StatMod4

Igor

31 мая 2019 г

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3

library(memisc)

## Warning: package 'memisc' was built under R version 3.5.3

## Loading required package: lattice

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 3.5.3

##   
## Attaching package: 'memisc'

## The following object is masked from 'package:ggplot2':  
##   
## syms

## The following objects are masked from 'package:stats':  
##   
## contr.sum, contr.treatment, contrasts

## The following object is masked from 'package:base':  
##   
## as.array

library(DescTools)

## Warning: package 'DescTools' was built under R version 3.5.3

##   
## Attaching package: 'DescTools'

## The following object is masked from 'package:memisc':  
##   
## %nin%

library(lmtest)

## Warning: package 'lmtest' was built under R version 3.5.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.5.3

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(caTools)

## Warning: package 'caTools' was built under R version 3.5.3

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:memisc':  
##   
## collect, recode, rename, syms

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readxl)

## Warning: package 'readxl' was built under R version 3.5.3

library(knitr)

## Warning: package 'knitr' was built under R version 3.5.3

library(kernlab)

## Warning: package 'kernlab' was built under R version 3.5.2

##   
## Attaching package: 'kernlab'

## The following object is masked from 'package:ggplot2':  
##   
## alpha

library(caret)

## Warning: package 'caret' was built under R version 3.5.3

##   
## Attaching package: 'caret'

## The following objects are masked from 'package:DescTools':  
##   
## MAE, RMSE

library(mfx)

## Warning: package 'mfx' was built under R version 3.5.3

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 3.5.3

## Loading required package: betareg

## Warning: package 'betareg' was built under R version 3.5.3

library(pROC)

## Warning: package 'pROC' was built under R version 3.5.3

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(ResourceSelection)

## Warning: package 'ResourceSelection' was built under R version 3.5.3

## ResourceSelection 0.3-4 2019-01-08

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.5.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.5.3

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(nortest)

## Warning: package 'nortest' was built under R version 3.5.2

data <- read.csv("C:/Users/Igor/Downloads/diabetes.csv")  
data <- select(data,"Pregnancies","Glucose","BloodPressure","BMI","DiabetesPedigreeFunction","Outcome")  
names(data)

## [1] "Pregnancies" "Glucose"   
## [3] "BloodPressure" "BMI"   
## [5] "DiabetesPedigreeFunction" "Outcome"

cor.test(data$Pregnancies,data$Glucose)

##   
## Pearson's product-moment correlation  
##   
## data: data$Pregnancies and data$Glucose  
## t = 3.6134, df = 766, p-value = 0.0003219  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.0592571 0.1983860  
## sample estimates:  
## cor   
## 0.1294587

cor.test(data$Pregnancies,data$BloodPressure)

##   
## Pearson's product-moment correlation  
##   
## data: data$Pregnancies and data$BloodPressure  
## t = 3.9498, df = 766, p-value = 8.542e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.07124984 0.20992804  
## sample estimates:  
## cor   
## 0.141282

cor.test(data$Pregnancies,data$BMI)

##   
## Pearson's product-moment correlation  
##   
## data: data$Pregnancies and data$BMI  
## t = 0.48949, df = 766, p-value = 0.6246  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.05312764 0.08831688  
## sample estimates:  
## cor   
## 0.01768309

cor.test(data$Pregnancies,data$DiabetesPedigreeFunction)

##   
## Pearson's product-moment correlation  
##   
## data: data$Pregnancies and data$DiabetesPedigreeFunction  
## t = -0.92832, df = 766, p-value = 0.3535  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.10402026 0.03731008  
## sample estimates:  
## cor   
## -0.03352267

cor.test(data$Glucose,data$BloodPressure)

##   
## Pearson's product-moment correlation  
##   
## data: data$Glucose and data$BloodPressure  
## t = 4.2732, df = 766, p-value = 2.17e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.08273846 0.22094875  
## sample estimates:  
## cor   
## 0.1525896

cor.test(data$Glucose,data$BMI)

##   
## Pearson's product-moment correlation  
##   
## data: data$Glucose and data$BMI  
## t = 6.2737, df = 766, p-value = 5.891e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1527152 0.2873218  
## sample estimates:  
## cor   
## 0.2210711

cor.test(data$Glucose,data$DiabetesPedigreeFunction)

##   
## Pearson's product-moment correlation  
##   
## data: data$Glucose and data$DiabetesPedigreeFunction  
## t = 3.8374, df = 766, p-value = 0.0001346  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.06724638 0.20607934  
## sample estimates:  
## cor   
## 0.1373373

cor.test(data$BloodPressure,data$BMI)

##   
## Pearson's product-moment correlation  
##   
## data: data$BloodPressure and data$BMI  
## t = 8.1289, df = 766, p-value = 1.738e-15  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2153543 0.3456585  
## sample estimates:  
## cor   
## 0.2818053

cor.test(data$BloodPressure,data$DiabetesPedigreeFunction)

##   
## Pearson's product-moment correlation  
##   
## data: data$BloodPressure and data$DiabetesPedigreeFunction  
## t = 1.1431, df = 766, p-value = 0.2534  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.02956563 0.11168319  
## sample estimates:  
## cor   
## 0.04126495

cor.test(data$BMI,data$DiabetesPedigreeFunction)

##   
## Pearson's product-moment correlation  
##   
## data: data$BMI and data$DiabetesPedigreeFunction  
## t = 3.9317, df = 766, p-value = 9.198e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.0706052 0.2093086  
## sample estimates:  
## cor   
## 0.140647

**Вывод**: переменные, значительно коррелирующие между собой отсутствуют, так как коэффициент корреляции между любыми двумя независимыми признаками <|0.9|.

Преобразуем зависимую переменную как целочисленную:

data$Outcome<-as.integer(data$Outcome)

Разделим выборку на тестовую и обучающую:

set.seed(1)  
split<-sample.split(data$Outcome,SplitRatio = 0.7)  
train<-subset(data,split==TRUE)  
test<-subset(data,split==FALSE)

Построим регрессионные модели Probit и Logit:

**Заметка**: после первой попытки построить модели пришлось исключить три независимых признака (“SkinThickness”,“Insulin”,“Age”), так как они были статистиечски не значимы.

Вторая попытка:

Logit модель:

model\_1<-glm(Outcome~.,train,family = binomial(link="logit"))  
summary(model\_1)

##   
## Call:  
## glm(formula = Outcome ~ ., family = binomial(link = "logit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5644 -0.7036 -0.3910 0.6664 3.1355   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -8.524517 0.855772 -9.961 < 2e-16 \*\*\*  
## Pregnancies 0.122825 0.033290 3.690 0.000225 \*\*\*  
## Glucose 0.038319 0.004217 9.088 < 2e-16 \*\*\*  
## BloodPressure -0.012711 0.005980 -2.126 0.033541 \*   
## BMI 0.088796 0.017290 5.136 2.81e-07 \*\*\*  
## DiabetesPedigreeFunction 1.093122 0.365595 2.990 0.002790 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 696.28 on 537 degrees of freedom  
## Residual deviance: 490.87 on 532 degrees of freedom  
## AIC: 502.87  
##   
## Number of Fisher Scoring iterations: 5

Probit модель:

model\_2<-glm(Outcome~.,train,family=binomial(link="probit"))  
summary(model\_2)

##   
## Call:  
## glm(formula = Outcome ~ ., family = binomial(link = "probit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5848 -0.7243 -0.3790 0.6975 3.4415   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.898971 0.460302 -10.643 < 2e-16 \*\*\*  
## Pregnancies 0.073572 0.019067 3.859 0.000114 \*\*\*  
## Glucose 0.021772 0.002318 9.391 < 2e-16 \*\*\*  
## BloodPressure -0.007601 0.003455 -2.200 0.027813 \*   
## BMI 0.052508 0.009748 5.387 7.18e-08 \*\*\*  
## DiabetesPedigreeFunction 0.583169 0.206858 2.819 0.004815 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 696.28 on 537 degrees of freedom  
## Residual deviance: 492.55 on 532 degrees of freedom  
## AIC: 504.55  
##   
## Number of Fisher Scoring iterations: 5

Гоммит модель:

model\_3<-glm(Outcome~.,train,family = binomial(link="cloglog"))  
summary(model\_3)

##   
## Call:  
## glm(formula = Outcome ~ ., family = binomial(link = "cloglog"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8294 -0.7108 -0.4662 0.6960 2.8078   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.122352 0.562341 -10.887 < 2e-16 \*\*\*  
## Pregnancies 0.077535 0.022482 3.449 0.000563 \*\*\*  
## Glucose 0.026791 0.002721 9.846 < 2e-16 \*\*\*  
## BloodPressure -0.010661 0.003969 -2.686 0.007222 \*\*   
## BMI 0.060981 0.011644 5.237 1.63e-07 \*\*\*  
## DiabetesPedigreeFunction 0.448611 0.238826 1.878 0.060326 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 696.28 on 537 degrees of freedom  
## Residual deviance: 498.95 on 532 degrees of freedom  
## AIC: 510.95  
##   
## Number of Fisher Scoring iterations: 9

Вывод: все независимые признаки статистически значимы, т.к. для всех признаков Pr(>|z|) <0.05.

Сравнение трех моделей по информационным критериям:

mtable(model\_1,model\_2,model\_3,summary.stats = c("AIC","BIC"))

##   
## Calls:  
## model\_1: glm(formula = Outcome ~ ., family = binomial(link = "logit"),   
## data = train)  
## model\_2: glm(formula = Outcome ~ ., family = binomial(link = "probit"),   
## data = train)  
## model\_3: glm(formula = Outcome ~ ., family = binomial(link = "cloglog"),   
## data = train)  
##   
## ================================================================  
## model\_1 model\_2 model\_3   
## ----------------------------------------------------------------  
## (Intercept) -8.525\*\*\* -4.899\*\*\* -6.122\*\*\*   
## (0.856) (0.460) (0.562)   
## Pregnancies 0.123\*\*\* 0.074\*\*\* 0.078\*\*\*   
## (0.033) (0.019) (0.022)   
## Glucose 0.038\*\*\* 0.022\*\*\* 0.027\*\*\*   
## (0.004) (0.002) (0.003)   
## BloodPressure -0.013\* -0.008\* -0.011\*\*   
## (0.006) (0.003) (0.004)   
## BMI 0.089\*\*\* 0.053\*\*\* 0.061\*\*\*   
## (0.017) (0.010) (0.012)   
## DiabetesPedigreeFunction 1.093\*\* 0.583\*\* 0.449   
## (0.366) (0.207) (0.239)   
## ----------------------------------------------------------------  
## AIC 502.870 504.554 510.948   
## BIC 528.597 530.281 536.676   
## ================================================================  
## Significance: \*\*\* = p < 0.001; \*\* = p < 0.01; \* = p < 0.05

**Вывод**: наименьшие информационные критерии у логит-модели.

Для логит-модели: Коэффициент детерминации МакФаддена:

PseudoR2(model\_1,which="McFadden")

## McFadden   
## 0.2950109

**Вывод**: коэффициент достаточно мал, ~30%, а это значит изменения зависимой переменной объясняются изменениями совокупности признаков всего лишь на 30%.

lrtest(model\_1)

## Likelihood ratio test  
##   
## Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction  
## Model 2: Outcome ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 6 -245.44   
## 2 1 -348.14 -5 205.41 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Вывод**: так как p < 0.05, следовательно модель является значимой.

Для пробит-модели:

PseudoR2(model\_2,which = "McFadden")

## McFadden   
## 0.2925925

**Вывод**: коэффициент достаточно мал, ~29%, а это значит изменения зависимой переменной объясняются изменениями совокупности признаков всего лишь на 29%.

lrtest(model\_2)

## Likelihood ratio test  
##   
## Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction  
## Model 2: Outcome ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 6 -246.28   
## 2 1 -348.14 -5 203.73 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Для гампит модели:

PseudoR2(model\_3,which = "McFadden")

## McFadden   
## 0.2834089

lrtest(model\_3)

## Likelihood ratio test  
##   
## Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction  
## Model 2: Outcome ~ 1  
## #Df LogLik Df Chisq Pr(>Chisq)   
## 1 6 -249.47   
## 2 1 -348.14 -5 197.33 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Вывод**: так как p < 0.05, следовательно модель является значимой.

Preds1<-predict(model\_1,test,type="response")  
hoslem.test(test$Outcome,Preds1,g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: test$Outcome, Preds1  
## X-squared = 16.847, df = 8, p-value = 0.03175

**Вывод**: для логит-модели степень соответствия между оцененными вероятностями, спрогнозированные моделью и реальными вероятностями очень мала, всего лишь 0.03.

Preds2<-predict(model\_2,test,type="response")  
hoslem.test(test$Outcome,Preds2,g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: test$Outcome, Preds2  
## X-squared = 19.548, df = 8, p-value = 0.01219

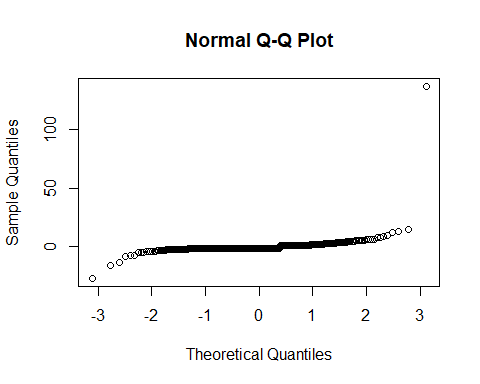
**Вывод**: для пробит-модели степень соответствия между оцененными вероятностями, спрогнозированные моделью и реальными вероятностями очень мала, всего лишь 0.01.

Preds3<-predict(model\_3,test,type="response")  
hoslem.test(test$Outcome,Preds3,g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: test$Outcome, Preds3  
## X-squared = 8.3456, df = 8, p-value = 0.4005

**Вывод**: для пробит-модели степень соответствия между оцененными вероятностями, спрогнозированные моделью и реальными вероятностями средняя, 0.04, что является лучшим показателем среди других моделей.

qqnorm(model\_1$residuals)



**Вывод**: судя по графику остатки не подчиняются нормальному закону распределения.

lillie.test(model\_1$residuals)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: model\_1$residuals  
## D = 0.30414, p-value < 2.2e-16

**Вывод**: поскольку p-value < 0.05 то нулевая гипотеза о согласии распределения остатков с нормальным законом распределения отвергается.

logitmfx(Outcome~.,data=test)

## Call:  
## logitmfx(formula = Outcome ~ ., data = test)  
##   
## Marginal Effects:  
## dF/dx Std. Err. z P>|z|   
## Pregnancies 0.0476313 0.0109780 4.3388 1.433e-05 \*\*\*  
## Glucose 0.0059406 0.0012286 4.8352 1.330e-06 \*\*\*  
## BloodPressure -0.0020503 0.0020566 -0.9970 0.318783   
## BMI 0.0160701 0.0052483 3.0620 0.002199 \*\*   
## DiabetesPedigreeFunction 0.1255835 0.1078624 1.1643 0.244305   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Вывод**: Маржинальный эффект для логит модели:

Признаки:

Беременности: 4.7%

Уровень глюкозы: 5.9%

Уровень кровеносного давления: -0.2%

БМИ: 1.6%

Диабет: 12.5%

probitmfx(Outcome~.,data=test)

## Call:  
## probitmfx(formula = Outcome ~ ., data = test)  
##   
## Marginal Effects:  
## dF/dx Std. Err. z P>|z|   
## Pregnancies 0.0476411 0.0106967 4.4538 8.436e-06 \*\*\*  
## Glucose 0.0056129 0.0011648 4.8188 1.445e-06 \*\*\*  
## BloodPressure -0.0017036 0.0020237 -0.8418 0.399894   
## BMI 0.0157139 0.0050645 3.1028 0.001917 \*\*   
## DiabetesPedigreeFunction 0.1035162 0.1049708 0.9861 0.324063   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Вывод**: маржинальный эффект для Пробит модели:

Признаки: 4.7%

Беременности: 4.7%

Уровень глюкозы: 0.5%

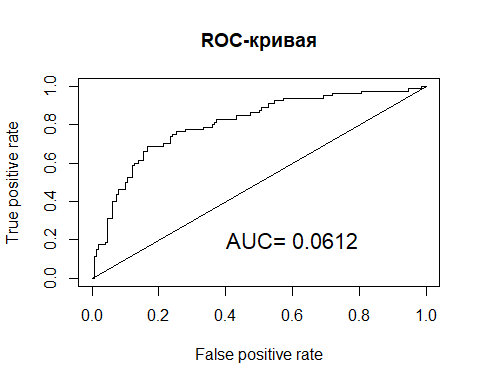
Кровеносное давление: -0.17%

БМИ: 1.5%

Диабет: 10%

ROC-кривая для логит-модели:

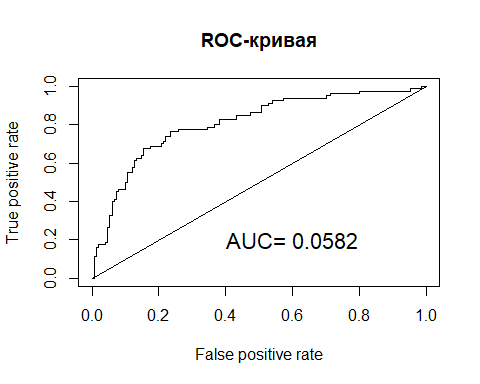
pr<-prediction(Preds1,test$Outcome)  
auc<-AUC(test$Outcome,Preds1)  
prf<-performance(pr,measure="tpr",x.measure = "fpr")  
plot(prf)  
lines(c(0,1),c(0,1))  
text(0.6,0.2,paste("AUC=",round(auc,4)),cex=1.4)  
title("ROC-кривая")



**Вывод**: так как коэффициент AUC = 0.06, значит качество логит-модель является неудовлетворительной.

ROC-кривая для Пробит модели:

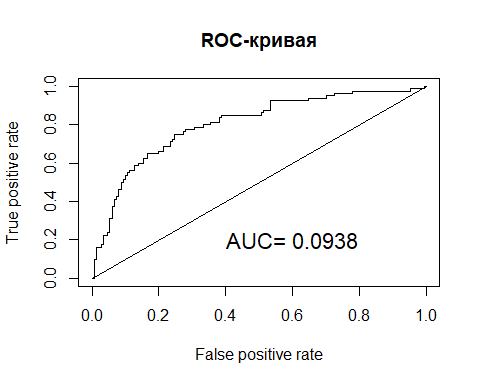
pr<-prediction(Preds2,test$Outcome)  
auc<-AUC(test$Outcome,Preds2)  
prf<-performance(pr,measure="tpr",x.measure = "fpr")  
plot(prf)  
lines(c(0,1),c(0,1))  
text(0.6,0.2,paste("AUC=",round(auc,4)),cex=1.4)  
title("ROC-кривая")



**Вывод**: так как коэффициент AUC = 0.06, значит качество пробит-модели является неудовлетворительной.

ROC-кривая для гампит-модели:

pr<-prediction(Preds3,test$Outcome)  
auc<-AUC(test$Outcome,Preds3)  
prf<-performance(pr,measure="tpr",x.measure = "fpr")  
plot(prf)  
lines(c(0,1),c(0,1))  
text(0.6,0.2,paste("AUC=",round(auc,4)),cex=1.4)  
title("ROC-кривая")



**Вывод**: так как коэффициент AUC = 0.09, значит качество логит-модели является неудовлетворительной, но это лучший показатель среди других моделей.