Statystyka w finansach i ubezpieczeniach
Sprawozdanie 2

Justyna Niedźwiedzka 229877

13 stycznia 2021

Część 1

Opis danych

Zbiór danych *germancredit* dostępny w pakiecie **scorecard** jest złożony z 1000 obserwacji (klientów) i zawiera 20 różnych charakterystyk opisujących klientów oraz binarną zmienną objaśnianą *creditability* przyjmującą wartość *good* w przypadku terminowego spłacania kredytu przez klienta oraz wartość *bad* w przypadku nieterminowego spłacania kredytu. Poniżej przedstawiono szczegółowy opis zmiennych:

- *status.of.existing.checking.account*- zmienna typu factor określająca status istniejącego rachunku rozliczeniowego, z 4 poziomami:
  - "... $<$ 0 DM ",
  - "0 $\leq$ ... $<$ 200 DM",
  - "... $\leq$ 200 DM / salary assignments for at least 1 year",
  - "no checking account".
- *duration.in.month*- zmienna numeryczna określająca czas trwania w miesiącach,
- *credit.history*- zmienna typu factor oznaczająca historię kredytową, przyjmująca 5 poziomów:
  - "no credits taken/ all credits paid back duly",
  - "all credits at this bank paid back duly",
  - "existing credits paid back duly till now",
  - "delay in paying off in the past",
  - "critical account/ other credits existing (not at this bank)".
- *purpose*- zmienna typu character określająca cel kredytu, przyjmująca następujące wartości: "business", "car (new)", "car (used)", "domestic appliances", "education", "furniture/equipment", "others", "radio/television", "repairs", "retraining".
- *credit.amount*- zmienna numeryczna oznaczająca kwotę kredytu.
- *savings.account.and.bonds*- zmienna typu factor opisująca ilość oszczędności, przyjmująca 5 poziomów:
  - "... $<$ 100 DM",
  - "100 $\leq$ ... $<$ 500 DM",
  - "500 $\leq$ ... $<$ 1000 DM",
  - "... $\leq$ 1000 DM",
  - "unknown/ no savings account".
- *present.employment.since*- zmienna typu factor określająca czas trwania obecnego zatrudnienia, z 5 poziomami:
  - "unemployed",

  - "... < 1 year",
  - "1 ≤ ... < 4 years",
  - "4 ≤ ... < 7 years",
  - "... ≤ 7 years".

- *installment.rate.in.percentage.of.disposable.income*- zmienna numeryczna oznaczająca wysokość rat wyrażona jako procent dochodu do dyspozycji.

- *personal.status.and.sex*- zmienna typu factor wskazująca status osobisty i płeć, z 5 poziomami:
  - "male : divorced/separated",
  - "female : divorced/separated/married",
  - "male : single",
  - "male : married/widowed",
  - "female : single".

- *other.debtors.or.guarantors*- zmienna typu factor przyjmująca 3 wartości: "none", "co-applicant", "guarantor".

- *present.residence.since*- zmienna numeryczna określająca czas zamieszkania w obecnym miejscu.

- *property*- zmienna typu factor określająca najwyższą wycenę nieruchomości klienta, z poziomami:
  - "real estate",
  - "building society savings agreement/ life insurance",
  - "car or other, not in attribute Savings account/bonds",
  - "unknown / no property".

- *age.in.years*- zmienna numeryczna oznaczająca wiek klienta.

- *other.installment.plans*- zmienna typu factor określająca inne plany ratalne posiadane przez klienta, z poziomami
  - "bank",
  - "stores",
  - "none".

- *housing*- zmienna typu factor oznaczająca typ opłat za mieszkanie, z poziomami "rent", "own", "for free".

- *number.of.existing.credits.at.this.bank*- zmienna numeryczna określająca liczbę istniejących kredytów w danym banku.

- *job*- zmienna typu factor wskazująca na status zatrudnienia, przyjmująca 4 poziomy:
  - "unemployed/ unskilled - non-resident",

  - "unskilled - resident",
  - "skilled employee / official",
  - "management/ self-employed/ highly qualified employee/ officer".

- *number.of.people.being.liable.to.provide.maintenance.for*- zmienna numeryczna oznaczająca liczbę osób na utrzymaniu.
- *telephone*- zmienna typu factor wskazująca, czy klient posiada zarejestrowany numer telefonu. Jest zmienną binarną przyjmującą wartości "none", "yes, registered under the customers name".
- *foreign.worker*- zmienna typu factor (binarna) określająca, czy klient jest pracownikiem zagranicznym, przyjmująca wartości "yes" oraz "no".
- *creditability*- zmienna typu factor wskazująca ryzyko kredytowe, określająca klienta jako złego- "bad" lub dobrego- "good".

Zbiór danych zawiera dane osób ubiegających się o kredyt w przeszłości. Kandydaci są oceniani jako dobrzy lub źli. Modele tych danych można wykorzystać do określenia, czy nowi wnioskodawcy stanowią dobre, czy złe ryzyko kredytowe.

Zbiór uczący i zbiór testowy

Dokonamy podziału danych na zbiór uczący i zbiór testowy. W tym celu skorzystamy z funkcji *split_df*. Zanim jednak to zrobimy, przekształcimy zmienną *purpose* na zmienną typu factor, a zmienną *creditability* na zmienną numeryczną.

```
library(scorecard)
data <- data.frame(germancredit)
attach(data)
data$purpose <- as.factor(data$purpose)
data$creditability <- as.numeric(data$creditability)
```

Po przekształceniu zmiennej *creditability* otrzymujemy 2 dla poziomu good oraz 1 dla bad. Jeszcze raz zmienimy oznaczenie przyjmując 0 jako good.

```
data$creditability[data$creditability==2] <- 0
```

Usuniemy jeszcze nieużywany poziom "female : single" ze zmiennej *personal.status.and.sex*

```
data$personal.status.and.sex <- droplevels(data$personal.status.and.sex)
```

Teraz zajmiemy się tworzeniem zbioru uczącego i testowego.

```
division <- split_df(data, y = data$creditability, ratio = c(0.7,0.3),
                     seed = 1234, name_dfs = c("train","test"))
train_set <- division$train
test_set <- division$test
```

Otrzymaliśmy w ten sposób zbiór uczący zawierający 698 obserwacji oraz testowy składający się z 302 obserwacji.

Funkcja klasyfikująca Fishera

Szczególnym przypadkiem analizy dyskryminacyjnej jest liniowa analiza dyskryminacyjna (linear discriminant analysis - LDA). Idea liniowej analizy dyskryminacyjnej polega na zredukowaniu wektora cech do jednego wymiaru przez rzutowanie danych dotyczących charakterystyk potencjalnego kredytobiorcy na prostą. Algebraicznie oznacza to zastąpienie wektora $\mathbf{X} = (X_1, \ldots, X_p)'$ kombinacją liniową wektora atrybutów aplikanta X i wektora w postaci

$$U = \mathbf{w}'\mathbf{X} = \sum_{j=1}^{p} w_j X_j.$$

Celem liniowej analizy dyskryminacyjnej jest taki wybór wektora $\mathbf{w}$, który w pewnym sensie gwarantuje najlepsze rozdzielenie wartości zmiennej losowej U w grupach.

Napisanie funkcji, która klasyfikowałaby jednostki zgodnie z regułą klasyfikującą Fishera jest możliwie, ale trudne.

W pakiecie **MASS** znajduje się funkcja klasyfikująca Fishera *lda*, jednak nie działa ona dla naszego zbioru danych.

```
library(MASS)

##
## Attaching package: 'MASS'
## The following object is masked from 'data':
##
##     housing

lda <- lda(formula = data$creditability~., data)

## Error in lda.default(x, grouping, ...): brakuje argumentu 'grouping', a nie ma
określonej wartości domyślnej

print(lda)

## function (x, ...)
## UseMethod("lda")
## <bytecode: 0x00000000196abda8>
## <environment: namespace:MASS>
```

Część 2

Model regresji logistycznej

Na podstawie danych ze zbioru uczącego wybierzemy model regresji logistycznej, przyjmując za zmienną zależną *creditability*, a pozostałe zmienne za potencjalne predyktory. Skorzystamy z regresji krokowej wstecznej oraz kryterium AIC.

```
library(stats)
model1 <- glm(creditability~., data = train_set,
              family = binomial(link = "logit"))
model1_AIC <- step(model1, direction = "backward", k = 2)

## Start:  AIC=726.48
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
##     personal.status.and.sex + other.debtors.or.guarantors + present.residence.since +
##     property + age.in.years + other.installment.plans + housing +
##     number.of.existing.credits.at.this.bank + job + number.of.people.being.liable.to.
##     telephone + foreign.worker
##
##                                                        Df Deviance    AIC
## - job                                                   3   628.78 720.78
## - property                                              3   630.35 722.35
## - present.residence.since                               1   628.49 724.49
## - personal.status.and.sex                               3   632.58 724.58
## - number.of.people.being.liable.to.provide.maintenance.for  1   628.68 724.68
## - housing                                               2   630.69 724.69
## - number.of.existing.credits.at.this.bank               1   628.73 724.73
## - age.in.years                                          1   629.03 725.03
## - present.employment.since                              4   635.52 725.52
## <none>                                                      628.48 726.48
## - telephone                                             1   631.68 727.68
## - other.debtors.or.guarantors                           2   634.21 728.21
## - savings.account.and.bonds                             4   638.27 728.27
## - foreign.worker                                        1   632.40 728.40
## - installment.rate.in.percentage.of.disposable.income   1   633.82 729.82
## - other.installment.plans                               2   637.83 731.83
## - credit.amount                                         1   635.85 731.85
## - credit.history                                        4   642.13 732.13
## - duration.in.month                                     1   636.49 732.49
## - purpose                                               9   674.51 754.51
## - status.of.existing.checking.account                   3   671.13 763.13
##
## Step:  AIC=720.78
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
```

```
##     personal.status.and.sex + other.debtors.or.guarantors + present.residence.since +
##     property + age.in.years + other.installment.plans + housing +
##     number.of.existing.credits.at.this.bank + number.of.people.being.liable.to.provid
##     telephone + foreign.worker
##
##                                                        Df Deviance    AIC
## - property                                              3   630.91 716.91
## - present.residence.since                               1   628.81 718.81
## - number.of.people.being.liable.to.provide.maintenance.for  1   628.93 718.93
## - number.of.existing.credits.at.this.bank               1   629.02 719.02
## - personal.status.and.sex                               3   633.05 719.05
## - housing                                               2   631.08 719.08
## - age.in.years                                          1   629.34 719.34
## - present.employment.since                              4   636.04 720.04
## <none>                                                      628.78 720.78
## - telephone                                             1   631.75 721.75
## - other.debtors.or.guarantors                           2   634.64 722.64
## - foreign.worker                                        1   632.66 722.66
## - savings.account.and.bonds                             4   638.99 722.99
## - installment.rate.in.percentage.of.disposable.income   1   634.61 724.61
## - other.installment.plans                               2   638.14 726.14
## - credit.history                                        4   642.57 726.57
## - credit.amount                                         1   636.84 726.84
## - duration.in.month                                     1   637.02 727.02
## - purpose                                               9   674.74 748.74
## - status.of.existing.checking.account                   3   671.22 757.22
##
## Step:  AIC=716.91
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
##     personal.status.and.sex + other.debtors.or.guarantors + present.residence.since +
##     age.in.years + other.installment.plans + housing + number.of.existing.credits.at.
##     number.of.people.being.liable.to.provide.maintenance.for +
##     telephone + foreign.worker
##
##                                                        Df Deviance    AIC
## - housing                                               2   632.70 714.70
## - present.residence.since                               1   630.93 714.93
## - personal.status.and.sex                               3   634.96 714.96
## - number.of.existing.credits.at.this.bank               1   631.06 715.06
## - number.of.people.being.liable.to.provide.maintenance.for  1   631.07 715.07
## - age.in.years                                          1   631.58 715.58
## - present.employment.since                              4   638.25 716.25
## <none>                                                      630.91 716.91
## - telephone                                             1   633.46 717.46
## - savings.account.and.bonds                             4   640.78 718.78
## - foreign.worker                                        1   634.93 718.93
```

```
## - other.debtors.or.guarantors                                    2   637.87 719.87
## - installment.rate.in.percentage.of.disposable.income            1   636.80 720.80
## - other.installment.plans                                        2   640.91 722.91
## - credit.history                                                 4   644.94 722.94
## - duration.in.month                                              1   639.46 723.46
## - credit.amount                                                  1   640.53 724.53
## - purpose                                                        9   678.37 746.37
## - status.of.existing.checking.account                            3   674.44 754.44
##
## Step:  AIC=714.7
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
##     personal.status.and.sex + other.debtors.or.guarantors + present.residence.since +
##     age.in.years + other.installment.plans + number.of.existing.credits.at.this.bank
##     number.of.people.being.liable.to.provide.maintenance.for +
##     telephone + foreign.worker
##
##                                                       Df Deviance    AIC
## - personal.status.and.sex                              3   636.65 712.65
## - number.of.people.being.liable.to.provide.maintenance.for  1   632.78 712.78
## - present.residence.since                              1   632.79 712.79
## - number.of.existing.credits.at.this.bank              1   632.85 712.85
## - age.in.years                                         1   633.92 713.92
## - present.employment.since                             4   640.58 714.58
## <none>                                                     632.70 714.70
## - telephone                                            1   635.19 715.19
## - savings.account.and.bonds                            4   642.24 716.24
## - foreign.worker                                       1   636.34 716.34
## - other.debtors.or.guarantors                          2   639.74 717.74
## - installment.rate.in.percentage.of.disposable.income  1   638.19 718.19
## - other.installment.plans                              2   643.05 721.05
## - duration.in.month                                    1   641.41 721.41
## - credit.history                                       4   647.53 721.53
## - credit.amount                                        1   642.05 722.05
## - purpose                                              9   679.91 743.91
## - status.of.existing.checking.account                  3   676.71 752.71
##
## Step:  AIC=712.65
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
##     other.debtors.or.guarantors + present.residence.since + age.in.years +
##     other.installment.plans + number.of.existing.credits.at.this.bank +
##     number.of.people.being.liable.to.provide.maintenance.for +
##     telephone + foreign.worker
##
##                                                       Df Deviance    AIC
```

```
## - number.of.people.being.liable.to.provide.maintenance.for  1   636.70 710.70
## - present.residence.since                                   1   636.75 710.75
## - number.of.existing.credits.at.this.bank                   1   636.84 710.84
## - present.employment.since                                  4   643.88 711.88
## - age.in.years                                              1   638.08 712.08
## <none>                                                          636.65 712.65
## - telephone                                                 1   639.05 713.05
## - savings.account.and.bonds                                 4   645.82 713.82
## - foreign.worker                                            1   640.46 714.46
## - other.debtors.or.guarantors                               2   643.50 715.50
## - installment.rate.in.percentage.of.disposable.income       1   642.55 716.55
## - duration.in.month                                         1   644.75 718.75
## - other.installment.plans                                   2   647.03 719.03
## - credit.amount                                             1   645.71 719.71
## - credit.history                                            4   652.35 720.35
## - purpose                                                   9   684.00 742.00
## - status.of.existing.checking.account                       3   679.69 749.69
##
## Step:  AIC=710.7
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
##     other.debtors.or.guarantors + present.residence.since + age.in.years +
##     other.installment.plans + number.of.existing.credits.at.this.bank +
##     telephone + foreign.worker
##
##                                                       Df Deviance    AIC
## - present.residence.since                              1   636.80 708.80
## - number.of.existing.credits.at.this.bank              1   636.92 708.92
## - present.employment.since                             4   643.89 709.89
## - age.in.years                                         1   638.09 710.09
## <none>                                                     636.70 710.70
## - telephone                                            1   639.09 711.09
## - savings.account.and.bonds                            4   645.84 711.84
## - foreign.worker                                       1   640.47 712.47
## - other.debtors.or.guarantors                          2   643.51 713.51
## - installment.rate.in.percentage.of.disposable.income  1   642.57 714.57
## - duration.in.month                                    1   644.79 716.79
## - other.installment.plans                              2   647.34 717.34
## - credit.amount                                        1   645.73 717.73
## - credit.history                                       4   653.00 719.00
## - purpose                                              9   684.29 740.29
## - status.of.existing.checking.account                  3   679.72 747.72
##
## Step:  AIC=708.8
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
```

```
##     other.debtors.or.guarantors + age.in.years + other.installment.plans +
##     number.of.existing.credits.at.this.bank + telephone + foreign.worker
##
##                                                 Df Deviance    AIC
## - number.of.existing.credits.at.this.bank         1   637.04 707.04
## - present.employment.since                        4   643.94 707.94
## - age.in.years                                    1   638.12 708.12
## <none>                                                636.80 708.80
## - telephone                                       1   639.15 709.15
## - savings.account.and.bonds                       4   645.89 709.89
## - foreign.worker                                  1   640.66 710.66
## - other.debtors.or.guarantors                     2   643.62 711.62
## - installment.rate.in.percentage.of.disposable.income  1   642.72 712.72
## - duration.in.month                               1   644.90 714.90
## - other.installment.plans                         2   647.38 715.38
## - credit.amount                                   1   645.81 715.81
## - credit.history                                  4   653.08 717.08
## - purpose                                         9   684.29 738.29
## - status.of.existing.checking.account             3   679.92 745.92
##
## Step:  AIC=707.04
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     present.employment.since + installment.rate.in.percentage.of.disposable.income +
##     other.debtors.or.guarantors + age.in.years + other.installment.plans +
##     telephone + foreign.worker
##
##                                                 Df Deviance    AIC
## - present.employment.since                        4   644.06 706.06
## - age.in.years                                    1   638.29 706.29
## <none>                                                637.04 707.04
## - telephone                                       1   639.32 707.32
## - savings.account.and.bonds                       4   646.14 708.14
## - foreign.worker                                  1   640.95 708.95
## - other.debtors.or.guarantors                     2   643.87 709.87
## - installment.rate.in.percentage.of.disposable.income  1   642.86 710.86
## - duration.in.month                               1   645.06 713.06
## - other.installment.plans                         2   647.54 713.54
## - credit.amount                                   1   646.05 714.05
## - credit.history                                  4   653.69 715.69
## - purpose                                         9   684.52 736.52
## - status.of.existing.checking.account             3   679.97 743.97
##
## Step:  AIC=706.06
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     installment.rate.in.percentage.of.disposable.income + other.debtors.or.guarantors
##     age.in.years + other.installment.plans + telephone + foreign.worker
```

```
##
##                                                    Df Deviance    AIC
## - age.in.years                                      1   645.15 705.15
## <none>                                                  644.06 706.06
## - telephone                                         1   646.16 706.16
## - foreign.worker                                    1   647.68 707.68
## - savings.account.and.bonds                         4   654.72 708.72
## - installment.rate.in.percentage.of.disposable.income  1   650.11 710.11
## - duration.in.month                                 1   650.26 710.26
## - other.debtors.or.guarantors                       2   652.31 710.31
## - other.installment.plans                           2   654.29 712.29
## - credit.amount                                     1   653.78 713.78
## - credit.history                                    4   661.90 715.90
## - purpose                                           9   688.97 732.97
## - status.of.existing.checking.account               3   689.02 745.02
##
## Step:  AIC=705.15
## creditability ~ status.of.existing.checking.account + duration.in.month +
##     credit.history + purpose + credit.amount + savings.account.and.bonds +
##     installment.rate.in.percentage.of.disposable.income + other.debtors.or.guarantors
##     other.installment.plans + telephone + foreign.worker
##
##                                                    Df Deviance    AIC
## <none>                                                  645.15 705.15
## - telephone                                         1   647.87 705.87
## - foreign.worker                                    1   648.74 706.74
## - savings.account.and.bonds                         4   655.95 707.95
## - installment.rate.in.percentage.of.disposable.income  1   651.01 709.01
## - other.debtors.or.guarantors                       2   653.44 709.44
## - duration.in.month                                 1   651.61 709.61
## - other.installment.plans                           2   655.09 711.09
## - credit.amount                                     1   654.69 712.69
## - credit.history                                    4   663.77 715.77
## - purpose                                           9   690.34 732.34
## - status.of.existing.checking.account               3   690.98 744.98
```

Wybieramy model z najmniejszym AIC = 705.15. W ten sposób otrzymujemy model

$$\text{creditability} = \text{status.of.existing.checking.account} + \text{duration.in.month} +$$
$$+ \text{credit.history} + \text{purpose} + \text{credit.amount} + \text{savings.account.and.bonds} +$$
$$+ \text{installment.rate.in.percentage.of.disposable.income} + \text{other.debtors.or.guarantors} +$$
$$+ \text{other.installment.plans} + \text{telephone} + \text{foreign.worker}.$$

Model posiada aż 11 zmiennych objaśniających. Sprawdzając dopasowanie modelu do danych na zbiorze testowym może okazać się, że jest on przeuczony.

Macierz pomyłek

W oparciu o wybrany model dla danych ze zbioru testowego obliczymy prawdopodobieństwa sukcesu oraz prognozowane wartości zmiennej *creditability*. Jako sukces przyjmujemy prawdopodobieństwo, że klient nie spłaci kredytu, czyli wartość zmiennej *creditability* = 1. Wyznaczymy także macierz pomyłek, dokładność (accuracy), błąd, precyzję przewidywania pozytywnego (Positive Prediction Value), precyzję przewidywania negatywnego (Negative Predicion Value), czułość (sensivity, in. True Positive Rate) oraz specyficzność (specificity, in. False Positive Rate).

Poniżej wyjaśnimy kryteria oceny modelu w oparciu o macierz pomyłek. Najpierw wprowadzimy oznaczenia

$$N = TN + FP,$$

$$P = FN + TP,$$

$$N^* = TN + FN,$$

$$P^* = FP + TP.$$

Dokładność (accuracy - ACC) oznacza liczbę obserwacji sklasyfikowanych poprawnie podzieloną przez liczbę wszystkich obserwacji.

$$ACC = \frac{TP + TN}{N + P}$$

Błąd możemy obliczyć ze wzoru $1 - ACC$.

Precyzja przewidywania pozytywnego (Positive Prediction Value - PPV) mierzy proporcję prawdziwie pozytywnych klasyfikacji względem wszystkich pozytywnych klasyfikacji.

$$PPV = \frac{TP}{P^*}$$

Precyzja przewidywania negatywnego (Negative Prediction Value – NPV) to stosunek liczby przypadków prawdziwie negatywnie sklasyfikowanych do wszystkich negatywnych klasyfikacji.

$$NPV = \frac{TN}{N^*}$$

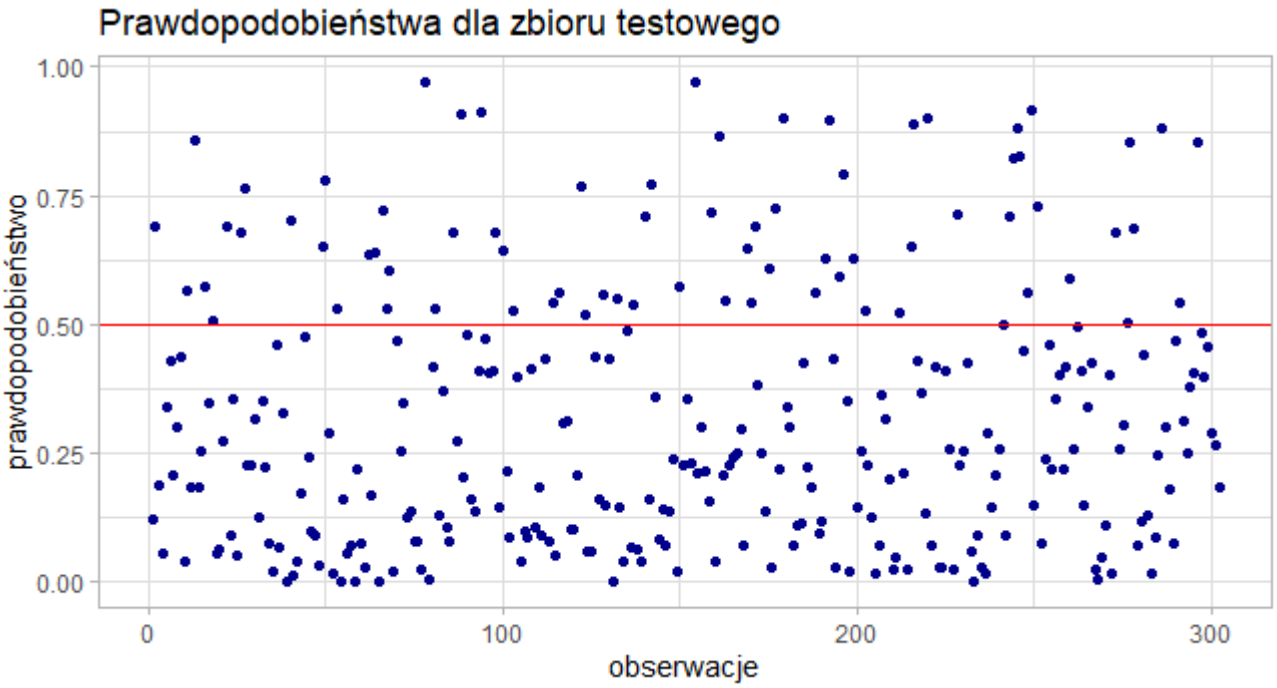
Czułość (sensitivity, True Positive Rate - TPR) to proporcja liczby poprawnych pozytywnych klasyfikacji względem liczby wszystkich (prawdziwie) pozytywnych przypadków.

$$TPR = \frac{TP}{P}$$

Specyficzność (specifity, False Positive Rate - FPR) to liczba prawdziwie negatywnych klasyfikacji względem wszystkich (prawdziwie) negatywnych przypadków.

$$FPR = \frac{TN}{N}$$

```
library(caret)
pred1 <- predict(model1_AIC, test_set, type = "response")
```



Rysunek 1: Prawdopodobieństwa dla zbioru testowego dla modelu 1

Możemy zauważyć, że otrzymaliśmy więcej wartości poniżej 0.5. Przypomnijmy, że sukcesem jest niespłacenie kredytu przez klienta, zatem jest większe prawdopodobieństwo, że kredyt zostanie spłacony.

Mając oszacowane prawdopodobieństwa wyznaczymy macierz pomyłek. Aby to zrobić skorzystamy z funkcji *confusionMatrix* z pakietu **caret**. Jako punkt odcięcia przyjmujemy wartość 0.5, to znaczy reguła klasyfikacyjna jest postaci

$$d(x) = \begin{cases} 1, & gdy \quad p(x) > 0.5, \\ 0, & gdy \quad p(x) \leq 0.5. \end{cases}$$

Oznacza to, że jeśli prawdopodobieństwo > 0.5, klasyfikujemy klienta jako "złego" (niespłacającego kredytu), więc przypisujemy mu wartość 1. Natomiast dla prawdopodobieństwa ≤ 0.5 przyjmujemy, że klient jest "dobry", przypisując mu wartość 0.

```
pred_labels1 <- ifelse(pred1 > 0.50, 1, 0)
pred_labels1 <- as.factor(pred_labels1)

real_labels1 <- test_set$creditability
real_labels1 <- as.factor(real_labels1)

conf_matrix1 <- confusionMatrix(pred_labels1, real_labels1, positive = "1")
print(conf_matrix1)

## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction   0   1
##          0 186  44
##          1  29  43
##
##                Accuracy : 0.7583
##                  95% CI : (0.7059, 0.8055)
##     No Information Rate : 0.7119
##     P-Value [Acc > NIR] : 0.04139
##
##                   Kappa : 0.3788
##
##  Mcnemar's Test P-Value : 0.10130
##
##             Sensitivity : 0.4943
##             Specificity : 0.8651
##          Pos Pred Value : 0.5972
##          Neg Pred Value : 0.8087
##              Prevalence : 0.2881
##          Detection Rate : 0.1424
##    Detection Prevalence : 0.2384
##       Balanced Accuracy : 0.6797
##
##        'Positive' Class : 1
##
```

```
accuracy1 <- conf_matrix1[["overall"]][["Accuracy"]]
print(accuracy1)

## [1] 0.7582781

blad1 <- 1 - conf_matrix1[["overall"]][["Accuracy"]]
print(blad1)

## [1] 0.2417219

positive_prediction_value1 <- conf_matrix1[["byClass"]][["Pos Pred Value"]]
print(positive_prediction_value1)

## [1] 0.5972222

negative_prediction_value1 <- conf_matrix1[["byClass"]][["Neg Pred Value"]]
print(negative_prediction_value1)

## [1] 0.8086957

sensitivity1 <- conf_matrix1[["byClass"]][["Sensitivity"]]
print(sensitivity1)
```

```
## [1] 0.4942529

specificity1 <- conf_matrix1[["byClass"]][["Specificity"]]
print(specificity1)

## [1] 0.8651163
```

Prawdopodobieństwo poprawnej klasyfikacji wynosi 0.7582781. Model prawidłowo klasyfikuje klientów dobrych z prawdopodobieństwem 0.8651163, natomiast klientów złych prawidłowo klasyfikuje z prawdopodobieństwem 0.4942529.

Część 3

Konstrukcja modeli na podstawie WoE oraz IV

W tej części sprawozdania wykorzystamy WoE (Weight of Evidence) oraz IV (Information Value) do budowy modeli scoringowych.

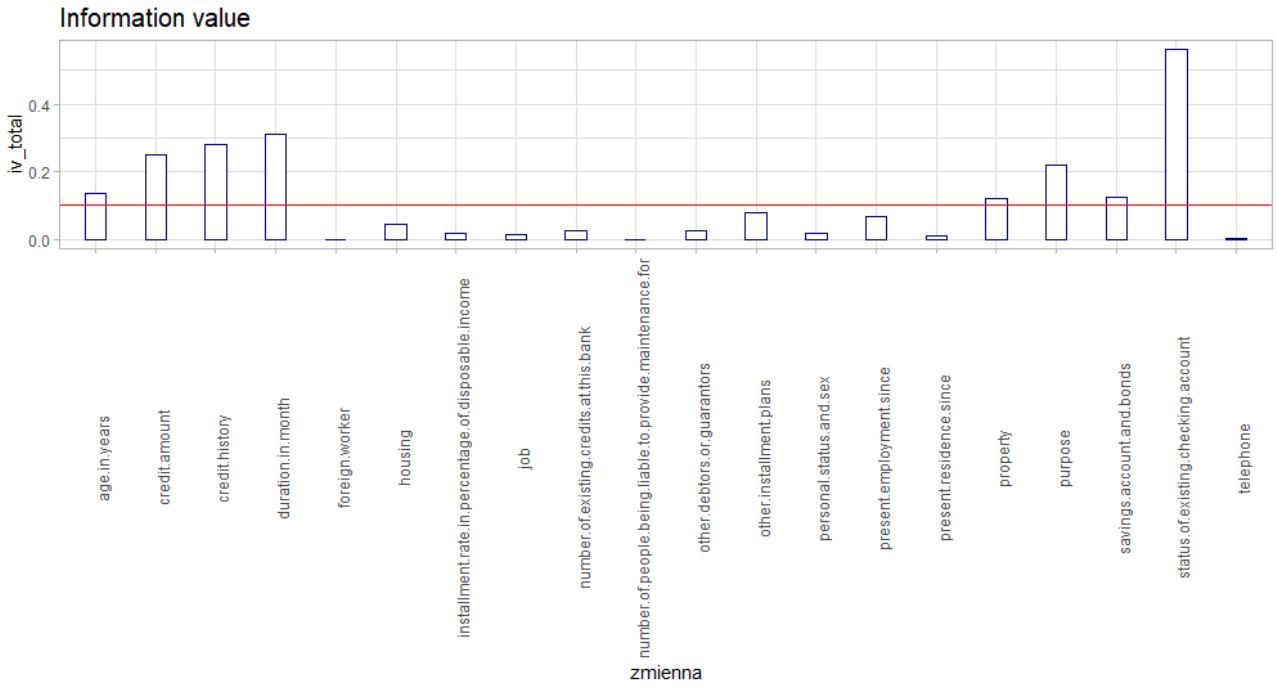
Najpierw skorzystamy z IV. W wyborze zmiennych tą metodą stosuje się tzw. regułę kciuka. Przyjmuje się, że jeżeli

- IV < 0.02, to zmienna X nie jest zmienną predykcyjną,
- IV $\in [0.02, 0.1)$, to zmienna X ma słabą moc predykcyjną,
- IV $\in [0.1, 0.3)$, to zmienna X ma średnią moc predykcyjną,
- IV ≥ 0.3, to zmienna X ma dużą moc predykcyjną.

```
woebin_train_set <- woebin(train_set, y = "creditability", positive = "1")

## [INFO] creating woe binning ...
```

Za zmienne istotne w budowie modelu przyjmujemy zmienne z IV ≥ 0.1, czyli ze średnią i dużą mocą predykcyjną.

```
namesofvariables <- c()
total_iv_values <-c()
qualityofprediction <- c()
for (i in 1:(length(woebin_train_set))){
  namesofvariables[i] <- unique(woebin_train_set[[i]]$variable)
  total_iv_values[i] <- unique(woebin_train_set[[i]]$total_iv)
  if (total_iv_values[i] > 0.1){
    qualityofprediction[i] <- 1
  } else {
    qualityofprediction[i] <- 0
  }
}
indexesofvariables_iv <- which(qualityofprediction==1)
namesofvariables_iv <- namesofvariables[indexesofvariables_iv]
print(namesofvariables_iv)

## [1] "status.of.existing.checking.account" "duration.in.month"
## [3] "credit.history"                      "purpose"
## [5] "credit.amount"                       "savings.account.and.bonds"
## [7] "property"                            "age.in.years"
```



Rysunek 2: Wartości Information Value dla poszczególnych zmiennych

Zmiennymi, które są istotne w budowie modelu na podstawie Information Value okazały się "status.of.existing.checking.account", "duration.in.month", "credit.history", "purpose", "credit.amount", "savings.account.and.bonds", "property" oraz "age.in.years". Wykorzystując te zmienne dokonamy podziału na zbiór uczący i testowy dla modelu 2.

```
train_set_iv <- train_set[,c("status.of.existing.checking.account",
                             "duration.in.month",
                             "credit.history",
                             "purpose",
                             "credit.amount",
                             "savings.account.and.bonds",
                             "property",
                             "age.in.years",
                             "creditability")]

test_set_iv <- test_set[,c("status.of.existing.checking.account",
                           "duration.in.month",
                           "credit.history",
                           "purpose",
                           "credit.amount",
                           "savings.account.and.bonds",
                           "property",
                           "age.in.years",
                           "creditability")]

woebin_train_set_iv <- woebin(train_set_iv, "creditability", positive = "1")

## [INFO] creating woe binning ...
```

```r
woebin_test_set_iv <- woebin(test_set_iv, "creditability", positive = "1")

## [INFO] creating woe binning ...

binned_train_iv <- woebin_ply(train_set_iv, woebin_train_set_iv)

## [INFO] converting into woe values ...

binned_test_iv <- woebin_ply(test_set_iv, woebin_test_set_iv)

## [INFO] converting into woe values ...

model2 <- glm(creditability~., data = binned_train_iv,
              family = binomial(link = "logit"))
```

```r
print(model2)

##
## Call:  glm(formula = creditability ~ ., family = binomial(link = "logit"),
##     data = binned_train_iv)
##
## Coefficients:
##                               (Intercept)
##                                   -0.8054
## status.of.existing.checking.account_woe
##                                    0.7925
##                   duration.in.month_woe
##                                    0.7189
##                      credit.history_woe
##                                    0.7762
##                                purpose_woe
##                                    1.1058
##                       credit.amount_woe
##                                    0.7479
##           savings.account.and.bonds_woe
##                                    0.6148
##                              property_woe
##                                    0.5743
##                         age.in.years_woe
##                                    0.8781
##
## Degrees of Freedom: 697 Total (i.e. Null);  689 Residual
## Null Deviance:     858.8
## Residual Deviance: 657.2  AIC: 675.2
```

Otrzymany model jest postaci

$$\text{creditability} = \text{status.of.existing.checking.account\_woe} +$$
$$+ \text{duration.in.month\_woe} + \text{credit.history\_woe} + \text{purpose\_woe} + \text{credit.amount\_woe} +$$
$$+ \text{savings.account.and.bonds\_woe} + \text{property\_woe} + \text{age.in.years\_woe}.$$

AIC tego modelu wynosi 675.2. Model zawiera 8 zmiennych objaśnianych, wydaje się że jest to optymalna liczba predyktorów. AIC jest niższe niż w przypadku modelu 1.

Teraz skonstruujemy model, wybierając tylko zmienne monotoniczne spośród tych zmiennych, których $IV \geq 0.1$.

```
monotonicity <- c()
for (i in indexesofvariables_iv){
  if(identical(sort(woebin_train_set[[i]]$woe, decreasing = TRUE),
               as.vector(woebin_train_set[[i]]$woe))|
     identical(sort(woebin_train_set[[i]]$woe, decreasing = FALSE),
               as.vector(woebin_train_set[[i]]$woe))){
    monotonicity <- append(monotonicity,1)
  } else {
    monotonicity <- append(monotonicity,0)
  }
}
indexesofvariables_monotonic_woe <- which(monotonicity==1)
namesofvariables_monotonic_woe <-
  namesofvariables_iv[indexesofvariables_monotonic_woe]
print(namesofvariables_monotonic_woe)

## [1] "status.of.existing.checking.account" "duration.in.month"
## [3] "property"
```

Zmiennymi monotonicznymi okazały się ”status.of.existing.checking.account”, ”duration.in.month” oraz ”property”. Na ich podstawie stworzymy zbiór uczący oraz zbiór testowy dla modelu 3.

```
train_set_woe <- train_set[,c("status.of.existing.checking.account",
                              "duration.in.month",
                              "property",
                              "creditability")]

test_set_woe <- test_set[,c("status.of.existing.checking.account",
                            "duration.in.month",
                            "property",
                            "creditability")]

woebin_train_set_woe <- woebin(train_set_woe, "creditability", positive = "1")

## [INFO] creating woe binning ...

woebin_test_set_woe <- woebin(test_set_woe, "creditability", positive = "1")
```

```
## [INFO] creating woe binning ...

binned_train_woe <- woebin_ply(train_set_woe,woebin_train_set_woe)

## [INFO] converting into woe values ...

binned_test_woe <- woebin_ply(test_set_woe,woebin_test_set_woe)

## [INFO] converting into woe values ...

model3 <- glm(creditability~., data = binned_train_woe,
              family = binomial(link = "logit"))
```

```
print(model3)

##
## Call:  glm(formula = creditability ~ ., family = binomial(link = "logit"),
##     data = binned_train_woe)
##
## Coefficients:
##                              (Intercept)
##                                  -0.8174
## status.of.existing.checking.account_woe
##                                   0.9737
##                   duration.in.month_woe
##                                   0.8756
##                             property_woe
##                                   0.6523
##
## Degrees of Freedom: 697 Total (i.e. Null);  694 Residual
## Null Deviance:     858.8
## Residual Deviance: 736.9  AIC: 744.9
```

Otrzymujemy model postaci

$$\text{creditability} = \text{status.of.existing.checking.account\_woe} +$$

$$+ \text{duration.in.month\_woe} + \text{property\_woe}.$$

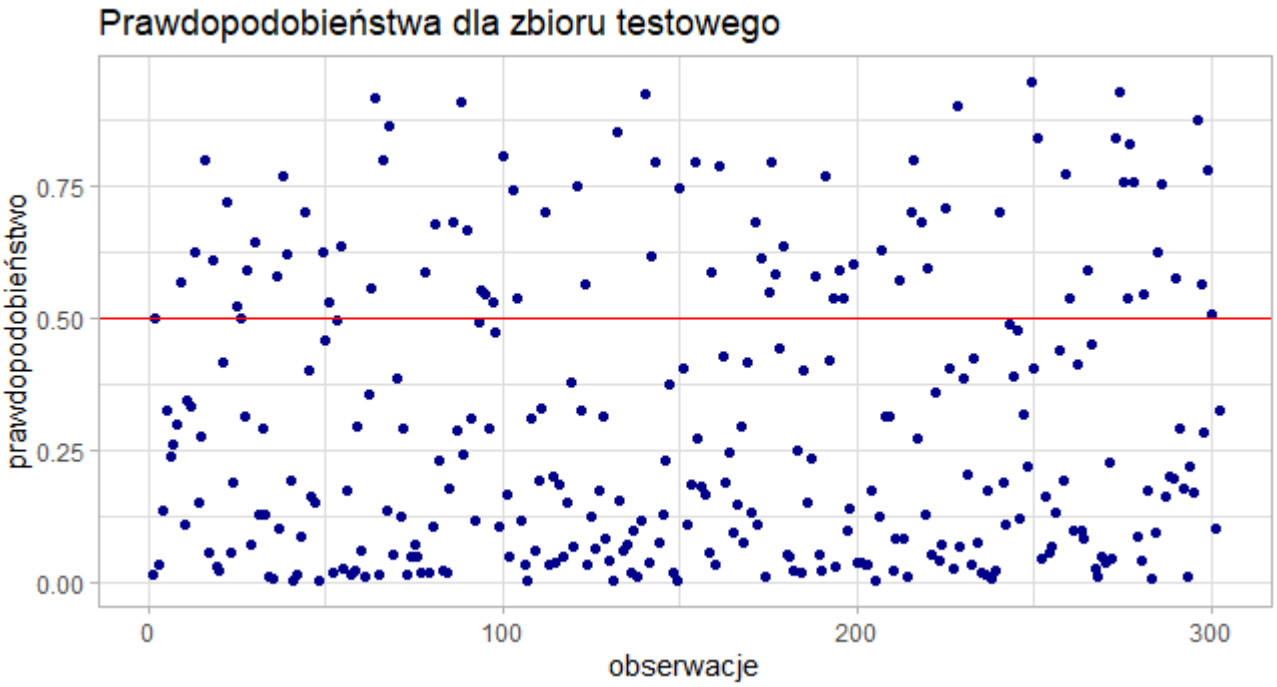
Powyższy model zawiera tylko 3 zmienne objaśniane. AIC tego modelu wynosi 744.9. Jest najwyższe spośród wszystkich rozpatrywanych modeli, zatem wypada najgorzej.

Porównanie modeli

W celu porównania oceny efektywności modeli scoringowych, skorzystamy z counting methods opartych na macierzy pomyłek opisanych w Części 1 oraz separability measures, takich jak krzywa ROC, AUC oraz współczynnik Giniego.

Najpierw wyznaczymy counting methods dla modelu 2.

```
model2_AIC <- step(model2, direction = "backward", trace = FALSE)
pred2 <- predict(model2_AIC, binned_test_iv, type = "response")
```



Rysunek 3: Prawdopodobieństwa dla zbioru testowego dla modelu 2

```
pred_labels2 <- ifelse(pred2 > 0.50, 1, 0)
pred_labels2 <- as.factor(pred_labels2)

real_labels2 <- test_set$creditability
real_labels2 <- as.factor(real_labels2)

conf_matrix2 <- confusionMatrix(pred_labels2, real_labels2, positive = "1")
print(conf_matrix2)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 188  32
##          1  27  55
##
```

```
##                Accuracy : 0.8046
##                  95% CI : (0.7554, 0.8478)
##     No Information Rate : 0.7119
##     P-Value [Acc > NIR] : 0.0001485
##
##                   Kappa : 0.5154
##
##  Mcnemar's Test P-Value : 0.6025370
##
##             Sensitivity : 0.6322
##             Specificity : 0.8744
##          Pos Pred Value : 0.6707
##          Neg Pred Value : 0.8545
##              Prevalence : 0.2881
##          Detection Rate : 0.1821
##    Detection Prevalence : 0.2715
##       Balanced Accuracy : 0.7533
##
##        'Positive' Class : 1
##
```

```
accuracy2 <- conf_matrix2[["overall"]][["Accuracy"]]
print(accuracy2)
```

```
## [1] 0.8046358
```

```
blad2 <- 1 - conf_matrix2[["overall"]][["Accuracy"]]
print(blad2)
```

```
## [1] 0.1953642
```

```
positive_prediction_value2 <- conf_matrix2[["byClass"]][["Pos Pred Value"]]
print(positive_prediction_value2)
```

```
## [1] 0.6707317
```

```
negative_prediction_value2 <- conf_matrix2[["byClass"]][["Neg Pred Value"]]
print(negative_prediction_value2)
```

```
## [1] 0.8545455
```

```
sensitivity2 <- conf_matrix2[["byClass"]][["Sensitivity"]]
print(sensitivity2)
```

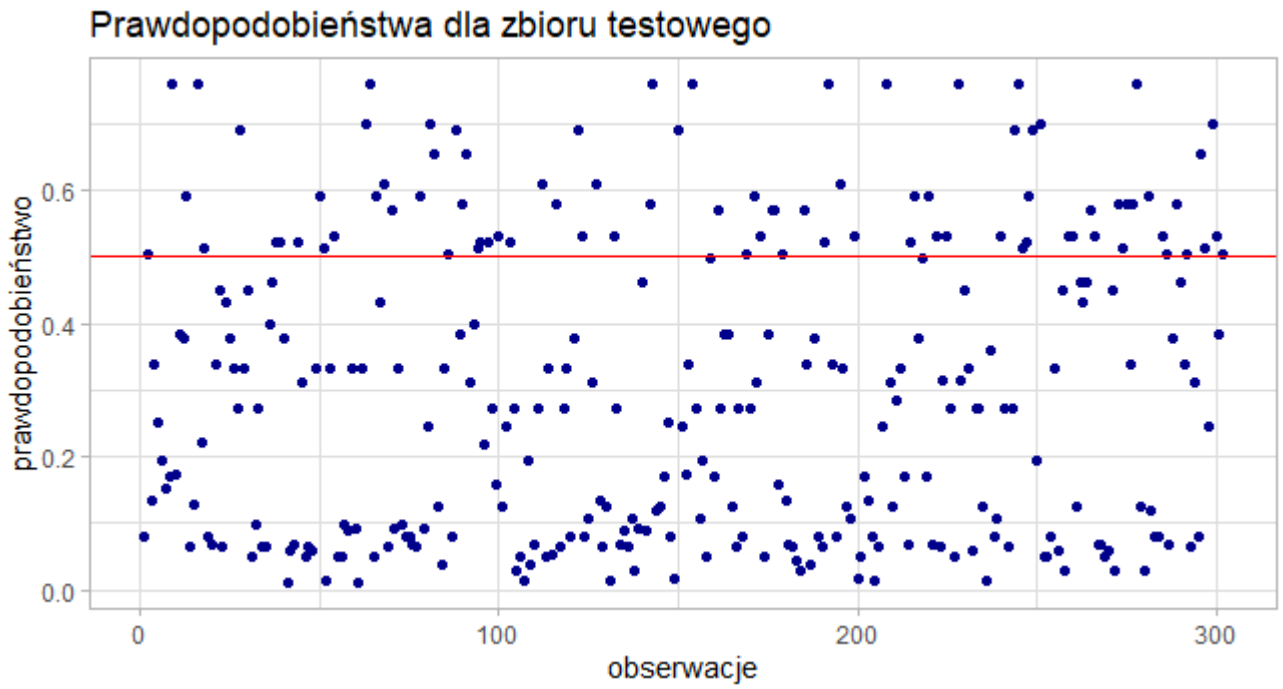
```
## [1] 0.6321839
```

```
specificity2 <- conf_matrix2[["byClass"]][["Specificity"]]
print(specificity2)
```

```
## [1] 0.8744186
```

Teraz wyznaczymy counting methods dla modelu 3.

```
model3_AIC <- step(model3, direction = "backward", trace = FALSE)
pred3 <- predict(model3_AIC, binned_test_woe, type = "response")
```



Rysunek 4: Prawdopodobieństwa dla zbioru testowego dla modelu 3

```
pred_labels3 <- ifelse(pred3 > 0.50, 1, 0)
pred_labels3 <- as.factor(pred_labels3)

real_labels3 <- test_set$creditability
real_labels3 <- as.factor(real_labels3)

conf_matrix3 <- confusionMatrix(pred_labels3, real_labels3, positive = "1")
print(conf_matrix3)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 181  36
##          1  34  51
##
##                Accuracy : 0.7682
##                  95% CI : (0.7165, 0.8146)
##     No Information Rate : 0.7119
##     P-Value [Acc > NIR] : 0.01654
```

```
##
##                  Kappa : 0.431
##
##  Mcnemar's Test P-Value : 0.90486
##
##             Sensitivity : 0.5862
##             Specificity : 0.8419
##          Pos Pred Value : 0.6000
##          Neg Pred Value : 0.8341
##              Prevalence : 0.2881
##          Detection Rate : 0.1689
##    Detection Prevalence : 0.2815
##       Balanced Accuracy : 0.7140
##
##        'Positive' Class : 1
##
```

```
accuracy3 <- conf_matrix3[["overall"]][["Accuracy"]]
print(accuracy3)
```

```
## [1] 0.7682119
```

```
blad3 <- 1 - conf_matrix3[["overall"]][["Accuracy"]]
print(blad3)
```

```
## [1] 0.2317881
```

```
positive_prediction_value3 <- conf_matrix3[["byClass"]][["Pos Pred Value"]]
print(positive_prediction_value3)
```

```
## [1] 0.6
```

```
negative_prediction_value3 <- conf_matrix3[["byClass"]][["Neg Pred Value"]]
print(negative_prediction_value3)
```

```
## [1] 0.8341014
```

```
sensitivity3 <- conf_matrix3[["byClass"]][["Sensitivity"]]
print(sensitivity3)
```

```
## [1] 0.5862069
```

```
specificity3 <- conf_matrix3[["byClass"]][["Specificity"]]
print(specificity3)
```

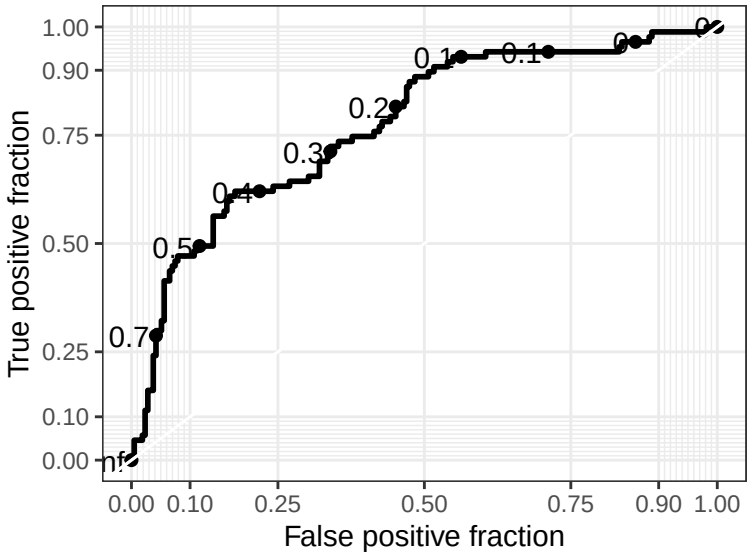
```
## [1] 0.8418605
```

Wyznaczymy jeszcze BIC dla rozpatrywanych modeli.

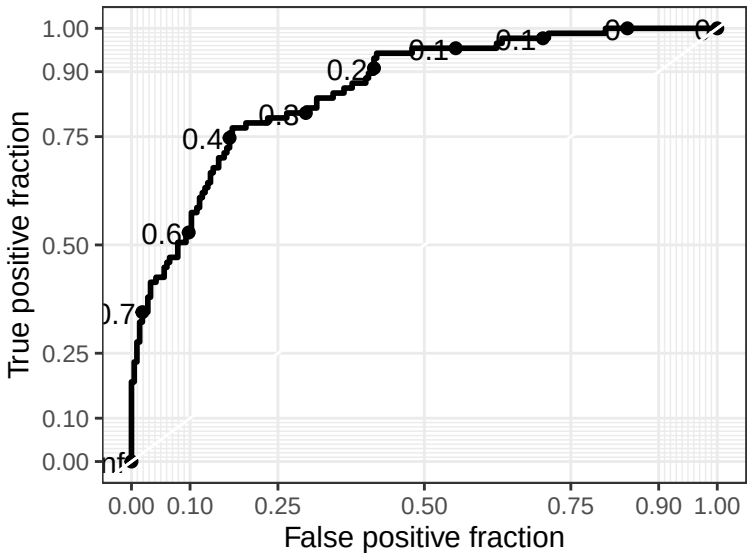
```
BIC(model1_AIC)

## [1] 841.5994

BIC(model2)

## [1] 716.1584

BIC(model3)

## [1] 763.0874
```

Poniżej przedstawiono krzywe ROC dla wszystkich 3 modeli.

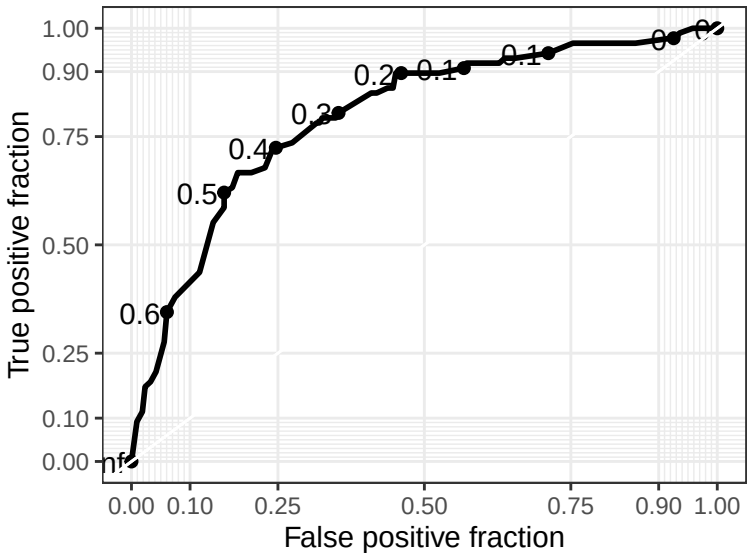
```
library(plotROC)
roc.estimate1 <- calculate_roc(pred1, real_labels1)
single.rocplot1 <- ggroc(roc.estimate1)
plot_journal_roc(single.rocplot1)
```

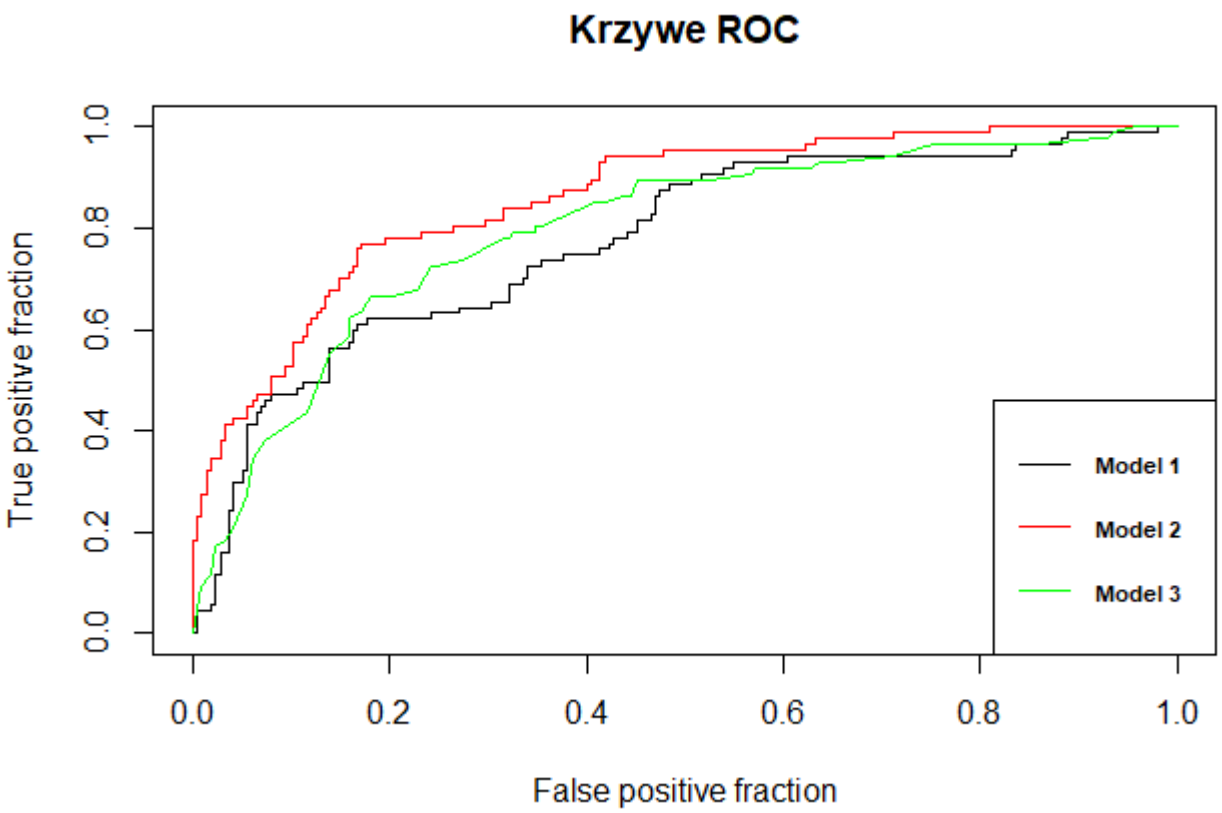



```
roc.estimate2 <- calculate_roc(pred2, real_labels2)
single.rocplot2 <- ggroc(roc.estimate2)
plot_journal_roc(single.rocplot2)
```




```
roc.estimate3 <- calculate_roc(pred3, real_labels3)
single.rocplot3 <- ggroc(roc.estimate3)
plot_journal_roc(single.rocplot3)
```





Rysunek 5: Krzywe ROC dla wszystkich 3 modeli

Zdecydowanie najlepiej wypada model 2, ponieważ krzywa ROC dla tego modelu znajduje się powyżej pozostałych dwóch krzywych. Na tej podstawie można już uznać model 2 za najlepszy, ale obliczymy jeszcze AUC oraz współczynnik Giniego dla wszystkich 3 modeli.

Podstawowym indeksem związanym z krzywą ROC jest AUC (ang. Area Under Curve) określony wzorem

$$AUC = \int_0^1 R(t)dt,$$

gdzie $R(t) = 1 - G(F^{-1}(1-t))$, przy czym F^{-1} oznacza dystrybuantę odwrotną do dystrybuanty F na zbiorze $[0, 1]$.

```
AUC1 <- abs(calc_auc(single.rocplot1)$AUC)
print(AUC1)
```

```
## [1] 0.7744453
```

```
AUC2 <- abs(calc_auc(single.rocplot2)$AUC)
print(AUC2)
```

```
## [1] 0.8587009
```

```
AUC3 <- abs(calc_auc(single.rocplot3)$AUC)
print(AUC3)
```

```
## [1] 0.7962577
```

Współczynnik Giniego wyznaczamy ze wzoru

$$G = 2AUC - 1.$$

```
Gini1 <- 2*AUC1-1
print(Gini1)

## [1] 0.5488907

Gini2 <- 2*AUC2-1
print(Gini2)

## [1] 0.7174018

Gini3 <- 2*AUC3-1
print(Gini3)

## [1] 0.5925154
```

Tabela 1: Porównanie modeli scoringowych

	model 1	model 2	model 3
dokładność	0.7582781	0.8046358	0.7682119
błąd	0.2417219	0.1953642	0.2317881
precyzja przewidywania pozytywnego	0.5972222	0.6707317	0.6000000
precyzja przewidywania negatywnego	0.8086957	0.8545455	0.8341014
czułość	0.4942529	0.6321839	0.5862069
specyficzność	0.8651163	0.8744186	0.8418605
AIC	705.15000	675.20000	744.90000
BIC	841.59940	716.15840	763.08740
AUC	0.7744453	0.8587009	0.7962577
współczynnik Giniego	0.5488907	0.7174018	0.5925154

Podsumowując wyniki z powyższej tabeli możemy jednoznacznie stwierdzić, że model 2 jest najlepiej dopasowany do danych. Tak jak już było wspomniane wcześniej, model 1 posiada najwięcej zmiennych objaśnianych spośród wszystkich rozważanych modeli, zatem po sprawdzeniu go na zbiorze testowym osiąga najgorsze oszacowania zarówno dla counting methods jak i separability measures.

Część 4

Dla każdego z 3 modeli uzyskanych w poprzednich zadaniach wyznaczymy optymalne punkty odcięcia (cutoffs). Skorzystamy z funkcji *optimalCutoff* z pakietu **InformationValue**.

```
library(InformationValue)
cutoff1_model1 <- optimalCutoff(real_labels1, pred1, "Ones")
cutoff2_model1 <- optimalCutoff(real_labels1, pred1, "Zeros")
cutoff3_model1 <- optimalCutoff(real_labels1, pred1, "Both")
cutoff4_model1 <- optimalCutoff(real_labels1, pred1, "misclasserror")

cutoff1_model2 <- optimalCutoff(real_labels2, pred2, "Ones")
cutoff2_model2 <- optimalCutoff(real_labels2, pred2, "Zeros")
cutoff3_model2 <- optimalCutoff(real_labels2, pred2, "Both")
cutoff4_model2 <- optimalCutoff(real_labels2, pred2, "misclasserror")

cutoff1_model3 <- optimalCutoff(real_labels3, pred3, "Ones")
cutoff2_model3 <- optimalCutoff(real_labels3, pred3, "Zeros")
cutoff3_model3 <- optimalCutoff(real_labels3, pred3, "Both")
cutoff4_model3 <- optimalCutoff(real_labels3, pred3, "misclasserror")

cutoff_model1 <- rbind(cutoff1_model1, cutoff2_model1,
                       cutoff3_model1, cutoff4_model1)
cutoff_model2 <- rbind(cutoff1_model2, cutoff2_model2,
                       cutoff3_model2, cutoff4_model2)
cutoff_model3 <- rbind(cutoff1_model3, cutoff2_model1,
                       cutoff3_model3, cutoff4_model3)

optimal_cutoff <- cbind(cutoff_model1, cutoff_model2, cutoff_model3)

optimal_cutoff <- as.data.frame(optimal_cutoff)
colnames(optimal_cutoff) <- c("Model 1", "Model 2", "Model 3")
rownames(optimal_cutoff) <- c("Ones", "Zeros", "Both", "misclasserror")
```

```
print(optimal_cutoff)

##                  Model 1    Model 2   Model 3
## Ones          0.02230802 0.02929409 0.0207784
## Zeros         0.90230802 0.79929409 0.9023080
## Both          0.43230802 0.38929409 0.4407784
## misclasserror 0.56230802 0.38929409 0.4407784
```