimate	Std. Error	z value	Pr(>|z|)	Signif. codes
(Intercept)	-8.1520	0.6706	-12.1570	< 2e-16	***
pregnant	0.1447	0.0283	5.1140	0.0000	***
glucose	0.0329	0.0034	9.6000	< 2e-16	***
mass	0.0861	0.0145	5.9220	0.0000	***

. ：0.5 ＜ P value ≦ 0.1

* ：0.01 ＜ P value ≦ 0.05

** ：0.001＜ P value ≦ 0.01

*** ：P value ≦ 0.001



How to Choose Cutting point from Training/Modeling dataset：

1. We will obtain training/modeling max. accuracy=0.773 when choosing one of 0.631、0.604 as cutting point. If we choose mean of them, **0.618**, as cutting point. We can obtain results as follows:
   By choosing 0.618 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

Prediction / True Label	Negative	Positive
Negative	423	27
Positive	131	110

Confusion matrix for testing data set：

Prediction / True Label	Negative	Positive
Negative	47	3
Positive	16	11

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.771	0.060	0.544
Testing	0.753	0.060	0.593

2. Considering the prior probability for positive and negative distribution, we set 0.3488 as cutting point to obtain the results as follows：more balance for False Positive Rate and False Negative Rate.

Confusion matrix for training/modeling data set：

Prediction / True Label	Negative	Positive
Negative	350	100
Positive	71	170

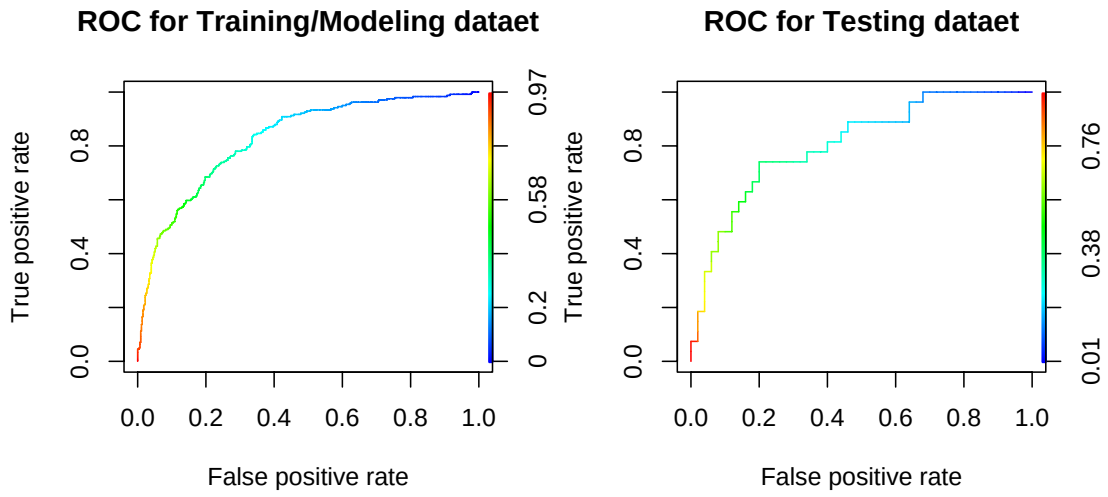
Confusion matrix for testing data set：

Prediction / True Label	Negative	Positive
Negative	37	13
Positive	7	20

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.753	0.222	0.295
Testing	0.740	0.260	0.259

ROC Curve：



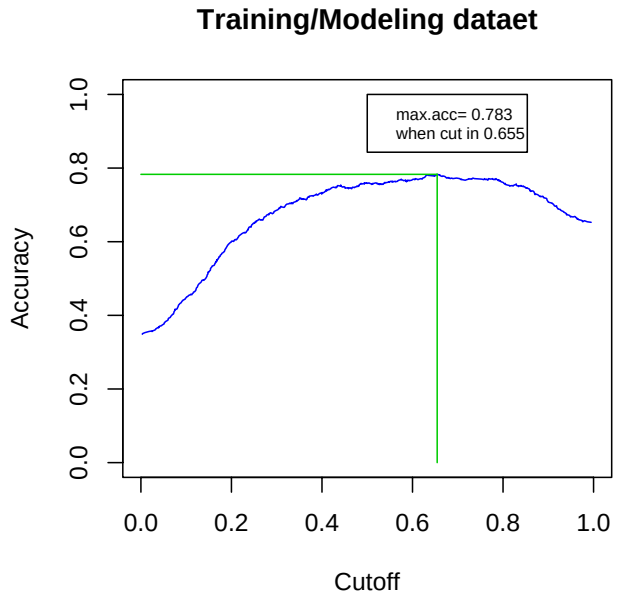
- **Fisher linear discriminant（FLD）**

1. If we don't consider the positive set and negative set distribution in Training/Modeling. Let Negative : Positive = 1 : 1

Model：

lda(Label ~ ., data=trn.data, prior=c(0.5,0.5))

Choosing Cutting point from Training/Modeling dataset：choose **0.655** as cutting point



By choosing 0.655 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

Prediction / True Label	Negative	Positive
Negative	401	49
Positive	101	140

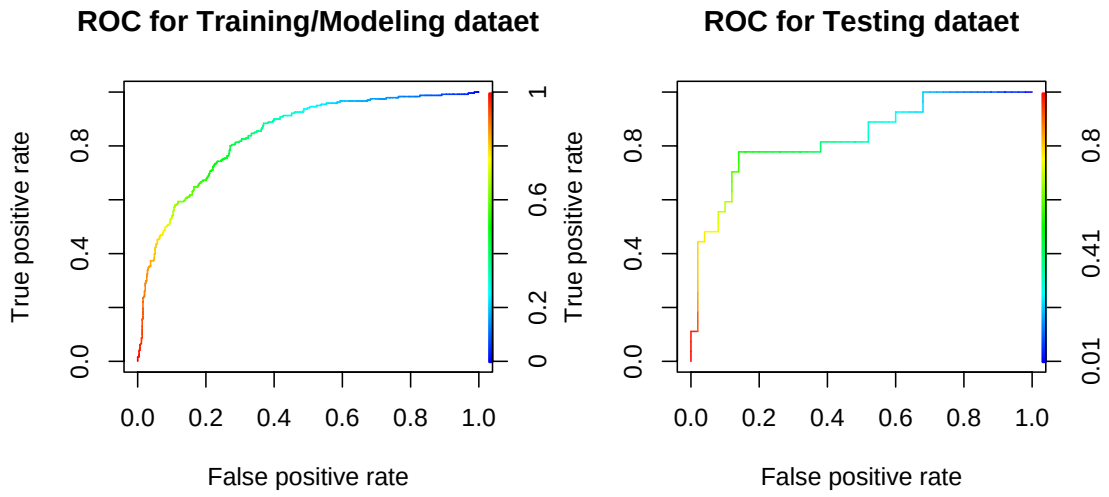
Confusion matrix for testing data set：

Prediction / True Label	Negative	Positive
Negative	44	6
Positive	11	16

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.783	0.109	0.419
Testing	0.779	0.120	0.407

ROC Curve：

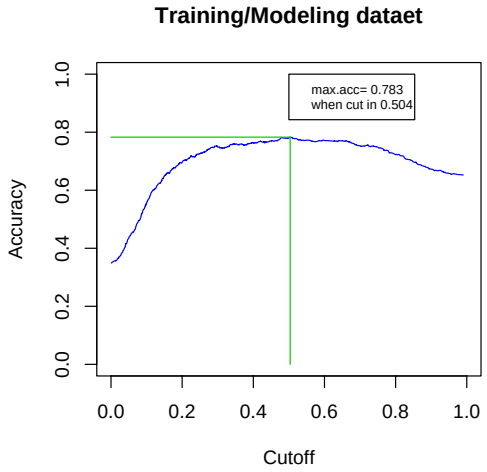


2. Considering the prior probabilities of class membership for Training/Modeling data set is Negative : Positive = 0.6512301 : 0.3487699.

Model：

lda(Label ~ ., data=trn.data, prior=c(sum(trny==0),sum(trny==1))/length(trny))

Choosing Cutting point from Training/Modeling dataset：choose **0.504** as cutting point



By choosing 0.504 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

True Label \ Prediction	Negative	Positive
Negative	401	49
Positive	101	140

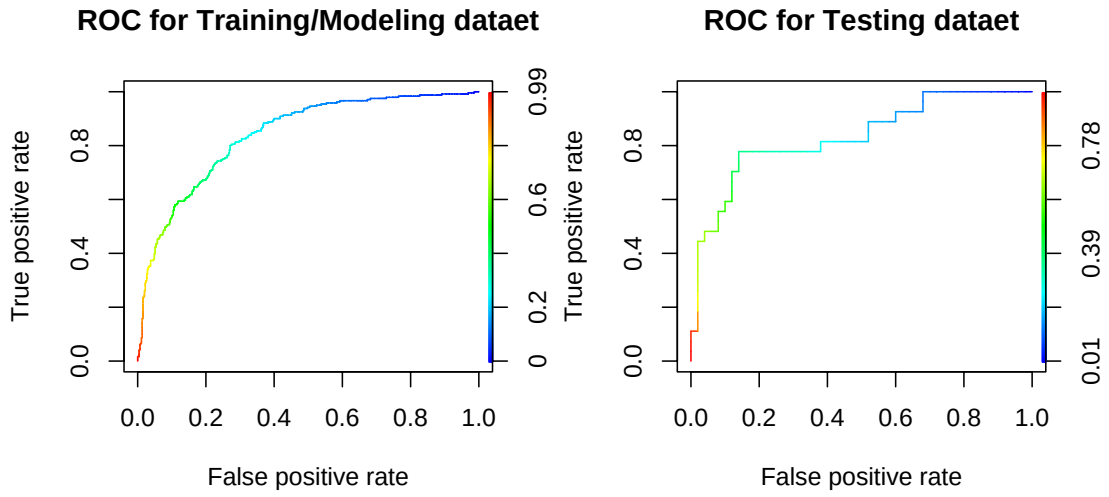
Confusion matrix for testing data set：

True Label \ Prediction	Negative	Positive
Negative	44	6
Positive	11	16

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.783	0.109	0.419
Testing	0.779	0.120	0.407

ROC Curve：



➢ **linear regression（LR）**

Model：

lm(formula = Label ~ ., data = trn.data)

Deviance Residuals：

Min	1Q	Median	3Q	Max
-0.99351	-0.29148	-0.09917	0.31735	1.19440

Coefficients：

	Estimate	Std. Error	z value	Pr(>|z|)	Signif. codes
(Intercept)	-0.8786	0.0902	-9.7370	< 2e-16	***
pregnant	0.0215	0.0054	3.9940	0.0001	***
glucose	0.0056	0.0005	10.5300	< 2e-16	***
pressure	-0.0019	0.0009	-2.2050	0.0278	*
triceps	0.0002	0.0012	0.2100	0.8341	
insulin	-0.0002	0.0002	-1.0580	0.2904	
mass	0.0140	0.0022	6.2110	0.0000	***
pedigree	0.1336	0.0472	2.8310	0.0048	**
age	0.0028	0.0016	1.7040	0.0889	.

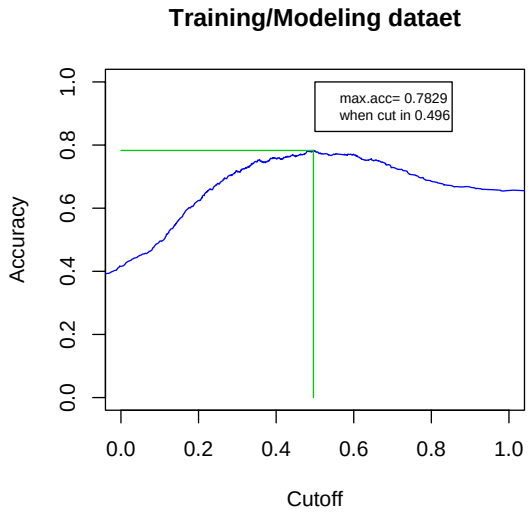
. ：0.5 ＜ P value ≦ 0.1

* ：0.01 ＜ P value ≦ 0.05

** ：0.001＜ P value ≦ 0.01

*** ：P value ≦ 0.001

Choosing Cutting point from Training/Modeling dataset：choose **0.496** as cutting point



By choosing 0.496 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

True Label \ Prediction	Negative	Positive
Negative	401	49
Positive	101	140

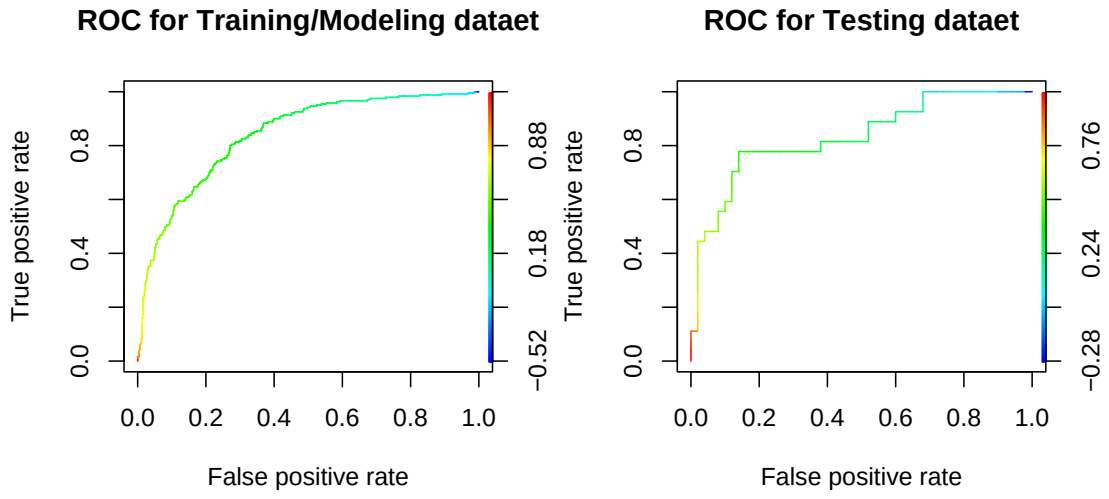
Confusion matrix for testing data set：

True Label \ Prediction	Negative	Positive
Negative	44	6
Positive	11	16

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.783	0.109	0.419
Testing	0.779	0.120	0.407

ROC Curve：



✧ LM_Variable Selection-1：

Choosing 5 variables – pregnant、glucose、pressure、mass、pedigree

Model：

lm(formula = Label ~ ., data = trn.data[,c(1,2,3,4,7,8)])

Deviance Residuals：

Min	1Q	Median	3Q	Max
-1.0908	-0.2931	-0.1017	0.3176	1.2156

Coefficients：

	Estimate	Std. Error	z value	Pr(>|z|)	Signif. codes
(Intercept)	-0.8218	0.0861	-9.5450	< 2e-16	***
pregnant	0.0269	0.0046	5.8420	0.0000	***
glucose	0.0056	0.0005	11.4580	< 2e-16	***
pressure	-0.0016	0.0008	-1.9590	0.0505	.
mass	0.0136	0.0021	6.4530	0.0000	***
pedigree	0.1325	0.0464	2.8540	0.0045	**

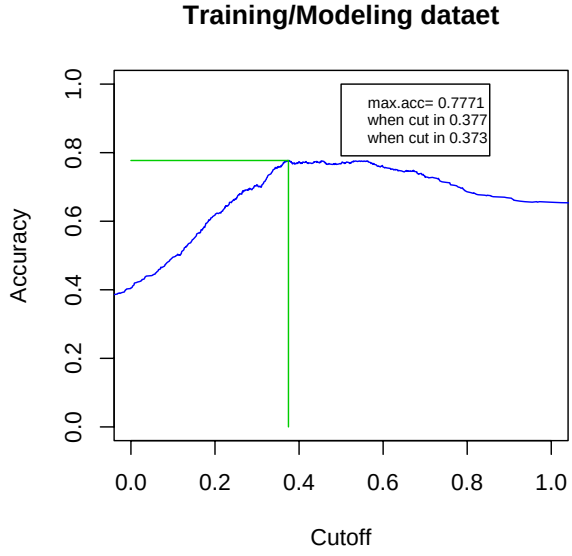
. ：0.5 ＜ P value ≦ 0.1

* ：0.01 ＜ P value ≦ 0.05

** ：0.001＜ P value ≦ 0.01

*** ：P value ≦ 0.001

Choosing Cutting point from Training/Modeling dataset：We will obtain training/modeling max. accuracy=0.777 when choosing one of 0.373、0.377 as cutting point. Finally, we decide to choose mean of them, **0.375**, as cutting point.



By choosing 0.375 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

Prediction / True Label	Negative	Positive
Negative	349	101
Positive	55	186

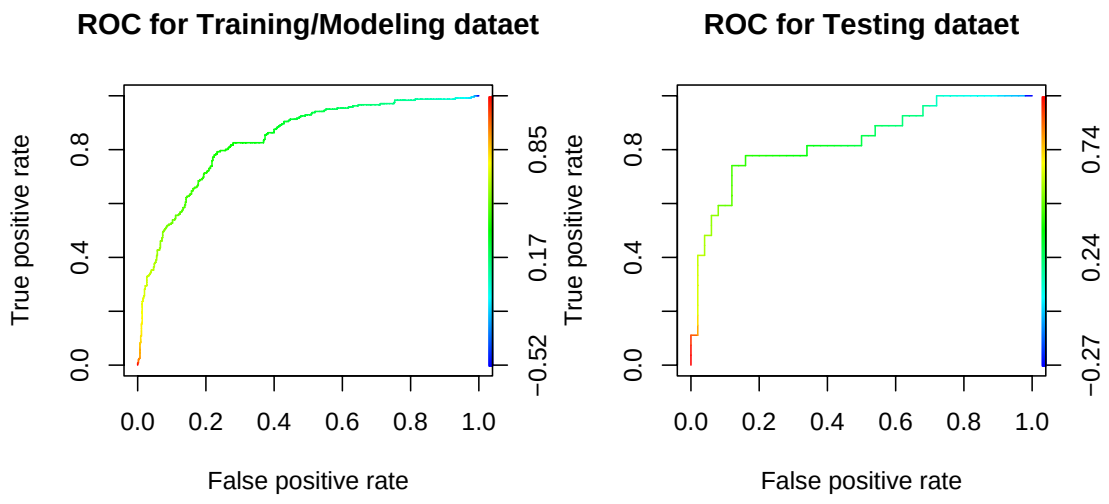
Confusion matrix for testing data set：

Prediction / True Label	Negative	Positive
Negative	37	13
Positive	6	21

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.774	0.224	0.228
Testing	0.753	0.260	0.222

ROC Curve：



✧ LM_Variable Selection-2：

Choosing 3 variables – pregnant、glucose、mass

Model：

lm(formula = Label ~ ., data = trn.data[,c(1,2,3,7)])

Deviance Residuals：

Min	1Q	Median	3Q	Max
-0.87643	-0.29007	-0.09268	0.32453	1.19923

Coefficients：

	Estimate	Std. Error	z value	Pr(>|z|)	Signif. codes
(Intercept)	-0.8653	0.0797	-10.8600	< 2e-16	***
pregnant	0.0247	0.0046	5.4000	0.0000	***
glucose	0.0057	0.0005	11.6910	< 2e-16	***
mass	0.0133	0.0020	6.5440	0.0000	***

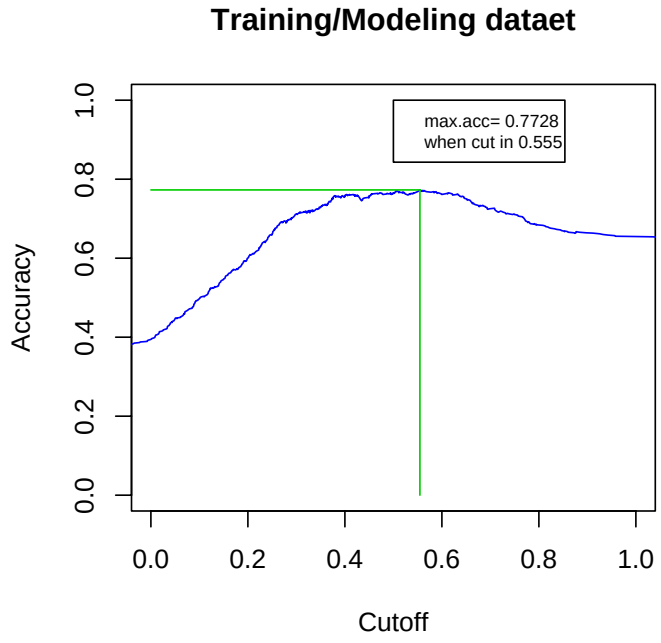
. ：0.5 ＜ P value ≦ 0.1

* ：0.01 ＜ P value ≦ 0.05

** ：0.001＜ P value ≦ 0.01

*** ：P value ≦ 0.001

Choosing Cutting point from Training/Modeling dataset：We will obtain training/modeling max. accuracy=0.773 when choosing **0.555** as cutting point.



By choosing 0.555 as the cutting point, we can obtain the results as follows：

Confusion matrix for training/modeling data set：

True Label \ Prediction	Negative	Positive
Negative	416	34
Positive	123	118

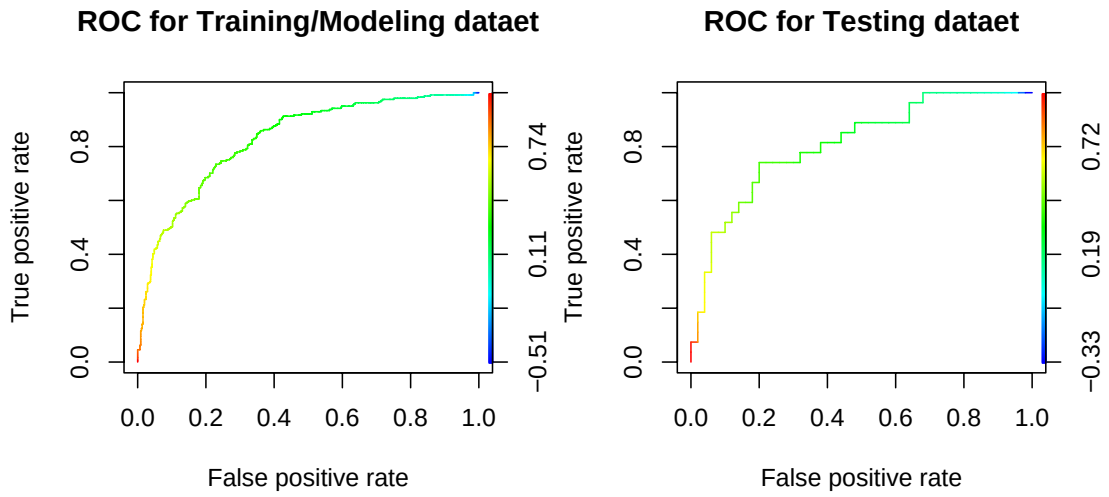
Confusion matrix for testing data set：

True Label \ Prediction	Negative	Positive
Negative	46	4
Positive	14	13

Summary the results of prediction：

	Accuracy	False Positive Rate	False Negative Rate
Training/Modeling	0.773	0.076	0.510
Testing	0.766	0.080	0.519

ROC Curve：



- Summary **Logistic Regression**、**Fisher Linear Discriminant**、**Linear Regression** three models with all variables：

	Logistic Regression		**Fisher Linear Discriminant**		**Linear Regression**	
	Training	Testing	Training	Testing	Training	Testing
False Positive Rate	0.109	0.120	0.109	0.120	0.109	0.120
False Negative Rate	0.415	0.407	0.419	0.407	0.419	0.407
Accuracy	0.784	0.779	0.783	0.779	0.783	0.779