Logistic regression for analyzing binary data in heart studies

南方科技大学
张昀,龙占廷,唐晔,孟楚寒,赵达
指导老师：陈欣,伍润雄

2019年5月14日

1.问题重述与初步分析

1.1样本数据

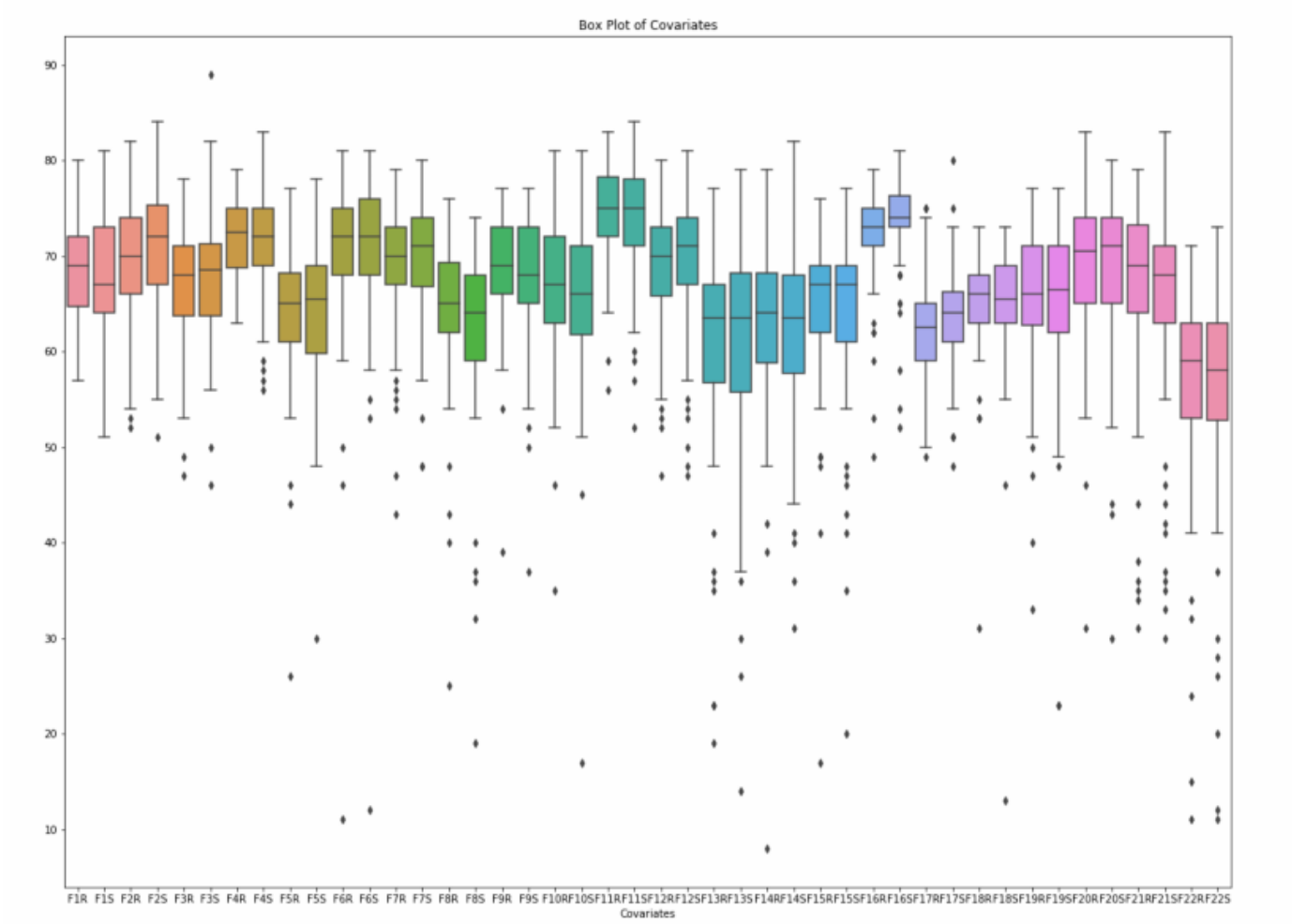
在本文中,我们分析了 SPECTF heart data 心脏数据,该数据集分为训练集和预测集两部分.数据集描述了心脏单质子发射计算机断层扫描（SPECT）图像的诊断. 每名患者分为两类：正常（诊断结果为1）和异常（诊断结果为0）,即响应变量. 对267个SPECT图像集（患者）的数据库进行处理,提取原始SPECT图像的特征. 因此,为每个患者创建了44个连续特征模式,即44个连续的自变量.

训练集: 45个变量, 80个观测, 第一列(OVERALL_DIAGNOSIS)为响应变量. 在响应变量的80个观测中有40个为0, 40个为1.

测试集: 186个观测, 第一列为响应变量的真实值. 在响应变量的186个观测中, 有15个为0, 有171个为1.

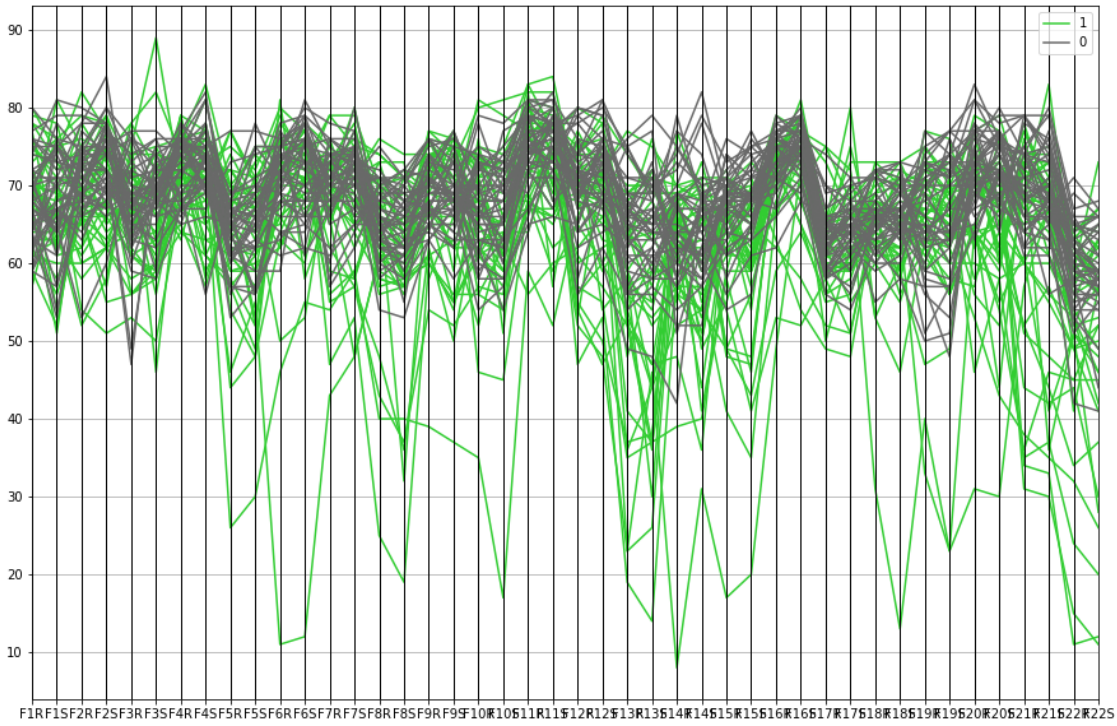
1.2数据可视化

- **箱图(boxplot):** 直观描述了44个自变量的数据特征,从图中我们可以看出44个协变量的取值范围是相近的,他们的度量单位不会影响模型估计,所以我们没有对原始数据进行标准化处理.



- **平行坐标系(parallel-coordinates):** 平行坐标也是一种多维可视化技术.它可以看到数据中的类别以及从视觉上估计其他的统计量.使用平行坐标时,每个点用线段联接.每个垂直的线代表一个属性.一组联接的线段表示一个数据点.可能是一类的数据点会更加接近.

图中,绿色线代表真实结果为1(Normal),灰色线代表真实结果为0 (Abnormal).



1.3建模目标

- 使用Losgitc 模型回归心脏数据；
- 分析全模型效果并对数据进行变量选择；
- 对模型结果进行讨论并选出最好的模型.

2.Logistic回归与全模型

2.1 Logistic回归模型

广义线性回归是探索“响应变量的期望”与“自变量”的关系,以实现对非线性关系的某种拟合.这里面涉及到一个“连接函数”和一个“误差函数”,“响应变量的期望”经过连接函数作用后,与“自变量”存在线性关系.选取不同的“连接函数”与“误差函数”可以构造不同的广义回归模型.当误差函数取“二项分布”而连接函数取“logit函数”时,就是常见的“logistic回归模型”,在0-1响应的问题中得到了大量的应用.

Logistic回归主要通过构造一个重要的指标：发生比来判定因变量的类别.在这里我们引入概率的概念,把事件发生定义为Y=1,事件未发生定义为Y=0,那么事件发生的概率为p,事件未发生的概率为1-p,把p看成x的线性函数.

回归中,最常用的估计是最小二乘估计,因为使得p在[0,1]之间变换,最小二乘估计不太合适,有木有一种估计法能让p在趋近与0和1的时候变换缓慢一些（不敏感）,这种变换是我们想要的,于是引入Logit变换,对p/(1-p)也就是发生与不发生的比值取对数,也称对数差异比.经过变换后,p对x就不是线性关系了. logistic回归的公式可以表示为：

$\mathrm{P}=\frac{\exp\left(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p\right)}{1+\exp\left(\beta_0+\beta_1 X_1+\cdots+\beta_p X_p\right)}$ 其中P是响应变量取1的概率,在0-1变量的情形中,这个概率就等于响应变量的期望.

2.2 模型评价

本文共涉及到11个模型,我们写了一个评估函数(见附录四),这个函数输出 混淆矩阵,准确度(accuracy),精确率(precison) ,召回率(recall),F1,AIC,BIC这7个指标.

下面我们分别对前五个指标和ROC进行分别说明:

- 混淆矩阵

真实类别	预测结果	
	类别1（正例）	类别2（反例）
类别1（正例）	真正例(True Positive) TP	假反例(False Negatibe) FN
类别2（反例）	假正例(False Positive)FP	真反例(True Negatibe) TN

- 准确率（Accuracy）表示正确分类的测试实例的个数占测试实例总数的比例,计算公式为:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

- 精确率（Precision）,也叫查准率,表示正确分类的正例个数占分类为正例的实例个数的比例,计算公式为:

$$Precision = \frac{TP}{TP+FP}$$

- 召回率（Recall）,也叫查全率,表示正确分类的正例个数占实际正例个数的比例,计算公式为:

$$Recall = \mathbf{R}\,\mathrm{ecall} = \frac{TP}{TP+FN}$$

- F1是基于召回率（Recall）与精确率（Precision）的调和平均,即将召回率和精确率综合起来评价,计算公式为:

$$F1 = \frac{2*\mathrm{Recall}*\mathrm{Precision}}{\mathrm{Recall}+\ \mathrm{Precision}}$$

- ROC曲线,用来度量分类中的非均衡性的工具.TPR（True Positive Rate）表示在所有实际为阳性的样本中,被正确地判断为阳性的比率,即：TPR=TP/(TP+FN)； FPR（ False Positive Rate）表示在所有实际为阴性的样本中,被错误地判断为阳性的比率,即：FPR=FP/(FP+TN).ROC曲线是以FPR作为X轴,TPR作为Y轴.FPR越大表明预测正类中实际负类越多,TPR越大,预测正类中实际正类越多.

2.3 全模型 (Full model)

在全模型中,我们对把44个变量纳入考量,进行回归.但是,从模型拟合的结果来看,只有几个系数是显著的.因为变量个数太多，变量之间关系很强的共线性;从模型预测角度来看,预测的准确度也很差,所以我们必须进行变量选择.

```
##     predict
## True   0   1
##    0   6   9
##    1  57 114
##       recall    precison    accuracy          F1         AIC         BIC
##    0.4000000   0.0952381   0.6451613   0.1538462  90.0000000 197.1911986
```

3.变量选择

我们在每次用不同的方法进行变量选择后，都进行了相应的stepwise 分析,每次stepwise处理后，预测结果都得到了相应的提升.基于患者心脏疾病预测的建模目的,我们保留了stepwise后的模型.

3.1 Stepwise

为了提升模型效果以及简化模型, 我们采用 stepwise 方法, 基于 AIC 进行变量筛选.

最终筛选出的变量为:

```
## OVERALL_DIAGNOSIS ~ F2S + F3S + F4R + F4S + F6R + F7R + F9R +
##     F11R + F11S + F13R + F14S + F16R + F17R + F18R + F18S + F19R +
##     F19S + F20R + F21R + F22R
```

筛选出的模型对测试集进行预测的效果为:

```
##     predict
## True   0   1
##    0   7   8
##    1  39 132
##     recall   precison   accuracy         F1        AIC        BIC
##  0.4666667  0.1521739  0.7473118  0.2295082 42.0000000 92.0225593
```

3.2 PCA

我们也尝试使用主成分分析法对协变量进行降维.

采用前11个主成分（累计贡献率为82.82%）作为协变量进行 Logistic 回归,所得的模型对测试集的预测效果为:

```
##     predict
## True   0   1
##    0   9   6
##    1  44 127
##      recall    precison    accuracy          F1         AIC         BIC
##   0.6000000   0.1698113   0.7311828   0.2647059  79.7146879 108.2990075
```

3.3 LASSO

LASSO试图通过惩罚模型的复杂性并将参数调整为0来解决这些问题. Lasso-Logistic回归模型是指在求解logistics回归的参数估计值时加入对参数的惩罚项以实现对变量的选择和参数估计. Lasso-Logistic回归模型中的参数估计可以表示为: $\max_{\beta_0,\beta}\left\{\sum_{i=1}^{n}\left[y_i\left(\beta_0+\beta^T x_i\right)-\ln\left(1+e^{\beta_0+\beta^T x_i}\right)\right]-\lambda\sum_{j=1}^{p}|\beta_j|\right\}$ 其中,β为p为列向量.

使用LASSO 回归, 所得模型对测试集的预测效果为:

```
##     predict
## True   0   1
##    0   8   7
##    1  38 133
##     recall   precison   accuracy         F1        AIC        BIC
##  0.5333333  0.1739130  0.7580645  0.2622951 60.1119564 91.0783027
```

```
## train$OVERALL_DIAGNOSIS ~ .
```

3.4 LASSO + Stepwise

对使用 LASSO 回归得到的模型利用 stepwise 方法进行变量选择, 可以得到一个效果更好的模型

```
##     predict
## True   0   1
##    0   8   7
##    1  36 135
##     recall   precison   accuracy         F1        AIC        BIC
##  0.5333333  0.1818182  0.7688172  0.2711864 53.0447669 74.4830066
```

```
## train$OVERALL_DIAGNOSIS ~ F2S + F4S + F11S + F14S + F15S + F17S +
##     F20S + F21R
```

3.5 Adaptive Lasso

一个改进Lasso的方法,称为Adaptive-Lasso方法,其参数估计定义为(对惩罚项进行了修改)

$\widehat{\beta}^{*(n)}=\arg\min\left\|Y-\sum_{j=1}^{p}X_j\beta_j\right\|^2+\lambda_n\sum_{j=1}^{p}\widehat{w}_j\left|\beta_j\right|$ 其中,$\widehat{w}_j=1/\left|\widehat{\beta}_j\right|^{\gamma}(\gamma>0),j=1,2,\ldots,p.$

$\widehat{\beta}=\left(\widehat{\beta}_1,\widehat{\beta}_2,\ldots,\widehat{\beta}_p\right)^{\top}$ 为普通最小二乘法所得系数估计值.记权重向量

$$\widehat{W}=(\widehat{w}_1,\widehat{w}_2,\ldots,\widehat{w}_p)^{\top}=\left(\frac{1}{\left|\widehat{\beta}_1\right|^{\gamma}},\frac{1}{\left|\widehat{\beta}_2\right|^{\gamma}},\cdots,\frac{1}{\left|\widehat{\beta}_p\right|^{\gamma}}\right)=\frac{1}{|\widehat{\beta}|^{\gamma}}\quad(\gamma>0).$$

```
##     predict
## True   0   1
##    0   8   7
##    1  41 130
##     recall   precison   accuracy         F1        AIC        BIC
##  0.5333333  0.1632653  0.7419355  0.2500000 57.0525756 97.5470284
```

```
## OVERALL_DIAGNOSIS ~ .
```

3.6 Adaptive Lasso+Stepwise

再对3.5中的模型进行stepwise,我们选择了如下的模型:

```
##     predict
## True   0   1
##    0  10   5
##    1  30 141
##     recall   precison   accuracy         F1        AIC        BIC
##  0.6666667  0.2500000  0.8118280  0.3636364 51.8356698 78.0379628
```

```
## OVERALL_DIAGNOSIS ~ F5R + F5S + F9R + F10R + F13S + F15R + F17R +
##     F18R + F20S + F21R
```

3.7 Elastic net

弹性网络模型是一种参数估计和变量选择同步的技术,其参数估计定义如下:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{\alpha}{2}\|\beta\|^2 + (1-\alpha)\|\beta\|_1$$

```
##     predict
## True   0   1
##    0   7   8
##    1  31 140
##      recall    precison    accuracy          F1         AIC         BIC
##   0.4666667   0.1842105   0.7903226   0.2641509  58.0000000 127.0787724
```

```
## train$OVERALL_DIAGNOSIS ~ .
```

3.8 Elastic net+Stepwise

对3.7的模型进行stepwise,我们得到如下模型:

```
##     predict
## True   0   1
##    0   8   7
##    1  37 134
##     recall   precison   accuracy         F1        AIC        BIC
##  0.5333333  0.1777778  0.7634409  0.2666667 30.0000000 65.7303995
```

```
## train$OVERALL_DIAGNOSIS ~ F1R + F2R + F4S + F5S + F13R + F14R +
##     F14S + F15R + F16R + F17R + F17S + F18S + F20S + F22R
```

3.9 Generalized elastic net

广义弹性网络模型是一种改扩展了的弹性网络模型,其参数估计如下:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + ln(\alpha + (1-\alpha)\|\beta\|_1)$$

```
##     predict
## True   0   1
##    0   8   7
##    1  50 121
##     recall   precison   accuracy         F1        AIC        BIC
##  0.5333333  0.1379310  0.6935484  0.2191781 42.0000000 92.0225593
```

```
## OVERALL_DIAGNOSIS ~ .
```

3.10 Generalized elastic net+Stepwise

对3.9的模型做stepwise,我们得到如下的模型:

```
##     predict
## True   0   1
##    0   9   6
##    1  44 127
##     recall   precison   accuracy         F1        AIC        BIC
##  0.6000000  0.1698113  0.7311828  0.2647059 28.0000000 61.3483729
```

```
## OVERALL_DIAGNOSIS ~ F2S + F5R + F5S + F9R + F11R + F13S + F14R +
##     F14S + F15S + F17R + F18R + F18S + F20S
```

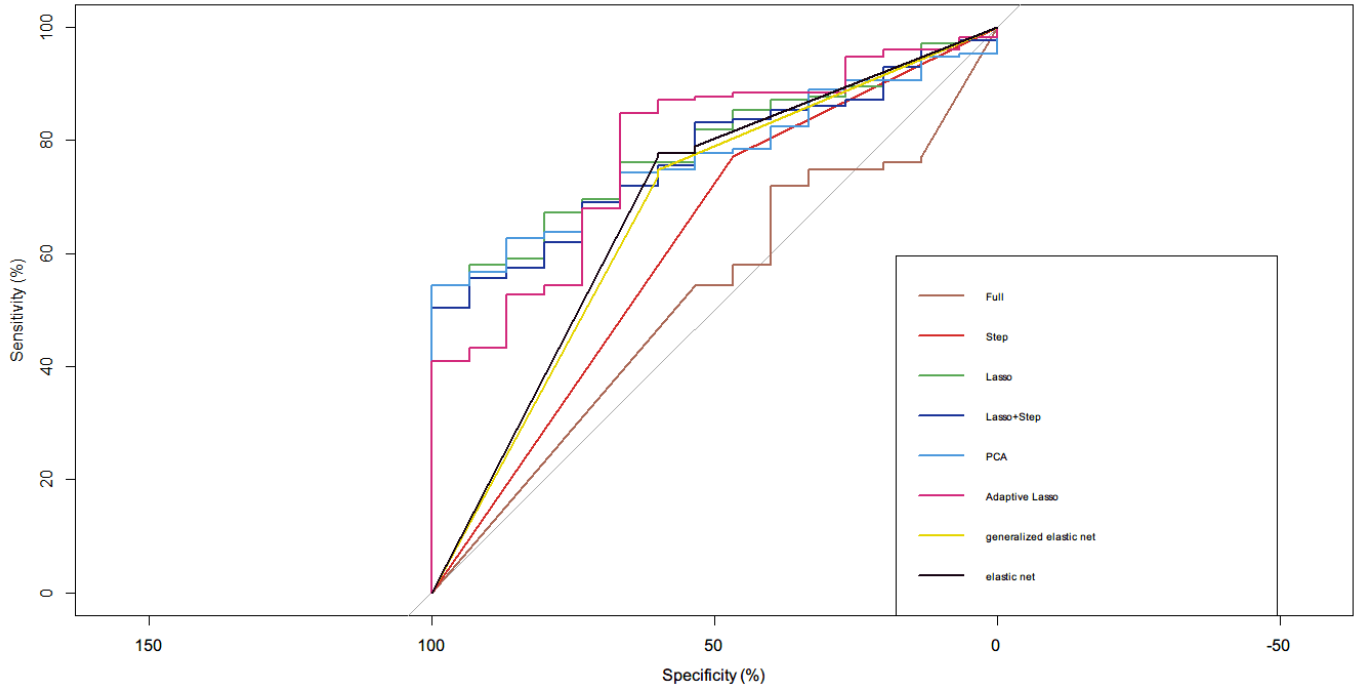
4.模型总结

我们将上述11种模型在预测集上的效果以及AIC和BIC指标总结如下表,通过对比分析,我们可以得出,预测效果最好的模型为 Adaptive Lasso+Stepwise模型,最终选出的变量为F5R ,F5S ,F9R ,F10R , F13S , F15R , F17R , F18R , F20S , F21R (共10个).从模型简洁程度来看,选出AIC和BIC最小的模型是Generalized elastic net+Stepwise模型,选出的变量为 F2S ,F5R , F5S , F9R , F11R , F13S , F14R , F14S , F15S ,F17R , F18R , F18S , F20S(共13个).这两个模型的anova分析分别见附录二和附录三.

考虑实际情况,我们建模的目的是对图像进行诊断,为了控制第一类误差,我们最后选择在预测集上效果最好的模型,即Adaptive Lasso+Stepwise模型.

Model	Recall	Precision	Accuracy	F1	AIC	BIC
Full Model	0.4000	0.0952	0.6452	0.1538	90.0000	197.1912
Full+Stepwise	0.4667	0.1522	0.7473	0.2295	42.0000	92.0226
Lasso	0.5333	0.1667	0.7473	0.2540	57.2234	81.0436
Lasso + Stepwise	0.5333	0.1818	0.7688	0.2712	53.0448	74.4830
PCA	0.6000	0.1698	0.7312	0.2647	79.7147	108.2990
Adaptiva Lasso	0.5333	0.1633	0.7419	0.2500	57.0526	97.5470
Adaptive Lasso + Stepwise	0.6667	0.2500	0.8118	0.3636	51.8357	78.0380
Generalized Elastic Net	0.5333	0.1379	0.6935	0.2192	42.0000	92.0226
Generalized Elastic Net + Stepwise	0.6000	0.1698	0.7312	0.2647	28.0000	61.3484
Elastic Net	0.5333	0.1778	0.7634	0.2667	30.0000	65.7304
Elastic Net + Stepwise	0.4667	0.1842	0.7903	0.2642	58.0000	127.0788

11个模型的ROC图如下：



附录

附录一,全模的结论分析：

```
##
## Call:
## glm(formula = OVERALL_DIAGNOSIS ~ ., family = binomial(link = logit),
##     data = train, control = list(maxit = 100))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -4.130e-06  -1.733e-07   0.000e+00   9.331e-07   3.906e-06
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.043e+03  6.128e+07       0        1
## F1R         -4.940e-01  4.386e+05       0        1
## F1S         -8.978e-01  3.826e+05       0        1
## F2R          4.064e-01  3.295e+05       0        1
## F2S         -2.969e+00  3.068e+05       0        1
## F3R         -3.230e+00  1.787e+05       0        1
## F3S          4.769e+00  4.558e+05       0        1
## F4R         -2.611e+00  9.015e+05       0        1
## F4S          2.651e+00  3.731e+05       0        1
## F5R          1.012e+01  2.731e+05       0        1
## F5S         -2.870e+00  1.064e+06       0        1
## F6R          2.920e+00  6.183e+05       0        1
## F6S         -1.638e+00  2.007e+05       0        1
## F7R          3.806e+00  1.740e+05       0        1
## F7S         -9.420e-01  9.960e+04       0        1
## F8R          1.621e+00  1.693e+05       0        1
## F8S          3.748e+00  4.727e+05       0        1
## F9R         -6.485e+00  4.612e+05       0        1
## F9S         -6.144e+00  4.412e+05       0        1
## F10R        -4.821e+00  2.823e+05       0        1
## F10S         1.986e+00  1.711e+05       0        1
## F11R         8.697e-01  5.975e+05       0        1
## F11S        -2.688e+00  3.275e+05       0        1
## F12R         6.529e-01  5.498e+05       0        1
## F12S         1.241e+00  3.248e+05       0        1
## F13R         1.749e+00  1.013e+06       0        1
## F13S        -7.158e+00  6.627e+05       0        1
## F14R         2.737e+00  4.229e+05       0        1
## F14S        -1.601e+00  2.383e+05       0        1
## F15R        -3.753e+00  5.039e+05       0        1
## F15S        -1.812e+00  1.132e+06       0        1
## F16R        -4.242e+00  1.359e+06       0        1
## F16S        -1.339e+00  6.228e+05       0        1
## F17R        -3.740e+00  1.016e+06       0        1
## F17S        -2.413e+00  1.123e+06       0        1
## F18R         3.003e+00  3.415e+05       0        1
```

```
## F18S          7.621e+00  8.588e+05       0        1
## F19R         -2.950e+00  4.256e+05       0        1
## F19S          2.359e+00  2.571e+05       0        1
## F20R          3.663e+00  7.647e+05       0        1
## F20S         -5.898e+00  9.896e+04       0        1
## F21R         -1.636e+00  2.997e+05       0        1
## F21S         -4.287e-01  1.855e+05       0        1
## F22R         -4.236e+00  7.265e+05       0        1
## F22S          6.585e+00  4.872e+05       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.1090e+02  on 79  degrees of freedom
## Residual deviance: 2.8946e-10  on 35  degrees of freedom
## AIC: 90
##
## Number of Fisher Scoring iterations: 27
```

附录二, Adaptive Lasso+Stepwise模型的结论分析:

```
##
## Call:
## glm(formula = OVERALL_DIAGNOSIS ~ F5R + F5S + F9R + F10R + F13S +
##     F15R + F17R + F18R + F20S + F21R, family = binomial(link = logit),
##     data = adaptivelassonewdata, control = list(maxit = 1000))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5186  -0.2390  -0.0003   0.0467   3.4153
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 127.86211   41.42951   3.086  0.00203 **
## F5R           0.91689    0.34670   2.645  0.00818 **
## F5S          -0.88134    0.29939  -2.944  0.00324 **
## F9R          -0.61493    0.25835  -2.380  0.01730 *
## F10R         -0.24715    0.14101  -1.753  0.07964 .
## F13S         -0.26163    0.09593  -2.727  0.00638 **
## F15R         -0.20820    0.12935  -1.610  0.10748
## F17R         -0.59845    0.22098  -2.708  0.00676 **
## F18R          0.79523    0.29366   2.708  0.00677 **
## F20S         -0.62751    0.20385  -3.078  0.00208 **
## F21R         -0.17563    0.12751  -1.377  0.16839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.904  on 79  degrees of freedom
## Residual deviance:  29.836  on 69  degrees of freedom
## AIC: 51.836
##
## Number of Fisher Scoring iterations: 9
```

附录三,Generalized elastic net+Stepwise模型的结论分析:

```
##
## Call:
## glm(formula = OVERALL_DIAGNOSIS ~ F2S + F5R + F5S + F9R + F11R +
##     F13S + F14R + F14S + F15S + F17R + F18R + F18S + F20S, family = binomial(link = logit),
##     data = generalizednewdata, control = list(maxit = 1000))
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -9.934e-06  -2.110e-08   0.000e+00   2.110e-08   9.065e-06
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.660e+04  4.861e+07   0.001    0.999
## F2S         -2.627e+01  4.855e+04  -0.001    1.000
## F5R          2.530e+02  3.384e+05   0.001    0.999
## F5S         -2.257e+02  2.998e+05  -0.001    0.999
## F9R         -2.777e+02  3.693e+05  -0.001    0.999
## F11R        -1.228e+02  1.651e+05  -0.001    0.999
## F13S        -6.696e+01  8.893e+04  -0.001    0.999
## F14R         1.468e+02  1.997e+05   0.001    0.999
## F14S        -9.846e+01  1.327e+05  -0.001    0.999
## F15S        -6.747e+01  9.498e+04  -0.001    0.999
## F17R        -1.266e+02  1.730e+05  -0.001    0.999
## F18R         1.345e+02  1.829e+05   0.001    0.999
## F18S         6.729e+01  9.646e+04   0.001    0.999
## F20S        -1.196e+02  1.629e+05  -0.001    0.999
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.1090e+02  on 79  degrees of freedom
## Residual deviance: 7.1453e-10  on 66  degrees of freedom
## AIC: 28
##
## Number of Fisher Scoring iterations: 32
```

附录四，评估函数代码：

```
eva_test = function(model, test.data, thereshold=0.5){
  library(pROC)
  library(caret)
  # Evaluate the model using the test data
  pred = predict(model, test[,-1], type="response")
  pred.class = as.integer(pred > thereshold)
  # Confusion Matrix
  cfm = table(True = test$OVERALL_DIAGNOSIS, predict = pred.class)
  colnames(cfm) = c("0", "1")
  rownames(cfm) = c("0", "1")
  print(cfm)
  TP = cfm[1,1]; FP = cfm[2,1]; FN = cfm[1,2]; TN = cfm[2,2]
  recall = TP/(TP+FN); precison = TP/(TP+FP)
  accuracy = (TP+TN)/sum(cfm); F1 = 2/(1/precison + 1/recall)
  performance = c(recall, precison, accuracy, F1)
  names(performance) = c("recall", "precison", "accuracy", "F1")
  print(performance)
  # print(confusionMatrix(cfm, positive = "1"))
  # ROC Curve
  plot(roc(test$OVERALL_DIAGNOSIS, pred))

}

eva_train = function(model){

}
```