Vilniaus Universitetas

Binarinio atsako modeliai

Laboratorinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 202č

# Naudoti metodai

Darbas atliktas naudojant R ir SAS.

Naudoti R paketai:

*tidyverse*

*caret*

*MASS*

*cutpointr*

*yardstick*

*effects*

# Duomenys ir jų šaltiniai

Pimų tautybės moterų diagnostiniai matavimai skirti nustatyti ar paciento sergamumą diabetu..

Duomenų šaltinis - Kaggle. Prieiga per internetą: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

„Pregnancies“ - neštumų kiekis.

„Glucose“ - gliukozės koncentracija plazmoje gliukozės tolerancijos testo metu.

„BloodPressure“ - diastolinis kraujo spaudimas.

„BMI“ – kūno masės indeksas.

„SkinThickness“ - tricepso odos plotis.

„Insulin“ - glukozės tolerancijos testo rezultatas.

„DiabetesPedigreeFunction“ - diabeto tikėtinumas pagal šeimos istoriją.

„Age“ – amžius.

„Outcome“ – diabeto diagnozė (klasifikuojantis kintamasis).

# Tikslas ir uždaviniai

Tikslas: Rasti kokią įtaką tam tikri požymiai daro tikimybei sirgti diabetu ir prognozuoti diagnozę ar pacientas serga diabetu.

Uždaviniai:

Parinkti binarinio atsako modelį (parinkti jungties funkciją).

Paprastesnio (turinčio mažiau kovariančių) modelio suradimas.

Slenkstinės reikšmės parinkimas.

Gautų modelio koeficientų interpretacija

Modelio tinkamumo analizė.

# Atliktos analizės aprašymas

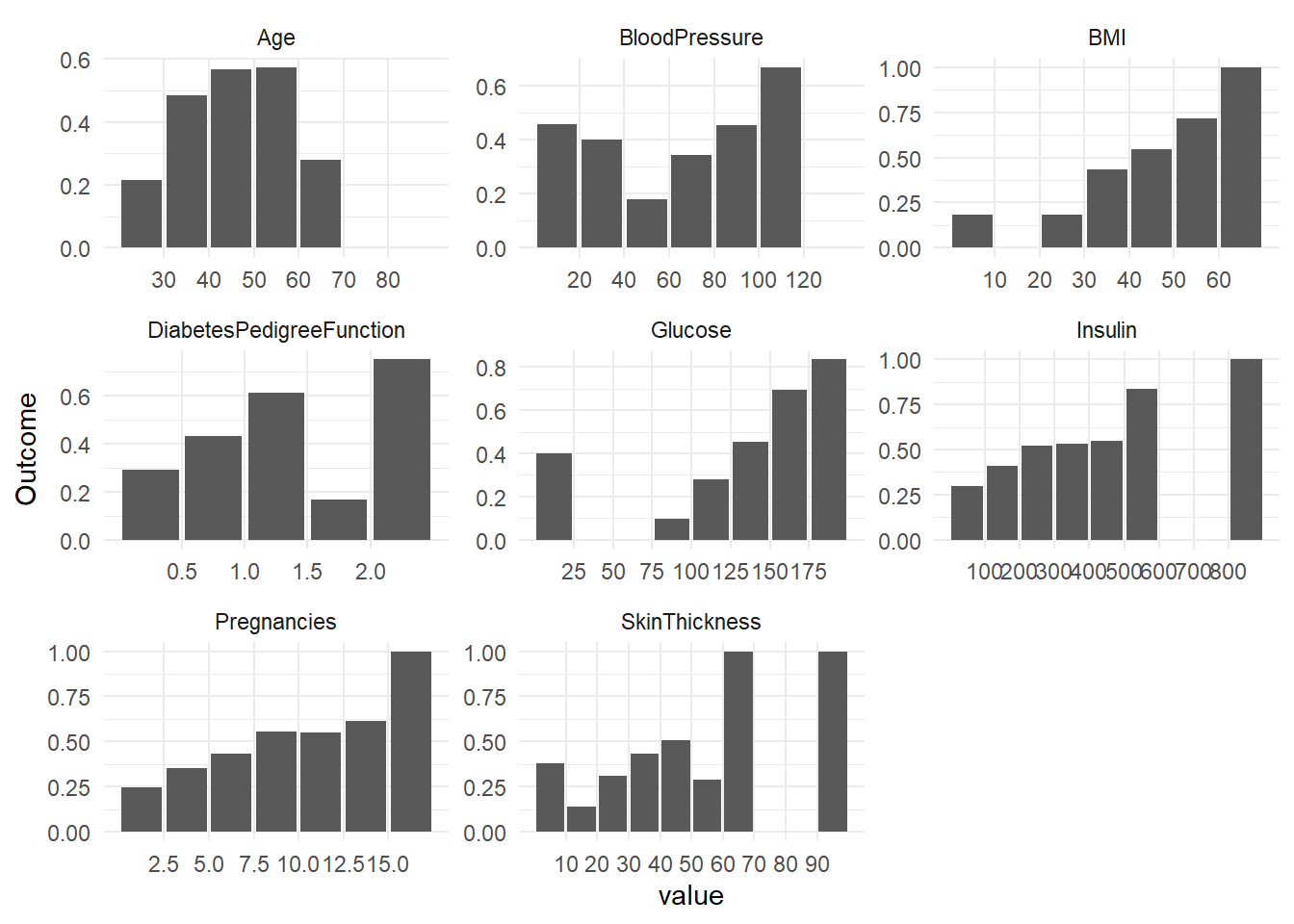
**1. Naudojant R**

Duomenų aibę sudaro duomenys apie 500 diabetų nesergančių ir 268 sergančių pacientų. Atliekant tiriamąją duomenų analizę palygintas kovariančių pasiskirstymas abiejose grupėse naudojant stačiakampes diagramas, pavaizduotos empirinės sirgimu diabetu tikimybės pagal kiekvieną kovariantę.

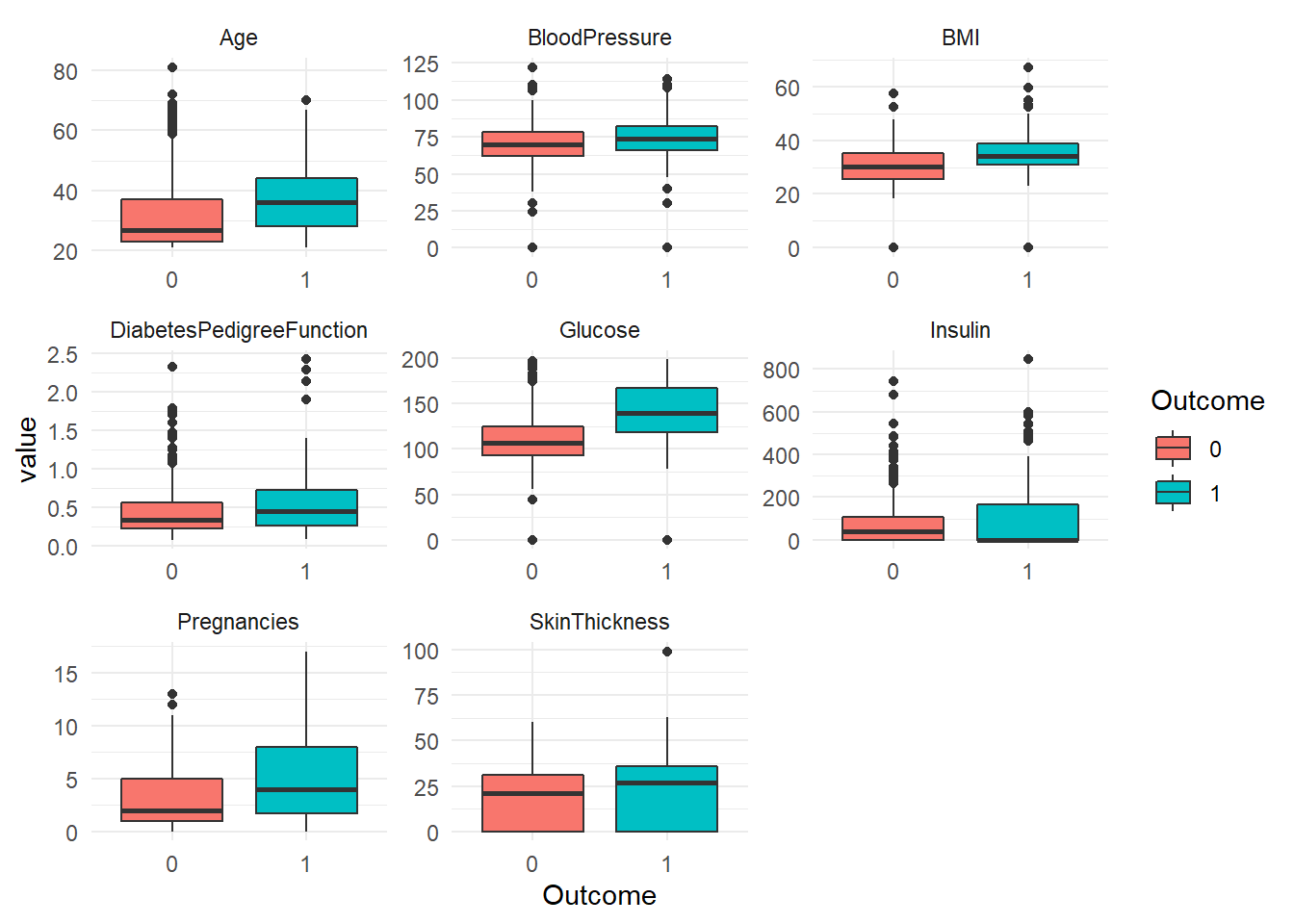
library(tidyverse)  
  
y <- read\_csv("diabetes.csv")  
  
table(y$Outcome)

##   
## 0 1   
## 500 268

# Empirinės tikimybės  
  
y\_plot <- y %>% pivot\_longer(1:8)  
  
  
y\_plot %>% ggplot(aes(value, Outcome)) +  
 stat\_summary(fun = mean, geom = "bar") +  
 facet\_wrap(vars(name), scales = "free") +  
 scale\_x\_binned(n.breaks = 8) +  
 theme\_minimal()



# stačiakampės diagramos  
y <- y %>% mutate(Outcome = factor(Outcome))  
y\_plot <- y\_plot %>% mutate(Outcome = factor(Outcome))  
  
y\_plot %>% ggplot(aes(Outcome, value, fill = Outcome)) +  
 geom\_boxplot() +  
 facet\_wrap(vars(name), scales = "free") +  
 theme\_minimal()



library(caret)  
library(yardstick)  
  
model <- glm(  
 formula = Outcome ~ ., family = binomial(logit),  
 data = y  
)  
  
1 - pchisq(model$deviance, model$df.residual) # goodness-of-fit testas

## [1] 0.8185965

model <- glm(  
 formula = Outcome ~ ., family = binomial(logit),  
 data = y  
)  
  
confusionMatrix(factor(as.numeric(model$fitted.values > 0.5)), factor(y$Outcome))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 445 112  
## 1 55 156  
##   
## Accuracy : 0.7826   
## 95% CI : (0.7517, 0.8112)  
## No Information Rate : 0.651   
## P-Value [Acc > NIR] : 1.373e-15   
##   
## Kappa : 0.4966   
##   
## Mcnemar's Test P-Value : 1.468e-05   
##   
## Sensitivity : 0.8900   
## Specificity : 0.5821   
## Pos Pred Value : 0.7989   
## Neg Pred Value : 0.7393   
## Prevalence : 0.6510   
## Detection Rate : 0.5794   
## Detection Prevalence : 0.7253   
## Balanced Accuracy : 0.7360   
##   
## 'Positive' Class : 0   
##

y\_2 <- y %>% mutate(pred = model$fitted.values)  
roc\_auc(y\_2, Outcome, pred, event\_level = "second")

## # A tibble: 1 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 roc\_auc binary 0.839

# kovariančių atranka  
model\_2 <- glm(  
 formula = Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction, family = binomial(logit),  
 data = y  
)  
  
anova(model, model\_2, test = "Chisq") # modelis statistiškai reikšmingai nesiskiria nuo modelio su visomis kovariantėmis

## Analysis of Deviance Table  
##   
## Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +   
## Insulin + BMI + DiabetesPedigreeFunction + Age  
## Model 2: Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 759 723.45   
## 2 762 728.56 -3 -5.1142 0.1636

model$aic

## [1] 741.4454

model\_2$aic

## [1] 740.5596

exp(coef(model\_2))

## (Intercept) Pregnancies Glucose   
## 0.0003509201 1.1658987706 1.0352654766   
## BloodPressure BMI DiabetesPedigreeFunction   
## 0.9880648446 1.0885341500 2.4858820555

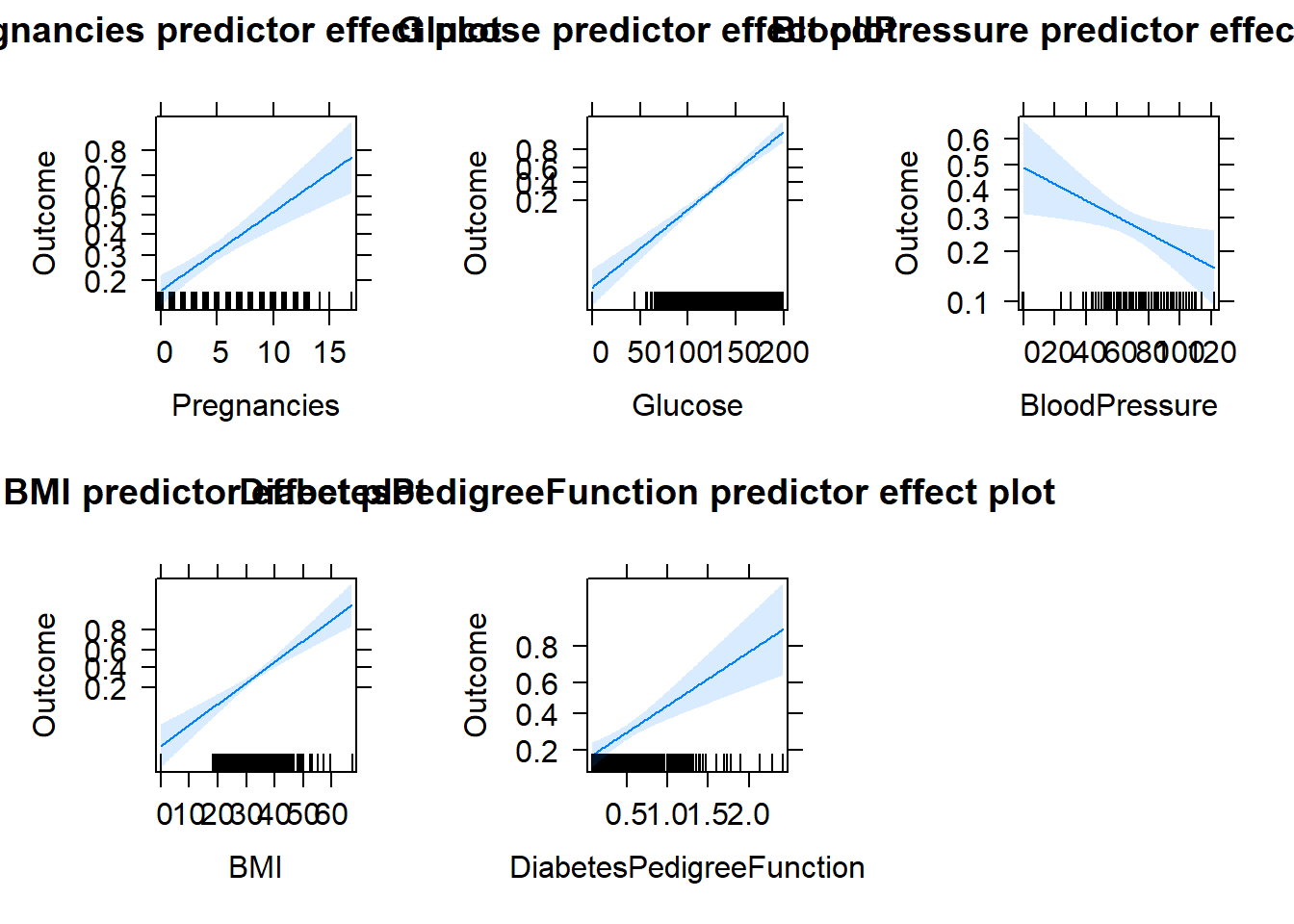
exp(confint(model\_2))

## 2.5 % 97.5 %  
## (Intercept) 8.892197e-05 0.001261429  
## Pregnancies 1.104688e+00 1.232240806  
## Glucose 1.028583e+00 1.042375734  
## BloodPressure 9.782664e-01 0.997813080  
## BMI 1.059521e+00 1.119907794  
## DiabetesPedigreeFunction 1.404567e+00 4.451068106

confusionMatrix(factor(as.numeric(model\_2$fitted.values > 0.5)), factor(y$Outcome))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 441 114  
## 1 59 154  
##   
## Accuracy : 0.7747   
## 95% CI : (0.7435, 0.8038)  
## No Information Rate : 0.651   
## P-Value [Acc > NIR] : 6.403e-14   
##   
## Kappa : 0.4795   
##   
## Mcnemar's Test P-Value : 4.034e-05   
##   
## Sensitivity : 0.8820   
## Specificity : 0.5746   
## Pos Pred Value : 0.7946   
## Neg Pred Value : 0.7230   
## Prevalence : 0.6510   
## Detection Rate : 0.5742   
## Detection Prevalence : 0.7227   
## Balanced Accuracy : 0.7283   
##   
## 'Positive' Class : 0   
##

# Modelio kovariačių efektai  
library(effects)  
plot(predictorEffects(model\_2))

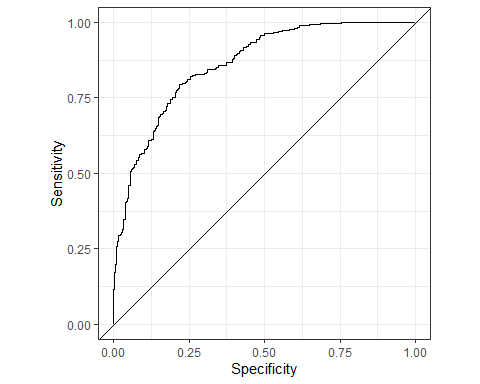


Pradinio modelio su visomis kovariantėmis (naudojant logit jungties funkciją) tikslumas (angl. accuracy) 78%, plotas po ROC kreive 0.84.

Rasta, kad modelis tik su kovariantėmis „Pregnancies“, „Glucose“ „BloodPressure“ „BMI“ ir „DiabetesPedigreeFunction“ statistiškai reikšmingai nesiskiria nuo modelio su visomis kovariantėmis (p=0.18). Modelio tikslumas 77%. Plotas po ROC kreive 0.84.

Modelio koeficientų interpretacija standartinė logit modeliui (pvz. Paciento kūno masės indeksui (angl. BMI) padidėjus 1, tikimybė, kad pacientas serga diabetu padidėja 1.08 kartus).

library(cutpointr)  
library(yardstick)  
  
y\_2 <- y %>% mutate(pred = model\_2$fitted.values)  
  
cp <- cutpointr(y\_2, pred, Outcome,  
 pos\_class = "1", direction = ">=",  
 method = maximize\_metric, metric = youden  
)  
  
  
cp$roc\_curve[[1]] %>%  
 ggplot(aes(x = 1 - tnr, y = tpr)) +  
 geom\_path() +  
 coord\_equal() +  
 geom\_abline() +  
 theme\_bw() +  
 xlab("Specificity") +  
 ylab("Sensitivity")

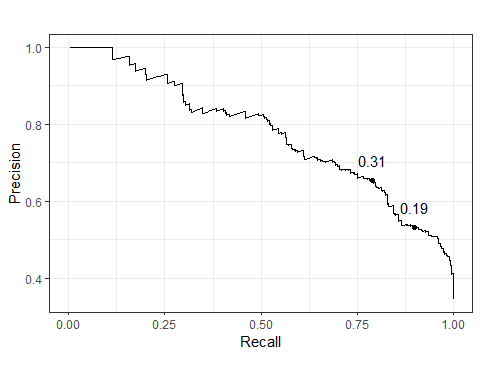


Atsižvelgiant į didesnį nesergančių pacientų kiekį duomenyse (stulp. “Outcome” reikšmė 0) laikyta, kad ROC kreivė gali teigti klaidinančia informaciją apie modelio kokybę. Papildomai pavaizduotas modelio Precision-Recall grafikas. Atsižvelgiant į uždavinio specifiką (laikyta, kad neteisingai diagnozuotos neigiamos diagnozės (False Negative) kaina didesnė už neteisingai diagnozuotą teigiamą diabeto diagnozę (False Positive)) modeliui siekta parinkti kitą slenkstinę reikšmę (angl. cutoff value).

roc\_auc(y\_2, Outcome, pred, event\_level = "second")

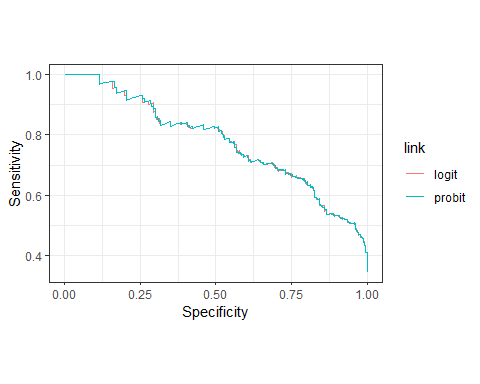
## # A tibble: 1 x 3  
## .metric .estimator .estimate  
## <chr> <chr> <dbl>  
## 1 roc\_auc binary 0.837

# TN skaičius nėra svarbus uždaviniui  
# be to, 0 skaičius šiek tiek didesnis už 1  
# todėl naudojama PR kreivė  
  
cutoff <- cp$roc\_curve[[1]] %>%  
 filter(tpr > 0.9) %>%  
 pull(m) %>%  
 max()  
  
labels <- filter(cp$roc\_curve[[1]], (m %in% c(max(m), cutoff))) %>% round(2)  
  
  
cp$roc\_curve[[1]] %>%  
 ggplot(aes(x = tpr, y = tp / (fp + tp))) +  
 geom\_point(data = labels) +  
 geom\_text(data = labels, aes(label = x.sorted), nudge\_y = 0.05) +  
 geom\_path() +  
 coord\_equal() +  
 theme\_bw() +  
 xlab("Recall") +  
 ylab("Precision")



# optimalios slenkstinės reikšmės pagal Youden indeksą ir pasirinkus ribą Sensitivity > 0.9

# Palyginimas su probit modeliu  
model\_3 <- glm(  
 formula = Outcome ~ Pregnancies + Glucose + BloodPressure + BMI + DiabetesPedigreeFunction, family = binomial(probit),  
 data = y  
)  
  
  
y\_3 <- y %>% mutate(pred = model\_3$fitted.values)  
  
  
cp\_2 <- cutpointr(y\_3, pred, Outcome,  
 method = maximize\_metric, metric = F1\_score  
)  
  
  
cp$roc\_curve[[1]] %>%  
 mutate(link = "logit") %>%  
 rbind((cp\_2$roc\_curve[[1]] %>% mutate(link = "probit"))) %>%  
 ggplot(aes(x = tpr, y = tp / (fp + tp), color = link)) +  
 geom\_path() +  
 coord\_equal() +  
 theme\_bw() +  
 xlab("Specificity") +  
 ylab("Sensitivity")



Rasta slenkstinė reikšmė pagal Joudeno (Youden) indeksą - 0.31. Naudojant kriterijų, siekiantį teisingai aptikti bent 90% procentų teigiamų diagnozių (Sensitivity > 0.9) - 0.19 (abi reikšmės pažymėtos Precision-Recall grafike). Modelis palygintas su modeliu su tokiomis pačiomis kovariantėmis, tačiau naudojančiu probit junties funkciją. Palygintos modelių ROC kreivės. Reikšmingų skirtumų tarp modelių nerasta.

Rezultatai

Naudojant logistinę regresiją siekta rasti kokie požymiai susiję su didžiausia rizika sirgti diabetu, prognozuoti šios ligos diagnozę.

Tyrimo metu rasta, kad modelis su kovariantėmis „Pregnancies“, „Glucose“ „BloodPressure“ „BMI“ ir „DiabetesPedigreeFunction“ statistiškai reikšmingai nesiskiria nuo modelio su visomis kovariantėmis (p=0.18). Modelio tikslumas (angl. accuracy) 0.77. Plotas po modelio ROC kreive = 0.84.

Atsižvelgiant į užduoties specifiką, pasirinktos kitos modelio slenkstinės reikšmės: siekiant teisingai aptikti bent 90% teigiamų diabeto diagnozių pasirinkta slenkstinė riba 0.19.

Reišmingų skirtumų tarp modelio naudojančio logit ir probit jungties funkcijas nerasta.

**2. Naudojant SAS**

PROC IMPORT DATAFILE='/home/u45871880/diabetes.csv'

DBMS=CSV

OUT=data;

GETNAMES=YES;

RUN;

%MACRO boxplot(column);

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.DATA;

vbox &column / category=Outcome;

yaxis grid;

run;

%MEND;

%boxplot(Pregnancies);

%boxplot(Glucose);

%boxplot(BloodPressure);

%boxplot(SkinThickness)

%boxplot(Insulin);

%boxplot(Age);

%boxplot(DiabetesPedigreeFunction);

%boxplot(BMI);

\* Modelis su visomis kovariantėmis;

PROC LOGISTIC DATA=data DESCENDING

plots(only)=(roc(ID=cutpoint) effect(X=(Pregnancies Glucose BloodPressure SkinThickness

Insulin Age DiabetesPedigreeFunction BMI) CLBAND=YES ALPHA=0.05));

MODEL Outcome = Pregnancies Glucose BloodPressure SkinThickness

Insulin BMI DiabetesPedigreeFunction Age /

RSQUARE CTABLE PPROB=(0.1 TO 0.9 BY 0.1) EXPB LACKFIT scale=none clparm=wald

RUN;

| **Response Profile** | | |
| --- | --- | --- |
| **Ordered Value** | **Outcome** | **Total Frequency** |
| **1** | 1 | 268 |
| **2** | 0 | 500 |

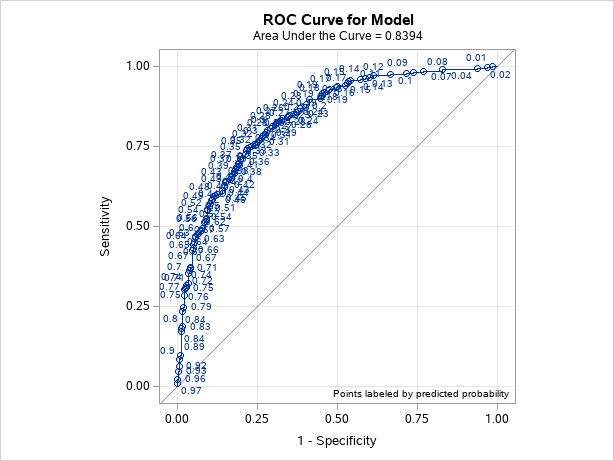
**Probability modeled is Outcome='1'.**

| **Testing Global Null Hypothesis: BETA=0** | | | |
| --- | --- | --- | --- |
| **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** |
| **Likelihood Ratio** | 270.0385 | 8 | <.0001 |
| **Score** | 232.8984 | 8 | <.0001 |
| **Wald** | 167.7255 | 8 | <.0001 |

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** | **Exp(Est)** |
| **Intercept** | 1 | -8.4047 | 0.7166 | 137.5452 | <.0001 | 0.000 |
| **Pregnancies** | 1 | 0.1232 | 0.0321 | 14.7466 | 0.0001 | 1.131 |
| **Glucose** | 1 | 0.0352 | 0.00371 | 89.8965 | <.0001 | 1.036 |
| **BloodPressure** | 1 | -0.0133 | 0.00523 | 6.4537 | 0.0111 | 0.987 |
| **SkinThickness** | 1 | 0.000619 | 0.00690 | 0.0080 | 0.9285 | 1.001 |
| **Insulin** | 1 | -0.00119 | 0.000901 | 1.7485 | 0.1861 | 0.999 |
| **BMI** | 1 | 0.0897 | 0.0151 | 35.3467 | <.0001 | 1.094 |
| **DiabetesPedigreeFunc** | 1 | 0.9452 | 0.2991 | 9.9828 | 0.0016 | 2.573 |
| **Age** | 1 | 0.0149 | 0.00933 | 2.5372 | 0.1112 | 1.015 |

| **Hosmer and Lemeshow Goodness-of-Fit Test** | | |
| --- | --- | --- |
| **Chi-Square** | **DF** | **Pr > ChiSq** |
| 8.3230 | 8 | 0.4026 |

| **Classification Table** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Prob Level** | **Correct** | | **Incorrect** | | **Percentages** | | | | |
| **Event** | **Non- Event** | **Event** | **Non- Event** | **Correct** | **Sensi- tivity** | **Speci- ficity** | **Pos Pred** | **Neg Pred** |
| **0.100** | 261 | 146 | 354 | 7 | 53.0 | 97.4 | 29.2 | 42.4 | 95.4 |
| **0.200** | 240 | 277 | 223 | 28 | 67.3 | 89.6 | 55.4 | 51.8 | 90.8 |
| **0.300** | 209 | 356 | 144 | 59 | 73.6 | 78.0 | 71.2 | 59.2 | 85.8 |
| **0.400** | 177 | 408 | 92 | 91 | 76.2 | 66.0 | 81.6 | 65.8 | 81.8 |
| **0.500** | 154 | 443 | 57 | 114 | 77.7 | 57.5 | 88.6 | 73.0 | 79.5 |
| **0.600** | 130 | 457 | 43 | 138 | 76.4 | 48.5 | 91.4 | 75.1 | 76.8 |
| **0.700** | 97 | 475 | 25 | 171 | 74.5 | 36.2 | 95.0 | 79.5 | 73.5 |
| **0.800** | 62 | 489 | 11 | 206 | 71.7 | 23.1 | 97.8 | 84.9 | 70.4 |
| **0.900** | 22 | 495 | 5 | 246 | 67.3 | 8.2 | 99.0 | 81.5 | 66.8 |



\* Pažingsninė regresija kovariančių atrinkimui;

PROC LOGISTIC DATA=data DESCENDING plots(only)=(roc);

MODEL Outcome = Pregnancies Glucose BloodPressure SkinThickness

Insulin BMI DiabetesPedigreeFunction Age /

CTABLE PPROB=(0.1 TO 0.9 BY 0.1) EXPB

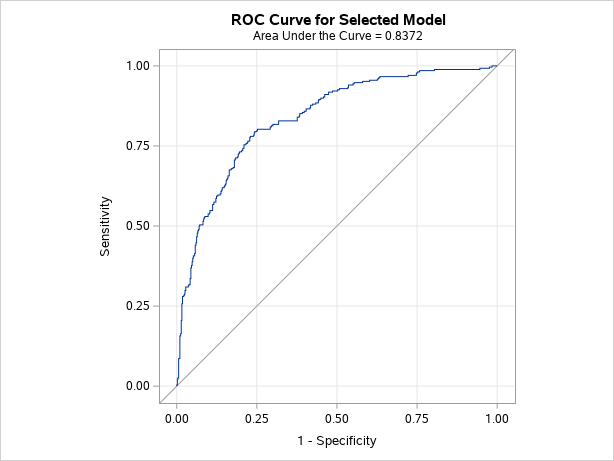
scale=none clparm=wald outroc=performance SELECTION=stepwise

RUN;

| **Summary of Stepwise Selection** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Effect** | | **DF** | **Number In** | **Score Chi-Square** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Entered** | **Removed** |
| **1** | **Glucose** |  | 1 | 1 | 167.1922 |  | <.0001 |
| **2** | **BMI** |  | 1 | 2 | 34.3033 |  | <.0001 |
| **3** | **Pregnancies** |  | 1 | 3 | 27.3305 |  | <.0001 |
| **4** | **DiabetesPedigreeFunc** |  | 1 | 4 | 9.6773 |  | 0.0019 |
| **5** | **BloodPressure** |  | 1 | 5 | 5.8123 |  | 0.0159 |

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** | **Exp(Est)** |
| **Intercept** | 1 | -7.9549 | 0.6758 | 138.5505 | <.0001 | 0.000 |
| **Pregnancies** | 1 | 0.1535 | 0.0278 | 30.4074 | <.0001 | 1.166 |
| **Glucose** | 1 | 0.0347 | 0.00339 | 104.3051 | <.0001 | 1.035 |
| **BloodPressure** | 1 | -0.0120 | 0.00503 | 5.6969 | 0.0170 | 0.988 |
| **BMI** | 1 | 0.0848 | 0.0141 | 36.0703 | <.0001 | 1.089 |
| **DiabetesPedigreeFunc** | 1 | 0.9106 | 0.2940 | 9.5919 | 0.0020 | 2.486 |

| **Classification Table** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Prob Level** | **Correct** | | **Incorrect** | | **Percentages** | | | | |
| **Event** | **Non- Event** | **Event** | **Non- Event** | **Correct** | **Sensi- tivity** | **Speci- ficity** | **Pos Pred** | **Neg Pred** |
| **0.100** | 259 | 138 | 362 | 9 | 51.7 | 96.6 | 27.6 | 41.7 | 93.9 |
| **0.200** | 239 | 275 | 225 | 29 | 66.9 | 89.2 | 55.0 | 51.5 | 90.5 |
| **0.300** | 215 | 362 | 138 | 53 | 75.1 | 80.2 | 72.4 | 60.9 | 87.2 |
| **0.400** | 181 | 412 | 88 | 87 | 77.2 | 67.5 | 82.4 | 67.3 | 82.6 |
| **0.500** | 152 | 439 | 61 | 116 | 77.0 | 56.7 | 87.8 | 71.4 | 79.1 |
| **0.600** | 133 | 462 | 38 | 135 | 77.5 | 49.6 | 92.4 | 77.8 | 77.4 |
| **0.700** | 96 | 477 | 23 | 172 | 74.6 | 35.8 | 95.4 | 80.7 | 73.5 |
| **0.800** | 63 | 492 | 8 | 205 | 72.3 | 23.5 | 98.4 | 88.7 | 70.6 |
| **0.900** | 23 | 496 | 4 | 245 | 67.6 | 8.6 | 99.2 | 85.2 | 66.9 |



\* Atsižvelgiai į uždavinio specifiką

\* Sukuriamas Precision-Recall grafikas alternativių slenksninių reikšmių parinkimui;

data precision\_recall;

set performance;

precision = \_POS\_/(\_POS\_ + \_FALPOS\_);

recall = \_POS\_/(\_POS\_ + \_FALNEG\_);

F\_stat = harmean(precision,recall);

if mod(\_N\_, 20) = 0 then \_PROB\_=\_PROB\_;

else \_PROB\_ = .;

run;

Proc SQL;

create table precision\_recall as

Select \*

From precision\_recall

having \_step\_ = max(\_step\_);

Quit;

proc sort data=precision\_recall;

by recall;

run;

ods graphics / reset width=6.4in height=4.8in imagemap;

proc sgplot data=WORK.PRECISION\_RECALL;

SERIES X = recall Y = precision / DATALABEL=\_PROB\_;

run;

ods graphics / reset;

