STATYSTYKA W FINANSACH I UBEZPIECZENIACH
LABORATORIUM

SPRAWOZDANIE

OPRACOWAŁA:
ALEKSANDRA GRZESZCZUK
NUMER ALBUMU: 255707

SPIS TREŚCI

1 CEL

W raporcie przedstawiona zostanie analiza zbioru danych `germancredit`. W oparciu o ten zbiór skonstruowane zostaną różne modele scoringowe. Następnie zostaną ocenione i porównane

2 ZADANIE 1

Kiedy bank otrzymuje wniosek o pożyczkę, na podstawie profilu wnioskodawcy musi podjąć decyzję, czy zatwierdzić pożyczkę, czy nie. Z tą decyzją wiążą się dwa rodzaje ryzyka:

- jeśli wnioskodawca ma dobre ryzyko kredytowe, czyli jest prawdopodobne, że spłaci pożyczkę, wówczas brak zatwierdzenia pożyczki danej osobie skutkuje utratą działalności dla banku
- jeśli wnioskodawca jest obarczony złym ryzykiem kredytowym, czyli nie jest prawdopodobne, aby spłacił pożyczkę, wówczas zatwierdzenie pożyczki dla tej osoby skutkuje stratą finansową dla banku

Aby zminimalizować stratę z punktu widzenia banku, potrzebna jest reguła decyzyjna dotycząca tego, komu udzielić zgody na pożyczkę, a komu nie. Profile demograficzne i społeczno-ekonomiczne wnioskodawcy są brane pod uwagę przez zarządzających pożyczkami przed podjęciem decyzji w sprawie jego wniosku o pożyczkę.

Niemieckie dane kredytowe [1] (*The German Credit Data*) zawierają dane dotyczące 21 zmiennych i klasyfikację, czy wnioskodawca jest uznawany za dobre, czy złe ryzyko kredytowe dla 1000 osób ubiegających się o pożyczkę. Oczekuje się, że model predykcyjny opracowany na podstawie tych danych zapewni kierownikowi banku wskazówki dotyczące podejmowania decyzji o zatwierdzeniu pożyczki potencjalnemu wnioskodawcy na podstawie jej/jego profili.

W zbiorze danych jest łącznie 21 atrybutów. Ich opisy i szczegóły zestawiono poniżej:

- *Status of existing checking account* - zmienna typu $factor$ - status istniejących rachunków bankowych klienta
- *Duration in month* - zmienna typu *numeric* - czas trwania kredytu w miesiącach
- *Credit history* - zmienna typu $factor$ - historia kredytowa
- *Purpose* - zmienna typu *character* - cel, na który brany jest kredyt
- *Credit amount* - zmienna typu *numeric* - ilość pieniędzy wzięta na kredyt
- *Savings account and bonds* - zmienna typu $factor$ - konto oszczędnościowe i obligacje
- *Present employment since* - zmienna typu $factor$ - okres obecnego zatrudnienia

[1] Dane `germancredit` dostępne w bibliotece `scorecard` w pakiecie `R`. Można także znaleźć wiele opracowań zbioru tych danych w programie `Python` na przykład `https://www.kaggle.com/datasets/uciml/german-credit/code`

- *Installment rate in percentage of disposable income* - zmienna typu *numeric* - stopa raty jako procent dochodu do dyspozycji
- *Personal status and sex* - zmienna typu $factor$ - status osobisty i płeć
- *Other debtors or guarantors* - zmienna typu $factor$ - inni dłużnicy lub poręczyciele
- *Present residence since* - zmienna typu *numeric* - okres obecnego miejsca zamieszkania
- *Property* - zmienna typu $factor$ - co kredytobiorca posiada na własność
- *Age in years* - zmienna typu *numeric* - wiek w latach
- *Other installment plans* - zmienna typu $factor$ -
- *Housing* - zmienna typu $factor$ - nieruchomość, czy jest na własność, wynajmowana itd
- *Number of existing credits at this bank* - zmienna typu *numeric* - liczba istniejących kredytów w tym banku
- *Job* - zmienna typu $factor$ - czy jest to wykwalifikowany pracownik, niewykwalifikowany
- *Number of people being liable to provide maintenance for* - zmienna typu *numeric* - liczba osób zobowiązana do utrzymania kredytobiorcy
- *Telephone* - zmienna typu $factor$ - numer telefonu
- *Foreign worker* - zmienna typu $factor$ - czy jest to obcokrajowy pracownik
- *Creditability* - zmienna typu $factor$ - zdolność kredytowa - dobra lub zła

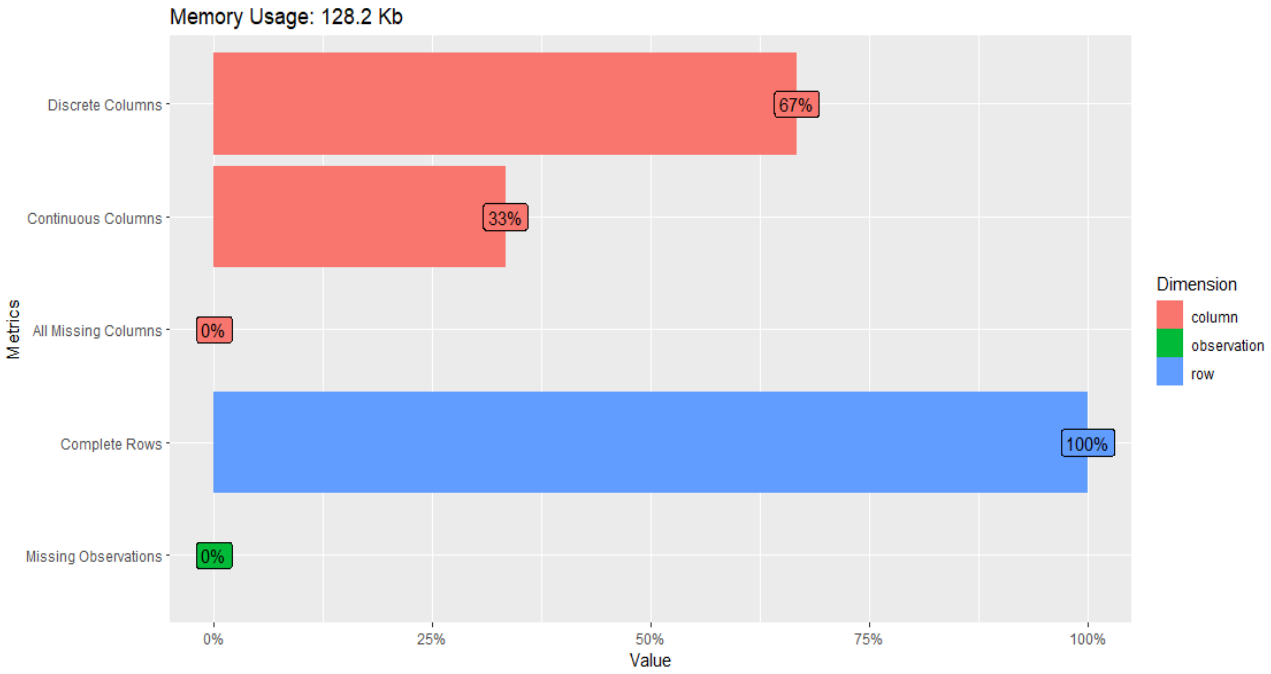
	Values
rows	1000.00
columns	21.00
discrete columns	14.00
continuous columns	7.00
all missing columns	0.00
total missing values	0.00
complete rows	1000.00
total observations	21000.00
memory usage	131320.00

Tabela 1: Podstawowe informacje dla danych `germancredit`

Dobrą wiadomością jest to, że w zbiorze danych nie ma brakujących wartości. Struktura wygląda na spójną.

Biblioteka `DataExplorer` [2] zajmuje się zautomatyzowanym procesem eksploracji danych na potrzeby zadań analitycznych oraz modelowania predykcyjnego. Dzięki temu można między innymi skupić się na rozumieniu danych. Funkcje dostępne w tym pakiecie skanują oraz analizują każdą zmienną, po czym je wizualizują za pomocą podstawowych, typowych technik graficznych.

Na podstawie funkcji `plot_into` dostępnej w bibliotece `DataExplorer` rysujemy partię podstawowych informacji dla danych wejściowych. Właściwie, większość informacji zawartych w tabeli (1) zostanie przedstawiona na poniższym wykresie (1).



Rysunek 1: Wykres podstawowych informacji dla danych `germancredit` - opracowanie własne

Niemal w każdym banku sprawdzenie kompletu dokumentacji do kredytu podzielone jest na cztery główne działy.

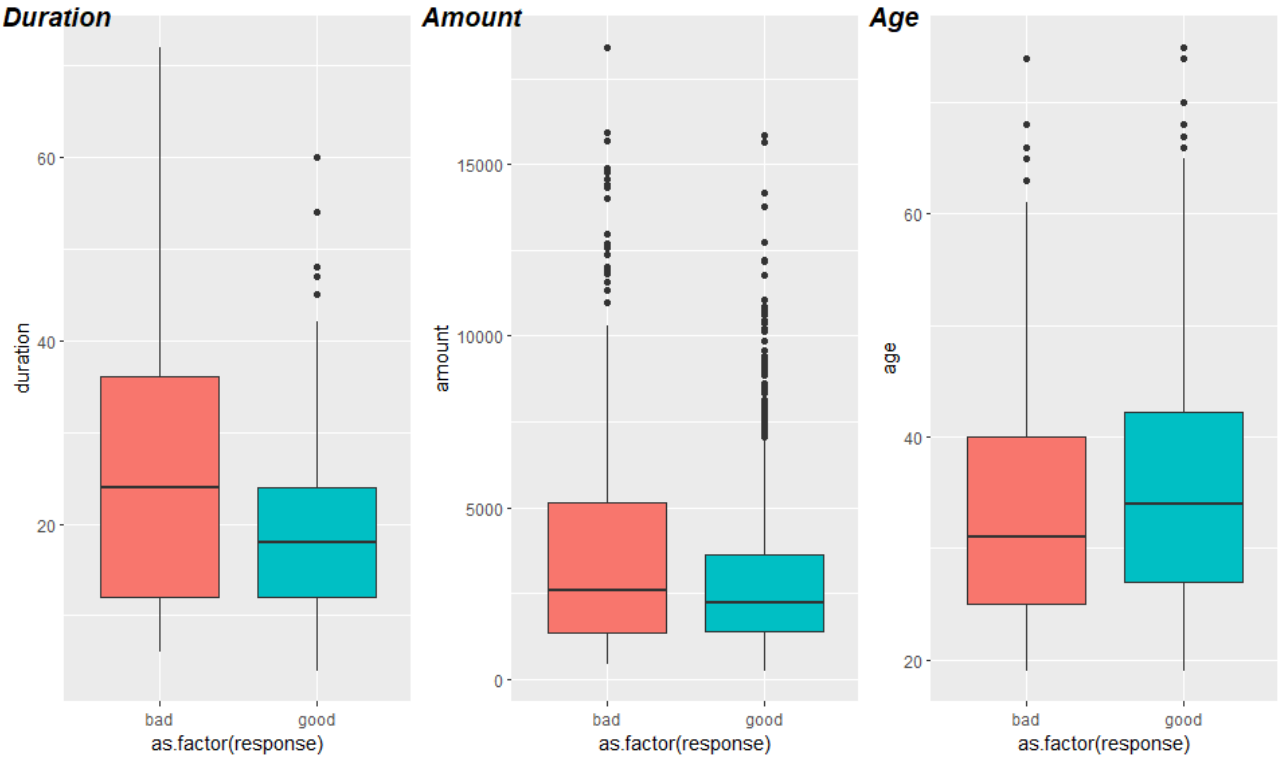
- Analiza osobista - polega na sprawdzeniu gospodarstwa domowego pod kątem liczby osób, obciążeń kredytowych, wieku czy sytuacji prawnej
- Analiza ekonomiczna - określa, czy dochody są stabilne oraz, czy dają możliwość prawidłowego regulowania przyszłego zobowiązania kredytowego
- Analiza prawna - bank kontroluje wszystkie dokumenty, sytuację prawną i inne kwestie mogące przeszkodzić w realizacji transakcji. Jeśli analityk kredytowy stwierdzi, że bank będzie narażony na ryzyko, decyzja odmowna dotycząca udzielenia kredytu będzie szybka i prosta

[2] W `https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html` znajduje się przykładowe wprowadzenie do biblioteki `DataExplorer` wykonane przez Boxuan Cui, w roku 2020, na podstawie danych `nycflights13` zawierających informacje dotyczące lotów, które odleciały w 2013 roku z Nowego Jorku.

- Analiza nieruchomości - dotyczy zweryfikowania wartości rynkowej zabezpieczenia poprzez wycenę dostarczoną przez wnioskodawcę kredytu

Poniższy rysunek (2) zawiera trzy wykresy pudełkowe zmiennych *duration* (oznaczającą, ile miesięcy zajmie spłata kredytu), *amount* (wskazującą na kwotę kredytu) oraz *age*. Każdy z wykresów podzielony jest na dwie części względem zmiennej *creditability* oznaczającej zdolność kredytową, która może być *bad* (zła) lub *good* (dobra).

Zwróćmy uwagę na wykres *duration*. W obydwu przypadkach zdolności kredytowej, czarna kreska oznaczająca medianę wykresów, jest symetryczna. Oznacza to, że wykres cech w tym przypadku jest symetryczny. Ponadto, w przypadku złej zdolności kredytowej, mamy do czynienia z dłuższym wykresem, co oznacza, że dane są bardziej rozproszone, to znaczy, mogą przyjmować bardziej różniące się wartości. Co więcej, dostrzegamy bardzo długie wąsy, co sugeruje, że mamy dużo obserwacji skrajnych. W przypadku dobrej zdolności kredytowej tego samego wykresu zauważamy, że posiada on bardzo odległe wartości odstające. Może to odzwierciedlać rzeczywisty rozkład lub być rezultatem przypadku, ale może także świadczyć o błędnym pomiarze czy pomyłkach we wprowadzaniu informacji do bazy danych.



Rysunek 2: Wykres pudełkowy zmiennych *duration*, *amount* oraz *age* danych `germancredit` - opracowanie własne

Przejdźmy do wykresu pudełkowego *amount* powyższego rysunku (2). Dokładnie widać, że mediana nie znajduje się na "środku" pudełka, zatem wykres nie jest symetryczny. Znacząco większa jest odległość górnego krańca pudełka od mediany, zatem mamy asymetrię prawostronną. Oznacza to, że obserwacje statystyczne skupiają się przy wartościach cechy mniejszych od

średniej arytmetycznej. Co więcej, w przypadku i dobrej i złej zdolności kredytowej, spotykamy się z bardzo dużą ilością obserwacji skrajnych.

Przyglądając się wykresowi *age* rysunku (2) dostrzegamy pewną niesymetryczność w obydwu przypadkach, ale w celu jej dokładnego określenia, korzystamy z funkcji `summary`. Tabela (2) zawiera podsumowanie zmiennej wiek w przypadku dobrej oraz złej zdolności kredytowej. Dostrzegamy, że wszystkie wartości (minimalne, maksymalne itd.) są ze sobą bardzo zbliżone. Ponownie w przypadku złej zdolności kredytowej spotykamy się z prawostronną asymetrią.

	age - good	age - bad
Min.	19.00	19.00
1st Qu.	32.00	30.50
Median	45.00	42.00
Mean	45.17	42.55
3rd Qu.	58.00	53.50
Max.	75.00	74.00

Tabela 2: Tabela podsumowująca zmienną wiek w przypadku dobrej oraz złej zdolności kredytowej danych `germancredit`

Kategoria wiekowa	liczba osób
18 - 30	411
30 - 50	476
50 - 80	113

Tabela 3: Kategoryzacja zmiennej wiek

Powyższa tabela (3) zawiera kategoryzację zmiennej wiek na trzy grupy - osoby w wieku między 18 a 30, między 30 a 50 oraz powyżej 50 roku życia aż do 80. Zauważamy, że najwięcej kredytów biorą osoby w drugiej kategorii wiekowej. Wydaje się to dość standardowe, ponieważ jest to wiek, kiedy ludzie najczęściej chcą się ustatkować i biorą kredyty na mieszkania. Według Macieja Kazimierskiego [3], analityka finansowego, znaczna większość kredytobiorców w momencie zaciągnięcia zobowiązania była w wieku od 25 do 44 lat. Sprawdzimy, czy w przypadku zmiany kategoryzacji zmiennej wiek, również uzyskamy podobne wnioski.

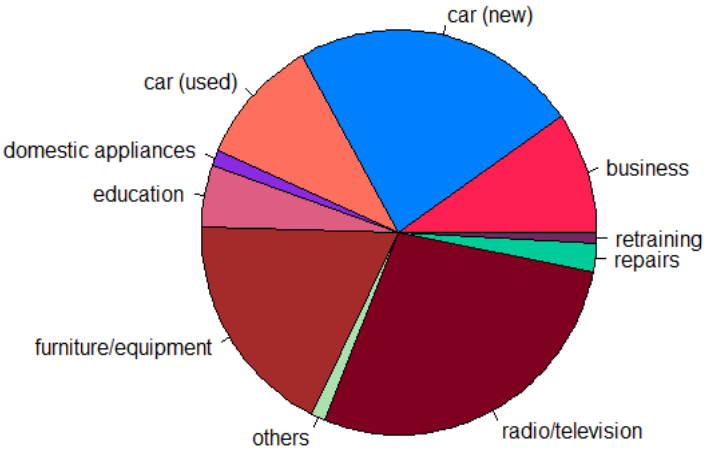
Kategoria wiekowa	liczba osób
18 - 25	190
25 - 44	609
44 - 80	201

Tabela 4: Kategoryzacja zmiennej wiek

Widzimy faktycznie, że ze wszystkich 1000 danych prawie 61% jest osobami w wieku od 25 do 44 lat. Z oczywistych względów osoba, która dopiero co ukończyła 18 rok życia nie uzyska

[3] Źródło `https://direct.money.pl/artykuly/porady/kto-w-polsce-bierze-kredyt-na-mieszkanie,247,0,2353655`

zbyt wielu punktów i jej ocena scoringowa może być niewystarczająca do uzyskania kredytu. W praktyce więc banki w ten sposób uniemożliwiają osobom poniżej 21 roku życia uzyskanie kredytu mieszkaniowego. Pokrywa się to z naszą tabelą (4), gdzie najmniej kredytów otrzymaliśmy w zadanej grupie wiekowej. Znacznie łatwiej mają osoby, które są w wieku 25 - 44 lat. Ta grupa kredytobiorców charakteryzuje się już względną stabilnością zawodową i rodzinną, co stanowi cechy bardzo pożądane przez banki. Te osoby mogą łatwiej uzyskać kredyt niż osoby dopiero rozpoczynające swoje dorosłe życie. Dlatego w tej kategorii wiekowej posiadamy najwięcej osób.



Rysunek 3: Wykres kołowy zmiennej *purpose* danych `germancredit` - opracowanie własne

purpose	number
business	97
car (new)	234
car (used)	103
domestic appliances	12
education	50
furniture/equipment	181
others	12
radio/television	280
repairs	22
retraining	9

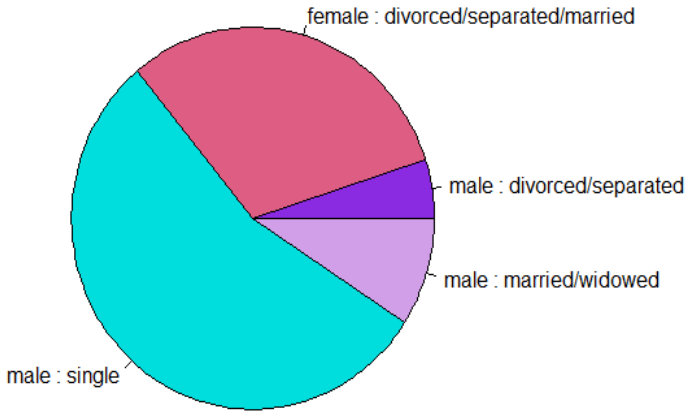
Tabela 5: Tabela zawierająca podsumowanie zmiennej *purpose* danych `germancredit`

Powyższy rysunek (3) oraz tabela (5) określają nam, na co najczęściej ludzie chcieli wziąć kredyt

w banku. Dostrzegamy, że najchętniej były to trzy rzeczy - radio/telewizor, nowy samochód bądź meble/wyposażenie, zapewne do nowego mieszkania.

purpose	number	response good	response bad
car (new)	234	145	89
education	50	28	22
furniture/equipment	181	123	58
radio/television	280	218	62

Tabela 6: Tabela zawierająca podsumowanie zmiennej *purpose* w kategorii dobrej bądź złej zdolności kredytowej danych `germancredit`



Rysunek 4: Wykres kołowy zmiennej *Personal status and sex* danych `germancredit` - opracowanie własne

Personal status and sex	number
male : divorced/separated	50
female : divorced/separated/married	310
male : single	548
male : married/widowed	92

Tabela 7: Tabela zawierająca zmienną *Personal status and sex* danych `germancredit`

Na podstawie powyższego rysunku (4) oraz tabeli (7) widzimy, że zdecydowaną większością kredytobiorców są samotni mężczyźni (54, 8%). Równocześnie, najmniejszą liczbą kredytobiorców są również mężczyźni, ale rozwiedzeni bądź w separacji (5%).

3 ZADANIE 2

Zbiór treningowy jest to zbiór, na którym dany model się uczy, na podstawie którego adaptuje swoje parametry. Model ma pełny dostęp do danych treningowych, także do informacji o klasach wszystkich wektorów.
Zbiór testowy jest to zbiór służący do sprawdzania poprawności zaprojektowania i stopnia nauczenia modelu. Dostęp do klas wektorów jest zabroniony, ponieważ celem działania modelu jest właśnie przewidzenie klasy. Po procesie uczenia można porównać klasę przewidywaną z prawdziwą, czyli stwierdzić czy wystąpił błąd w klasyfikacji (badając tym samym dokładność klasyfikatora).

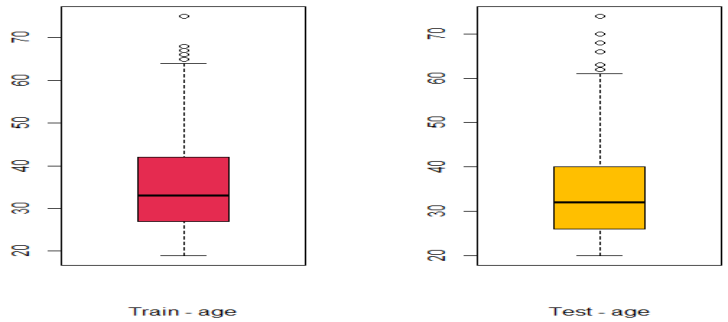
Pakiet `scorecard` [4] umożliwia opracowanie karty oceny ryzyka kredytowego łatwiej i wydajniej, udostępniając funkcje dla niektórych typowych zadań, takie jak podział danych, wybór zmiennych, skalowanie karty wyników czy ocena wydajności. W celu podziału naszego zbioru danych `germancredit` na zbiór treningowy oraz testowy skorzystamy z funkcji `split_df`, dostępnego w omówionej wyżej bibliotece, który podzieli nasz zbiór w proporcjach 70 : 30

	Good creditability	Bad creditability	Number
Train	479	202	681
Test	221	98	319

Tabela 8: Tabela zawierająca podział ze względu na zmienną *creditability* danych testowych i treningowych `germancredit`

Przyjrzymy się najpierw zbiorowi treningowemu *train*. Ze wszystkich obserwacji, osoby określone jako dobrzy klienci stanowią 71%, analogicznie osoby określone jako źli klienci wynoszą 29%. W przypadku zbioru testowego *test* otrzymujemy następujące proporcje. Klienci określeni jako dobrzy liczą 70%, zaś osoby określone jako źli klienci wyznaczają 30%.

Stosunek dobrych oraz złych klientów zbioru uczącego, oraz testowego jest niemal identyczny. Wnioskujemy zatem, że podział danych `germancredit` jest poprawny.



Rysunek 5: Wykres pudełkowy zmiennej *age* danych `germancredit` - opracowanie własne

[4]Pełny opis biblioteki oraz lista dostępnych funkcji `https://cran.r-project.org/web/packages/scorecard/scorecard.pdf`

4 ZADANIE 3

W scoringu kredytowym stosowane są między innymi następujące metody

- statystyczne, na przykład regresja dla danych binarnych
- niestatystyczne, na przykład sieci neuronowe bądź algorytmy genetyczne

Aktualnie zajmiemy się omówieniem modelu regresji dla danych binarnych. Ogólnie, można go przedstawić w następującej postaci

$$g(p(\mathbf{x})) = \mathbf{x}'\beta$$

Gdzie $p(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$ jest prawdopodobieństwem sukcesu w przypadku, gdy wektorem atrybutów klienta jest $\mathbf{x}$, g jest pewną odpowiednią funkcją nazywaną łączącą, zaś parametr β definiujemy jako wektor nieznanych parametrów modelu.

Do najpopularniejszych modeli regresji dla danych binarnych należą

- model regresji logistycznej, który otrzymujemy przyjmując za η dystrybuantę rozkładu logistycznego
$$g(p(\mathbf{x})) = \ln \frac{p(\mathbf{x})}{1 - p(\mathbf{x})}$$
- model regresji probitowej, który otrzymujemy przyjmując za η dystrybuantę rozkładu normalnego
$$g(p(\mathbf{x})) = \Phi^{-1}(p(\mathbf{x}))$$
- model $log - log$, który otrzymujemy przyjmując za η dystrybuantę rozkładu Gumbela,
$$g(p(\mathbf{x})) = -\ln(-\ln(p(\mathbf{x})))$$
- model *complementary* $log - log$, który otrzymujemy przyjmując za η dystrybuantę rozkładu $\ln(Z)$ gdzie Z ma rozkład wykładniczy $\mathcal{E}$ (1)
$$g(p(\mathbf{x})) = \ln(-\ln(1 - p(\mathbf{x})))$$

Zadanie wyboru modelu w przypadku parametrycznych modeli regresji polega właściwie na wyborze zmiennych objaśniających. Wyboru zmiennych do modelu możemy dokonać w oparciu o testy statystyczne lub przy pewnych kryteriach.

Kryterium informacyjne Akaik'a (AIC) określone jest wzorem

$$AIC = -2 \ln L\left(\hat{\theta}|data\right) + 2k$$

Gdzie L jest funkcją wiarygodności związaną z modelem parametrycznym, $\hat{\theta}$ jest estymatorem największej wiarygodności parametru θ modelu, zaś *data* jest wektorem danych. Parametr k jest liczbą estymowanych parametrów modelu.

Warto dodać, że kryterium AIC zazwyczaj przyjmuje wartości dodatnie.

Bayesowskie kryterium informacyjne (BIC) określone jest wzorem

$$BIC = -2 \ln L\left(\hat{\theta}|data\right) + k \ln n$$

Gdzie L jest funkcją wiarygodności związaną z modelem parametrycznym, $\hat{\theta}$ jest estymatorem największej wiarygodności parametru θ modelu, zaś *data* jest wektorem danych. Parametr k jest liczbą estymowanych parametrów modelu, natomiast parametr n określa liczbę danych.

Nadmieńmy, że wartość $k \ln n$ w kryterium BIC jest większą wartością niż $2k$ w kryterium AIC. Stąd model wybrany na podstawie kryterium BIC może być modelem prostszym (czyli będzie miał mniej parametrów) od modelu wybranego na podstawie kryterium AIC.

Regresja logistyczna jest przydatna w sytuacjach, w których wymagane jest przewidywanie obecności lub braku cechy bądź wyniku na podstawie wartości zestawu predyktorów. Współczynniki regresji logistycznej mogą być używane do oszacowania ilorazów szans dla każdej zmiennej niezależnej w modelu. Głównym czynnikiem wpływającym na fakt, że do omówienia naszych danych wybraliśmy regresję logistyczną, jest łatwość w interpretacji wyników.

W ramach najlepszego modelu staramy się znaleźć ten najbardziej odpowiedni za pomocą regresji logistycznej z wykorzystaniem kryteriów AIC, BIC. Korzystając z funkcji `glm`[5] z biblioteki `stats` tworzymy model regresji logistycznej.

```
model_glm<-glm(formula = response~., data = train, family = binomial('logit'))
```

Rysunek 6: Kod tworzący model regresji logistycznej dla pełnego modelu - opracowanie własne

Wartość AIC z zadanego modelu (6) wyniosła 677.7961. Jak już jednak wspomnieliśmy, odpowiedni model postaramy się dobrać z wykorzystaniem kryteriów AIC oraz BIC.
Dla kryterium informacyjnego Akaik'a tworzymy trzy funkcje - AIC model forward, AIC model backward oraz AIC model both. Analogicznie postępujemy w przypadku kryterium informacyjnego Bayesowskiego.

Najpierw jednak zdefiniujemy sobie określone wyżej procedury.

- procedura forward - startujemy z pustego modelu i dodajemy zmienne pojedynczo w taki sposób, aby w każdym kroku maksymalnie poprawiać wartość kryterium
- procedura backward - startujemy z modelu ze wszystkimi zmiennymi i usuwamy zmienne w taki sposób, aby w każdym kroku maksymalnie poprawiać wartość kryterium
- procedura both - na zmianę dodajemy i usuwamy zmienne zgodnie z powyższą regułą

[5]Funkcja `glm` służy do dopasowywania uogólnionych modeli liniowych określonych przez podanie symbolicznego opisu predyktora liniowego, oraz opisu rozkładu błędów. Więcej informacji można znaleźć w `https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm`

```
AIC_model_forward <- step(
  object = model_glm_simple,
  scope=list(lower=model_glm_simple,upper=model_glm),
  data = train,
  direction = "forward",
  family = binomial('logit'),
  trace = FALSE)

AIC_model_backward <- step(object = model_glm, direction = "backward", trace = FALSE)

AIC_model_both <- step(object = model_glm, direction = "both", trace = FALSE)
```

Rysunek 7: Kryterium informacyjne AIC - opracowanie własne

W wszystkich ustalonych powyżej trzech przypadkach wartość kryterium AIC wyniosła 660.427.

```
BIC_model_forward <- step(
  object = model_glm_simple,
  scope=list(lower=model_glm_simple,upper=model_glm),
  data = train,
  direction = "forward",
  family = binomial('logit'),
  trace = FALSE,
  k=log(nrow(train)))

BIC_model_backward <- step(object = model_glm, direction = "backward", trace = FALSE,
  k = log(nrow(train)))

BIC_model_both <- step(object = model_glm, direction = "both", trace = FALSE,
  k = log(nrow(train)))
```

Rysunek 8: Kryterium informacyjne BIC - opracowanie własne

Podobnie jak wcześniej otrzymaliśmy wszędzie jednakową wartość, AIC dla wszystkich przypadków wyniosła 716.5272.
Dokonamy teraz wyboru zmiennych do modelu. Mniejsza liczba zmiennych w modelu pozwala na łatwiejszą wizualizację i interpretacje wyników. Mniejsza ilość zmiennych skutkuje także mniejszą liczbą obserwacji, na których uczy się model. Ponadto wybierając odpowiedni model jesteśmy w stanie wywnioskować nieistotność pewnych zmiennych i usunąć je z modelu.

```
## Start:  AIC=677.8
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + sex + other_debtor +
##     present_resid + property + age + other_install + housing +
##     n_credits + job + n_people + telephone + foreign
##
##                    Df Deviance    AIC
## - sex               3   581.52 673.52
## - property          3   581.59 673.59
## - job               3   582.64 674.64
## - present_resid     1   579.80 675.80
## - n_people          1   579.89 675.89
## - telephone         1   580.00 676.00
## <none>                  579.80 677.80
## - installment_rate  1   583.14 679.14
## - saving_acct       4   589.51 679.51
## - amount            1   583.83 679.83
## - n_credits         1   584.19 680.19
## - housing           2   586.61 680.61
## - other_debtor      2   587.26 681.26
## - duration          1   585.71 681.71
## - present_emp       4   591.96 681.96
## - other_install     2   588.80 682.80
## - credit_his        4   593.20 683.20
## - age               1   589.72 685.72
## - foreign           1   590.36 686.36
## - purpose           9   618.33 698.33
## - chk_acct          3   629.77 721.77
##
## Step:  AIC=673.52
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     present_resid + property + age + other_install + housing +
##     n_credits + job + n_people + telephone + foreign
##
##                    Df Deviance    AIC
## - property          3   583.06 669.06
## - job               3   584.40 670.40
## - present_resid     1   581.53 671.53
## - n_people          1   581.60 671.60
## - telephone         1   581.79 671.79
## <none>                  581.52 673.52
```

Rysunek 9: Wybór zmiennych do modelu AIC backward - opracowanie własne

```
## - installment_rate  1   584.58 674.58
## - saving_acct       4   590.96 674.96
## - amount            1   585.51 675.51
## - n_credits         1   585.71 675.71
## - housing           2   588.26 676.26
## - present_emp       4   593.09 677.09
## - other_debtor      2   589.22 677.22
## - duration          1   587.40 677.40
## - credit_his        4   594.76 678.76
## - other_install     2   590.85 678.85
## - age               1   591.54 681.54
## - foreign           1   591.84 681.84
## - purpose           9   620.43 694.43
## - chk_acct          3   630.95 716.95
##
## Step:  AIC=669.06
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     present_resid + age + other_install + housing + n_credits +
##     job + n_people + telephone + foreign
##
##                    Df Deviance    AIC
## - job               3   586.09 666.09
## - present_resid     1   583.08 667.08
## - n_people          1   583.13 667.13
## - telephone         1   583.30 667.30
## <none>                  583.06 669.06
## - saving_acct       4   592.11 670.11
## - installment_rate  1   586.34 670.34
## - n_credits         1   587.06 671.06
## - amount            1   587.55 671.55
## - present_emp       4   595.11 673.11
## - other_debtor      2   591.25 673.25
## - duration          1   589.51 673.51
## - housing           2   592.30 674.30
## - credit_his        4   596.37 674.37
## - other_install     2   592.68 674.68
## - age               1   592.92 676.92
## - foreign           1   593.48 677.48
## - purpose           9   622.54 690.54
## - chk_acct          3   633.48 713.48
##
## Step:  AIC=666.09
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     present_resid + age + other_install + housing + n_credits +
```

Rysunek 10: Wybór zmiennych do modelu AIC backward - opracowanie własne

```
##     n_people + telephone + foreign
##
##                   Df Deviance    AIC
## - n_people          1   586.12 664.12
## - present_resid     1   586.14 664.14
## - telephone         1   586.34 664.34
## <none>                  586.09 666.09
## - saving_acct       4   595.54 667.54
## - installment_rate  1   589.71 667.71
## - n_credits         1   589.97 667.97
## - amount            1   590.92 668.92
## - present_emp       4   597.08 669.08
## - other_debtor      2   594.18 670.18
## - housing           2   594.90 670.90
## - duration          1   592.99 670.99
## - other_install     2   595.58 671.58
## - credit_his        4   599.72 671.72
## - age               1   596.61 674.61
## - foreign           1   597.21 675.21
## - purpose           9   625.20 687.20
## - chk_acct          3   635.09 709.09
##
## Step:  AIC=664.12
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     present_resid + age + other_install + housing + n_credits +
##     telephone + foreign
##
##                   Df Deviance    AIC
## - present_resid     1   586.18 662.18
## - telephone         1   586.39 662.39
## <none>                  586.12 664.12
## - saving_acct       4   595.55 665.55
## - installment_rate  1   589.73 665.73
## - n_credits         1   590.06 666.06
## - amount            1   590.93 666.93
## - present_emp       4   597.09 667.09
## - other_debtor      2   594.18 668.18
## - housing           2   594.93 668.93
## - duration          1   593.02 669.02
## - other_install     2   595.60 669.60
## - credit_his        4   599.87 669.87
## - age               1   596.62 672.62
## - foreign           1   597.23 673.23
## - purpose           9   625.41 685.41
## - chk_acct          3   635.43 707.43
```

Rysunek 11: Wybór zmiennych do modelu AIC backward - opracowanie własne

```
##
## Step:  AIC=662.18
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     age + other_install + housing + n_credits + telephone + foreign
##
##                   Df Deviance    AIC
## - telephone         1   586.43 660.43
## <none>                  586.18 662.18
## - saving_acct       4   595.56 663.56
## - installment_rate  1   589.86 663.86
## - n_credits         1   590.15 664.15
## - amount            1   591.00 665.00
## - present_emp       4   597.16 665.16
## - other_debtor      2   594.18 666.18
## - duration          1   593.03 667.03
## - other_install     2   595.61 667.61
## - credit_his        4   599.87 667.87
## - housing           2   596.29 668.29
## - age               1   596.64 670.64
## - foreign           1   597.27 671.27
## - purpose           9   625.42 683.42
## - chk_acct          3   635.70 705.70
##
## Step:  AIC=660.43
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     age + other_install + housing + n_credits + foreign
##
##                   Df Deviance    AIC
## <none>                  586.43 660.43
## - saving_acct       4   595.81 661.81
## - installment_rate  1   590.04 662.04
## - n_credits         1   590.31 662.31
## - amount            1   591.00 663.00
## - present_emp       4   597.63 663.63
## - other_debtor      2   594.36 664.36
## - duration          1   593.58 665.58
## - other_install     2   595.84 665.84
## - credit_his        4   600.05 666.05
## - housing           2   596.76 666.76
## - age               1   597.34 669.34
## - foreign           1   597.38 669.38
## - purpose           9   626.03 682.03
## - chk_acct          3   636.90 704.90
```

Rysunek 12: Wybór zmiennych do modelu AIC backward - opracowanie własne

Najmniejszą wartość AIC otrzymaliśmy w przypadku modelu $response \rightarrow chk\_acct + duration + credit\_his + purpose + amount + saving\_acct + present\_emp + installment\_rate + other\_debtor + age + other\_install + housing + n\_credits + foreign$.

Analogicznie postępujemy w przypadku kryterium AIC forward, AIC both, BIC backward, BIC forward, BIC both. Wszystkie wyniki otrzymanych wartości AIC zapisujemy w tabeli. Wybierzemy najmniejszą wartość i na tej podstawie określimy najlepsze kryterium oraz najlepszy model.

	AIC.backward	AIC.forward	AIC.both	BIC.backward	BIC.forward	BIC.both
AIC	660.43	677.8	660.43	611.1	613.76	611.1

Tabela 9: Wartość AIC

Widzimy, że model AIC backward oraz AIC both oddają te same wartości AIC, tak samo, jak BIC backward i BIC both.
Poniżej, na rysunkach, prezentujemy które zmienne wchodzą w skład najlepszych modeli względem kryterium AIC oraz kryterium BIC.

```
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     age + other_install + housing + n_credits + foreign
```

Rysunek 13: Wybór zmiennych do modelu AIC - opracowanie własne

```
## response ~ chk_acct + duration + credit_his + purpose + amount +
##     saving_acct + present_emp + installment_rate + other_debtor +
##     age + other_install + housing + n_credits + job + foreign
```

Rysunek 14: Wybór zmiennych do modelu BIC - opracowanie własne

5 ZADANIE 4

Do oceny efektywności modelu scoringowego używa się między innymi macierze pomyłek. Macierze pomyłek używane są do określenia wydajności modelu klasyfikacji, gdzie N oznacza liczbę klas docelowych. Macierz porównuje rzeczywiste wartości docelowe z przewidywanymi przed model. Daje nam to całościowy obraz tego, jak dobrze działa nasz model klasyfikacji i jakie rodzaje błędów popełnia.

Tablica 1: Macierz pomyłek

		Zaklasyfikowani		
		Bad	Good	
Real	Bad	TN	FP	N
	Good	FN	TP	P
		N*	P*	

Rysunek 15: Macierz pomyłek - źródło

Kryteria oceny modelu scoringowego w oparciu o macierz pomyłek są następujące:

- Dokładność (*Accuracy*) pozwala nam ocenić jakość klasyfikacji testu. Daje nam informacje na temat tego, jaka część testów, ze wszystkich zaklasyfikowanych, została oceniona poprawnie. Im wyższa wartość dokładności, tym lepiej. `ACC` $= 1$ oznacza idealnie dopasowanie i brak pomyłki ani razu.

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

- Prawdziwie pozytywna wartość (*True Positive Rate*) zwana inaczej czułością (*sensivity*). Mówi nam o tym, jaki jest udział prawidłowo prognozowanych przypadków pozytywnych wśród wszystkich przypadków pozytywnych. Wartość ta powinna być jak najbliższa 1.

$$TPR = \frac{TP}{P} \tag{1}$$

$$P(d(\mathbf{X}) = 1|E = 1) = P(s(\mathbf{X}) > c|E = 1)$$

- Prawdziwie negatywna wartość (*True Negative Rate*) inaczej specyficzność (*specifity*) mierzy, jak dużo ze wszystkich negatywnych przypadków zostało rzeczywiście zaklasyfikowanych do tej kategorii.

$$TNR = \frac{TN}{TN + FP}$$

$$P(d(\mathbf{X}) = 0|E = 0) = P(s(\mathbf{X}) \leq c|E = 0)$$

- Precyzja przewidywania pozytywnego (*positive prediction value*) jest to miara precyzji wskazująca, z jaką pewnością możemy ufać przewidywaniom pozytywnym

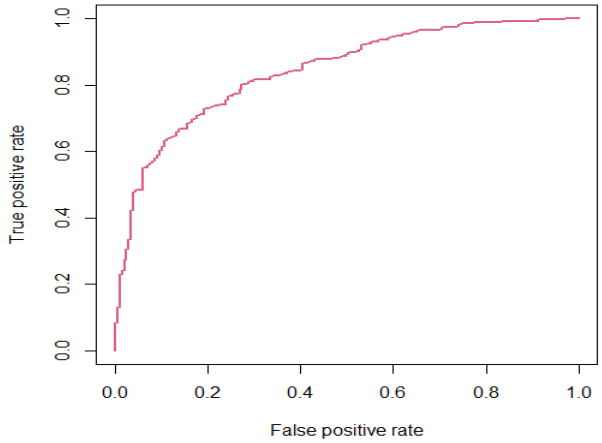
$$PPV = \frac{TP}{P*}$$

- Precyzja przewidywania negatywnego (*negative prediction value*) jest to miara precyzji wskazująca, z jaką pewnością możemy ufać przewidywaniom negatywnym

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

- Błąd

$$1 - ACC = 1 - \frac{TP + TN}{N + P}$$



Rysunek 16: Wykres krzywej ROC zbioru treningowego - opracowanie własne

Krzywa ROC wygląda dobrze i obserwujemy również 0.8413075 jako wartość AUC. Zgodnie z praktyczną regułą wartość AUC $> 0,70$ można uznać za zadowalającą. Krzywa ROC oznacza ogólną miarę dobroci klasyfikacji, stąd wyższa wartość oznacza, że model ma dobrą zdolność klasyfikacyjną.

```
model_glm <- glm(formula = response~., data = train, family = binomial('logit'))
AIC_model_both <- step(object = model_glm, direction = "both", trace = FALSE)
M1 <- AIC_model_both
AIC_pred <- predict.glm(object=M1, newdata=test, type='response')
AIC_conf_mat <- table(real = test$response, pred = as.numeric(AIC_pred>0.5))
```

Rysunek 17: Kod źródłowy do stworzenia macierzy pomyłek dla AIC_both - opracowanie własne

	bad	good
bad	51	47
good	40	181

Tabela 10: Macierz pomyłek dla AIC_both w pełnym modelu

accuracy	error	sensitivity	specificity	positive prediction	negative prediction
0.7272727	0.2727273	0.8190045	0.5204082	0.7938596	0.5604396

Tabela 11: Tabela z wyznaczonymi wartościami metryk

```
model_glm <- glm(formula = response~., data = train, family = binomial('logit'))
BIC_model_both <- step(object = model_glm, direction = "both", trace = FALSE,
  k=log(nrow(train)))
M2 <- BIC_model_both
BIC_pred <- predict.glm(object=M2, newdata=test, type='response')
BIC_model_both <- step(object = model_glm, direction = "both", trace = FALSE,
  k=log(nrow(train)))
```

Rysunek 18: Kod źródłowy do stworzenia macierzy pomyłek dla BIC_both - opracowanie własne

	bad	good
bad	39	59
good	34	187

Tabela 12: Macierz pomyłek dla BIC_both w pełnym modelu

accuracy	error	sensitivity	specificity	positive prediction	negative prediction
0.7084639	0.2915361	0.8461538	0.3979592	0.7601626	0.5342466

Tabela 13: Tabela z wyznaczonymi wartościami metryk

Przeanalizujemy obydwie tabelki (11) oraz (27). Im wyższa dokładność tym lepiej. Dokładność 1.00 oznacza, że wszystko jest idealnie dopasowane i algorytm nie pomylił się ani razu. W naszym przypadku, w obydwu modelach wartość accuracy wynosi około 0.7. Wnioskujemy więc, że nasz algorytm jest nawet poprawny. Im mniejsza wartość error, tym dokładniejszy jest nasz algorytm. W naszym przypadku wynosi on około 0.3, co wydaje się dość dużym błędem. Czułość określamy jako zdolność testu do prawidłowego rozpoznania na przykład choroby tam, gdzie ona występuje. Nasza sensitivity ponownie w obydwu przypadkach jest bardzo podobne i wynosi około 0.82, co uznajemy za dobrą wartość. Dalej analizujemy swoistość. Test o wysokiej swoistości cechuje niski błąd pierwszego rodzaju [6]. W przypadku modelu AIC_both swoistość wynosi 0.52, zaś w przypadku modelu BIC_both wynosi ona zaledwie 0.39. Idealna wartość PPV przy idealnym teście wynosi 1, zaś najgorsza możliwa wartość to 0. Ponownie, w przypadku obydwu modelu uzyskaliśmy zbliżoną wartość tego wskaźnika, wynoszącą około 0.8. W przypadku idealnego testu, który nie zwraca wyników fałszywie ujemnych, wartość NPV wynosi 1, a w przypadku testu, który nie zwraca wyników prawdziwie ujemnych, wartość NPV wynosi 0. U nas negative prediction wynosi około 0.56.

Uzyskane wyniki uważam, że nie są najlepsze i zapewne wybierając inny model, można uzyskać lepsze.

[6]Błędem pierwszego rodzaju nazywamy błąd polegający na odrzuceniu hipotezy zerowej, która w rzeczywistości nie jest fałszywa

6 ZADANIE 5

Na podstawie zbioru treningowego wyznaczonego w zadaniu 3 (tabelka (8)) dokonamy wyboru zmiennych do karty scoringowej w oparciu o IV ($InformationValue$). Za zmienną zależna przyjmiemy zmienną creditability, zaś pozostałe zmienne określamy jako potencjalne predyktory.

Wyjaśnimy najpierw kilka pojęć. Scoring kredytowy definiujemy jako proces, w wyniku którego informacje dotyczące ocenianej jednostki są w sposób obiektywny przekształcane na ciąg liczb, które po zsumowaniu stanowią miarę oceny ryzyka kredytowego. Przy pomocy scoringu ocenia się na przykład reakcję klienta na nowy produkt bądź prawdopodobieństwo, że klient będzie korzystał z produktu po zakończeniu okresu promocyjnego.

Punktową ocenę ryzyka można sklasyfikować według następujących kryteriów

- ze względu na oceniany podmiot
- ze względu na rodzaj kredytu
- ze względu na podmiot wykorzystujący scoring
- ze względu na dziedzinę wykorzystania scoringu

Waga dowodów (WOE - $Weight\ of\ Evidence$) jest miarą tego, jak bardzo dowody wspierają lub podważają hipotezę. WoE mierzy względne ryzyko atrybutu poziomu kategoryzacji. Wartość zależy od tego, czy wartość zmiennej docelowej jest zerem, czy zdarzeniem.

Analiza wartości informacji ($Information\ Value$) to technika eksploracji danych, która pomaga określić, które kolumny w zbiorze danych mają moc predykcyjną lub wpływ na wartość określonej zmiennej zależnej.

W praktyce oszacowanie IV wykorzystuje się do wyboru predyktorów do modelu scoringowego. Przyjmuje się, że jeżeli

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

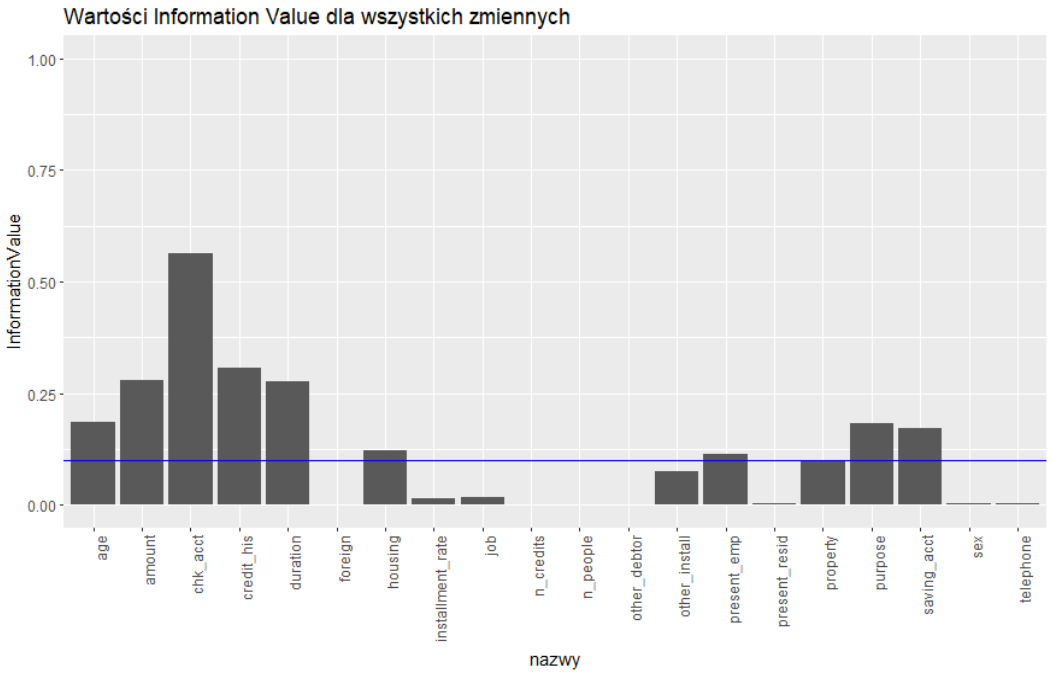
W głównej mierze korzystać będziemy z trzech pakietów

- `scorecard` [7]
- `woeBinning` [8]
- `InformationValue` [9]

Rysunek (19) zawiera skategoryzowane predyktory typu ciągłego, dla których $IV > 0.1$. Rysunek (20) jest wykresem, który pokazuje nam wszystkie zmienne z modelu oraz ich wartość IV. W dalszej części na podstawie zmiennych, których wartość IV przekracza niebieską linię na rysunku (20) stworzymy nowy zbiór treningowy oraz nowy zbiór testowy.

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

Rysunek 19: Wybrane zmienne, dla których IV > 0.1 - opracowanie własne



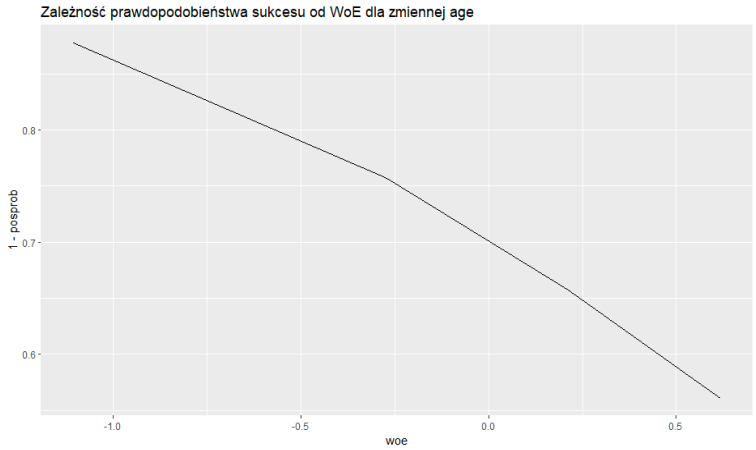
Rysunek 20: Wykres mocy predykcyjnej dla wszystkich zmiennych z modelu - opracowanie własne

[7] umożliwia opracowanie karty oceny ryzyka kredytowego łatwiej i wydajniej, udostępniając funkcje dla niektórych typowych zadań takich jak podział danych czy wybór zmiennych. `https://cran.r-project.org/web/packages/scorecard/scorecard.pdf`

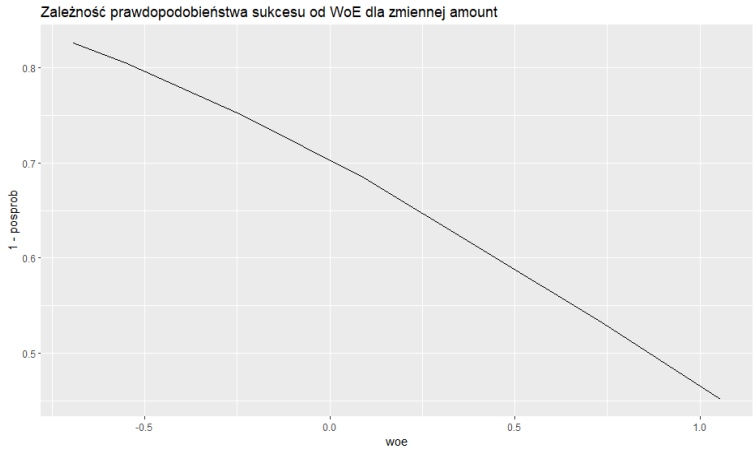
[8] Pakiet ten generuje, wizualizuje, zestawia w tabele i wdraża kategoryzację zmiennych pod nadzorem ciężaru dowodu (WOE) `https://www.rdocumentation.org/packages/woeBinning/versions/0.1.6/topics/woeBinning`

[9] Diagnostyka przewidywanych wyników prawdopodobieństwa, analiza wydajności, funkcje wspomagające poprawę dokładności. `https://cran.r-project.org/src/contrib/Archive/`

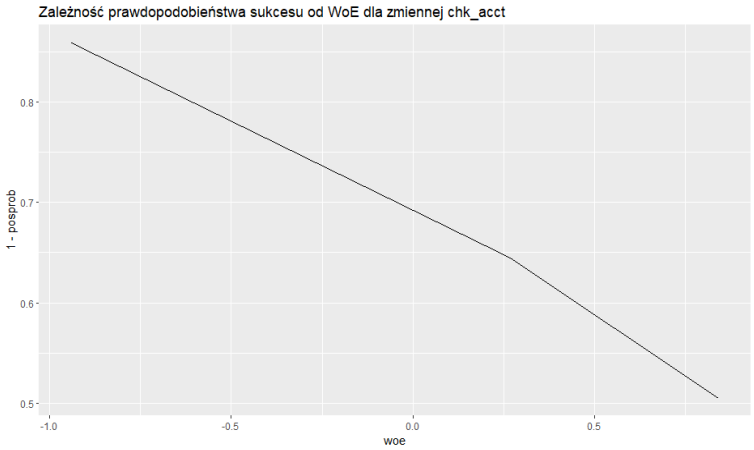
Stworzyliśmy nowy zbiór treningowy i testowy ze skategoryzowanymi zmiennymi, dla których $IV > 0.1$. Na ich podstawie będziemy sprawdzać założenia dotyczące monotoniczności względem prawdopodobieństwa sukcesu estymowanego.



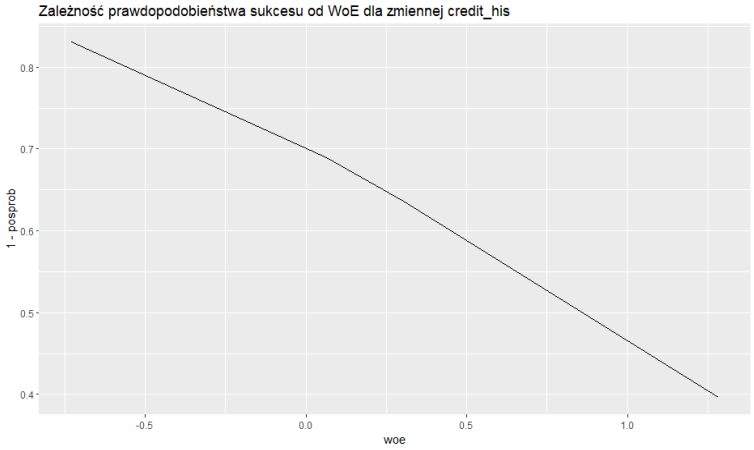
Rysunek 21: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *age* - opracowanie własne



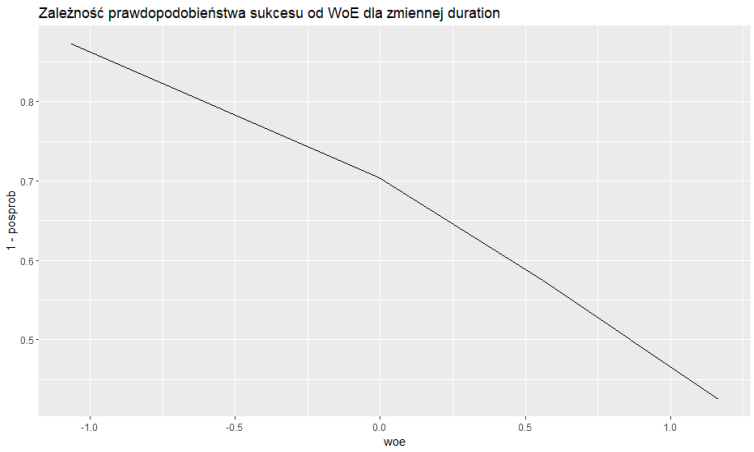
Rysunek 22: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *amonut* - opracowanie własne



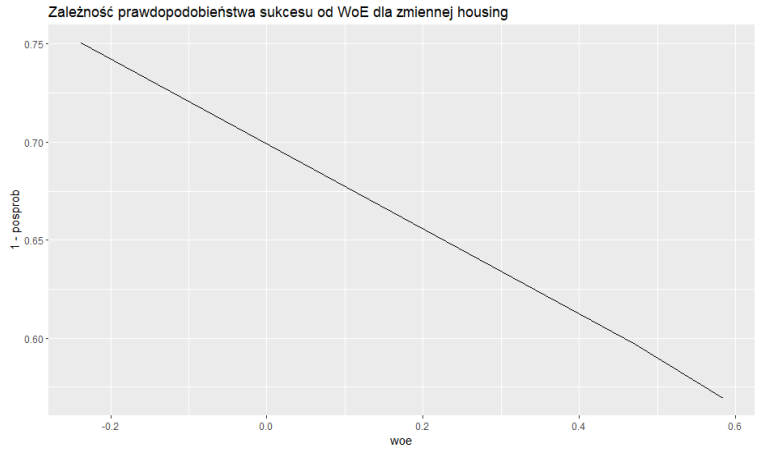
Rysunek 23: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *chk_acct* - opracowanie własne



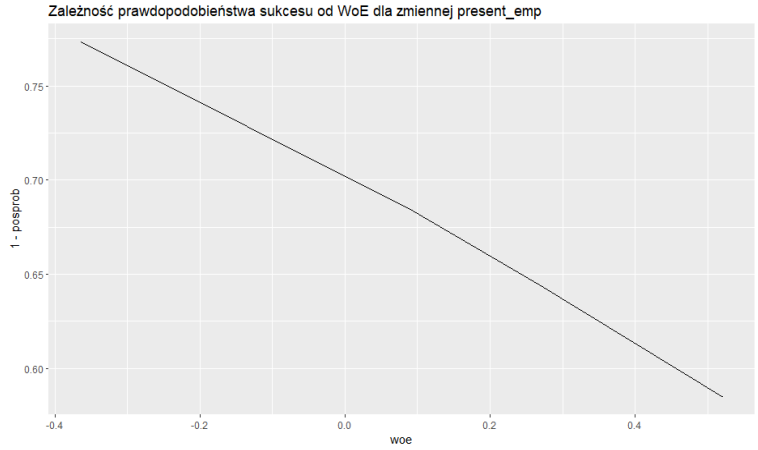
Rysunek 24: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *credit_hist* - opracowanie własne



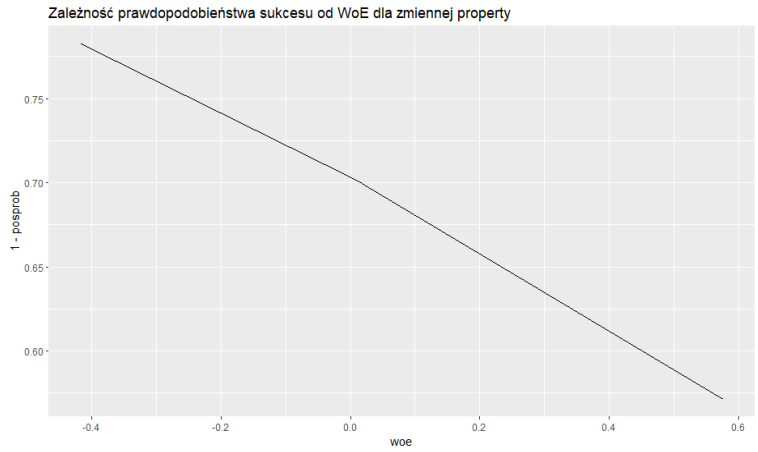
Rysunek 25: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *duration* - opracowanie własne



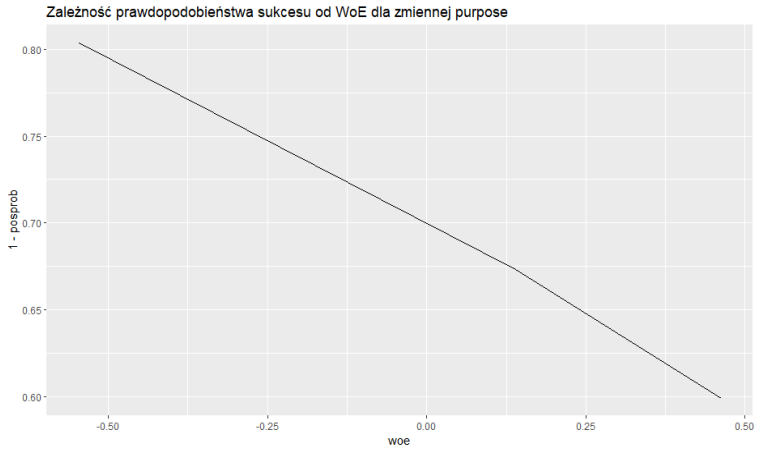
Rysunek 26: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *housing* - opracowanie własne



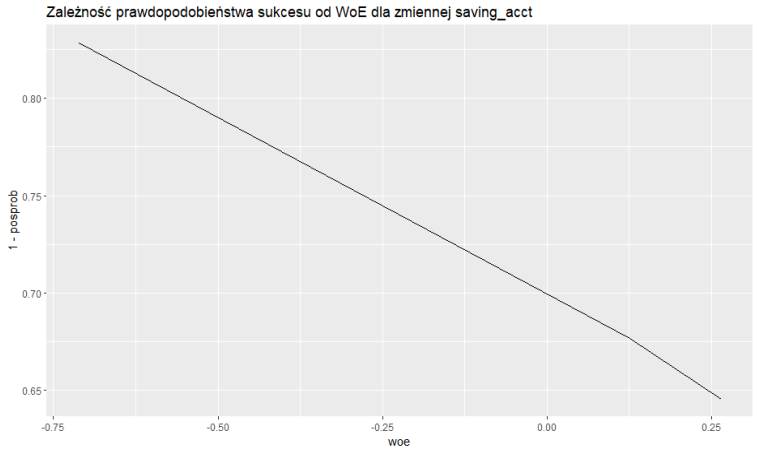
Rysunek 27: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *present_emp* - opracowanie własne



Rysunek 28: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *property* - opracowanie własne



Rysunek 29: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *purpose* - opracowanie własne



Rysunek 30: Zależność prawdopodobieństwa sukcesu od WoE dla zmiennej *saving_acct* - opracowanie własne

Przyglądając się powyższym rysunkom (21) - (31) możemy dostrzec zależność, że wraz ze wzrostem WoE spada prawdopodobieństwo sukcesu. Przypuszczamy, że estymowane wartości współczynników w modelach będą miały wartości ujemne.

	V1
(Intercept)	0.4518600
chk_acct_bin... >= 200 DM / salary assignments for at least 1...	1.6352550
chk_acct_bin0 <= ... < 200 DM	0.7343383
duration_bin[12,32)	-1.2806993
duration_bin[32,44)	-1.6134581
duration_bin[44, Inf)	-2.2972376
credit_his_bindelay in paying off in the past	-0.8083735
credit_his_binexisting credits paid back duly till now	-0.6190783
credit_his_binno credits taken/ all credits paid back duly%,%...	-1.2550576
purpose_binothers%,%education%,%car (new)%,%repairs	-0.8393374
purpose_binretraining%,%domestic appliances%,%car (used...	0.5409325
amount_bin[1800,2800)	-0.1328483
amount_bin[2800,4000)	0.9358821
amount_bin[4000,5200)	-0.8478681
amount_bin[5200,6800)	0.9221348
amount_bin[6800, Inf)	-0.5604641
saving_acct_bin100 <= ... < 500 DM	0.1693752
saving_acct_bin500 <= ... < 1000 DM%,%... >= 1000 DM%,%...	0.7395558
present_emp_bin1 <= ... < 4 years	0.5463963
present_emp_bin4 <= ... < 7 years%,%... >= 7 years	0.9116666
present_emp_binunemployed	0.5082517
property_bincar or other, not in attribute Savings account/b...	-0.2828197
property_binreal estate	0.1112562
property_binunknown / no property	-0.6156298
age_bin[26,28)	1.0368892
age_bin[28,35)	0.3978480
age_bin[35,37)	1.6460973
age_bin[37, Inf)	0.8736893
housing_binown	0.3086795
housing_binrent	-0.2004767

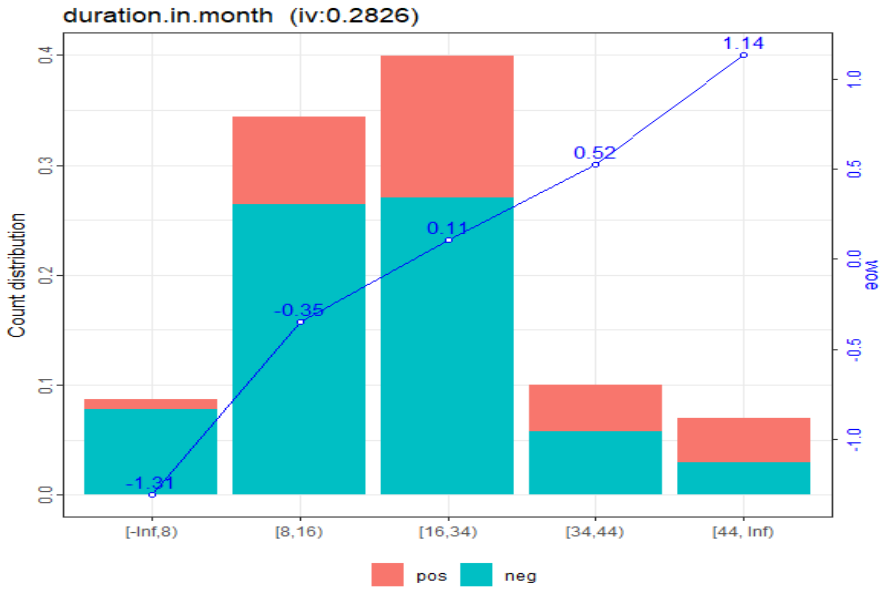
Rysunek 31: Parametry regreji modelu M3 - opracowanie własne

Zauważmy, że w modelu M3 uzyskujemy wiele parametrów regresji do estymowania. Aby zmniejszyć ich liczbę, każdej klasie pochodzącej z danej zmiennej przyporządkowujemy odpowiadającą jej wartość WoE. W ten sposób zamieniamy każdą zmienną dyskretną na ciągłą i otrzymujemy model M4.

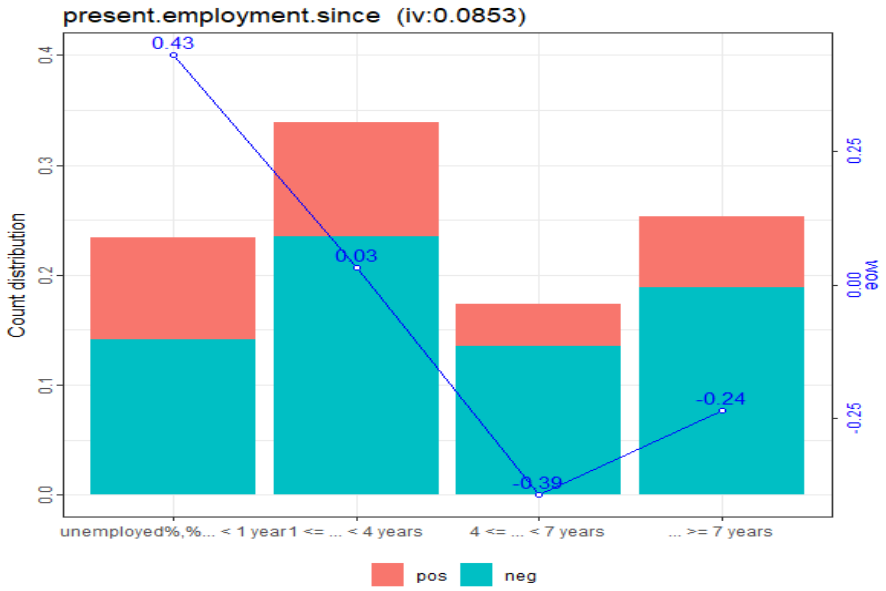
	Oszacowanie
(Intercept)	0.882
chk_acct_woe	-0.844
duration_woe	-0.796
credit_his_woe	-0.598
purpose_woe	-1.201
amount_woe	-0.848
saving_acct_woe	-0.765
present_emp_woe	-0.903
property_woe	-0.422
age_woe	-1.040
housing_woe	-0.610

Tabela 14: Oszacowania współczynników w modelu M4

Poniżej przedstawiamy estymowane wartości współczynników w modelu M4. Widzimy, że wcześniejsze wnioski na temat ujemnych współczynników okazały się prawdziwe.



Rysunek 32: UZUPEŁNIĆ - opracowanie własne



Rysunek 33: UZUPEŁNIĆ - opracowanie własne

7 ZADANIE 6

Biorąc pod uwagę zmienne wybrane w poprzednim zadaniu dokonamy predykcji wartości zmiennej *creditability* w oparciu o model regresji logistycznej ze wszystkimi predyktorami typu dyskretnego oraz w oparciu o model regresji logistycznej ze wszystkimi predyktorami typu ciągłego.

Tabela (21) jest wyznaczoną macierzą pomyłek na podstawie modelu regresji logistycznej ze wszystkimi predyktorami typu dyskretnego.

	0	1
bad	43	55
good	38	183

Tabela 15: Macierz pomyłek dla danych testowych w modelu M3

Tabela (22) jest wyznaczoną macierzą pomyłek na podstawie modelu regresji logistycznej ze wszystkimi predyktorami typu ciągłego z "nowymi" wartościami tych zmiennych, które odpowiadają WoE.

	0	1
bad	46	52
good	41	180

Tabela 16: Macierz pomyłek dla danych testowych w modelu M4

accuracy	error	sensitivity	specificity	positive prediction	negative prediction
0.7084639	0.2915361	0.8280543	0.4387755	0.7689076	0.5308642

Tabela 17: Tabela z wyznaczonymi wartościami metryk dla modelu M3

accuracy	error	sensitivity	specificity	positive prediction	negative prediction
0.7018634	0.2981366	0.8144796	0.4693878	0.7758621	0.5287356

Tabela 18: Tabela z wyznaczonymi wartościami metryk dla modelu M4

Obydwie tabele (17), (18) zawierają wyznaczone z macierzy pomyłek wartości konkretnych metryk. Widzimy, że wartości te są bardzo zbliżone. Zakładamy zatem, że między predykcją wartości zmiennej *creditability* na podstawie modelu regresji logistycznej ze wszystkimi predyktorami typu dyskretnego oraz w oparciu o model regresji logistycznej ze wszystkimi predyktorami typu ciągłego nie ma zbyt dużych różnic.

8 ZADANIE 7

Stosując cztery różne kryteria oceny efektywności modelu scoringowego dokonamy porównania modeli scoringowych uzyskanych w poprzednich zadaniach.

Porównane zostaną następujące cztery modele regresji logistycznej:

- M1 - oparty na "wyjściowych" zmiennych objaśniających i najlepszy na podstawie AIC
- M2 - oparty na "wyjściowych" zmiennych objaśniających i najlepszy na podstawie BIC
- M3 - oparty na skategoryzowanych zmiennych objaśniających,
- M4 - oparty na zmiennych objaśniających, którym przypisano WoE

Ustalmy, że F_m oraz G_n są dystrybuantami empirycznymi. Statystyką Kołmogorowa-Smirnowa nazywamy statystykę

$$D_{mn} = sup_{t \in \mathbf{R}} |F_m(t) - G_n(t)|$$

Jeśli próba pochodzi z rozkładu o dystrybuancie $G_n(t)$ to D_{mn} dąży prawie wszędzie do zera. Z definicji statystyki Kołmogorowa-Smirnowa wynika, że porównując dwie funkcje scoringowe za lepszą uznamy tę funkcję, której odpowiada większa wartość tej statystyki, bo lepiej "rozdziela" dobrych klientów od złych.

Pierwszym sposobem oceny efektywności modelu scoringowego jest macierz pomyłek oraz związane z nią metryki - dokładność, błąd, prawdziwie pozytywna wartość (czułość), prawdziwie negatywna wartość (specyficzność), precyja przewidywania pozytywnego oraz precyzja przewidywania negatywnego.

	0	1
bad	51	47
good	40	181

Tabela 19: macierz pomyłek dla danych testowych w modelu M1

	0	1
bad	39	59
good	34	187

Tabela 20: macierz pomyłek dla danych testowych w modelu M2

	0	1
0	43	55
1	38	183

Tabela 21: macierz pomyłek dla danych testowych w modelu M3

	0	1
0	46	52
1	41	180

Tabela 22: macierz pomyłek dla danych testowych w modelu M4

	accuracy	error	sensitivity	specificity	positive prediction	negative prediction
M1	0.7272727	0.2727273	0.8190045	0.5204082	0.7938596	0.5604396
M2	0.7084639	0.2915361	0.8461538	0.3979592	0.7601626	0.5342466
M3	0.7084639	0.2915361	0.8280543	0.4387755	0.7689076	0.5308642
M4	0.7018634	0.2981366	0.8144796	0.4693878	0.7758621	0.5287356

Tabela 23: Tabela z wyznaczonymi wartościami metryk wszystkich modeli

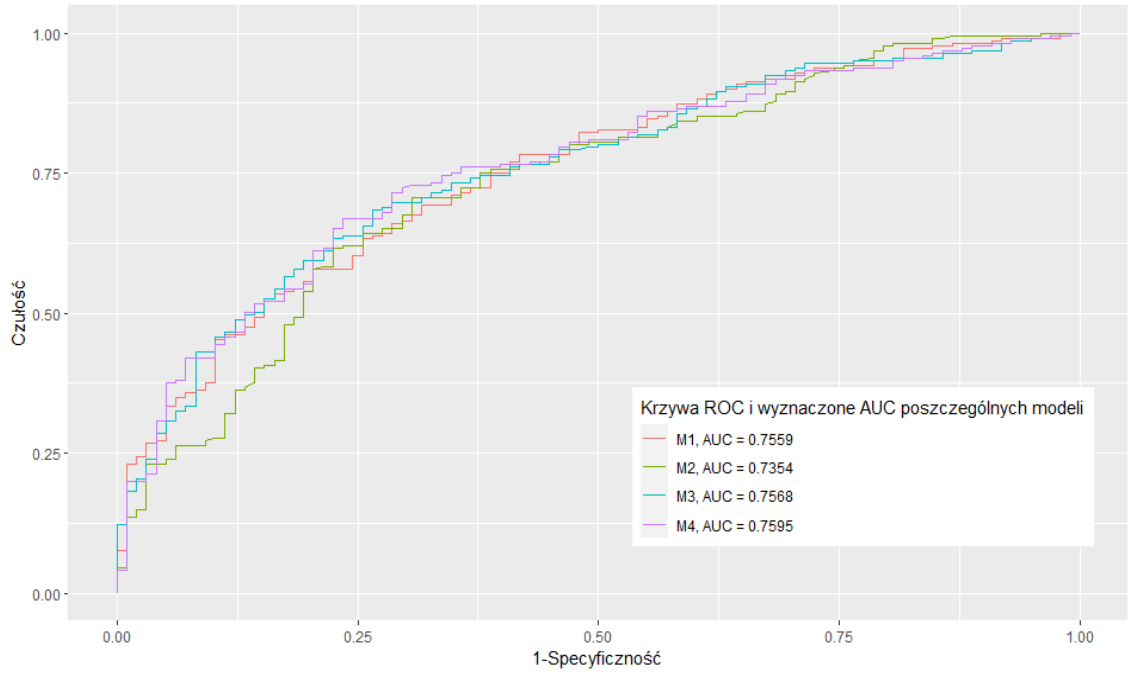
Na powyższej tabeli widzimy, że najlepszą dokładność, precyzje posiada model M1. Wskaźnik czułości najlepszy jest dla modelu M2. Duża swoistość pokazuje, że klasyfikator rzadko się myli, jeśli chodzi o negatywne przypadki. Tak, więc jeśli pokaże, że coś jest pozytywne, to możemy z dużym prawdopodobieństwem się spodziewać, że takie rzeczywiście jest. W naszym przypadku największą swoistość posiada model M1, jednakże jest ona na poziomie 0.5 co nie jest zbyt dobrym wynikiem. Postivie prediction value powinna być wartością jak najbliższą 1, tak samo, jak wartość negative prediction value. Dokładnie widzimy, że najbliższe 1 wartości występują w modelu $M1$.

Możemy zatem podsumować, że najlepszym modelem jest model $M1$. Jednakże nie przeważa on znacząco nad pozostałymi modelami - w gruncie rzeczy, wszystkie wartości we wszystkich modelach są do siebie bardzo zbliżone.

Drugi sposób oceny efektywności modelu scoringowego opiera się na separability measures rozkładów klientów dobrych i złych. Związane są one z krzywą ROC oraz statystyką Kołmogorowa-Smirnowa. Statystyka Kołmogorowa-Smirnowa została wyjaśniona na początku tego zadania. Zdefiniujemy teraz krzywą ROC. Krzywa ROC obrazuje, jak duży będzie odsetek błędnych klasyfikacji (pozytywnych i negatywnych) dla danego punktu odcięcia.

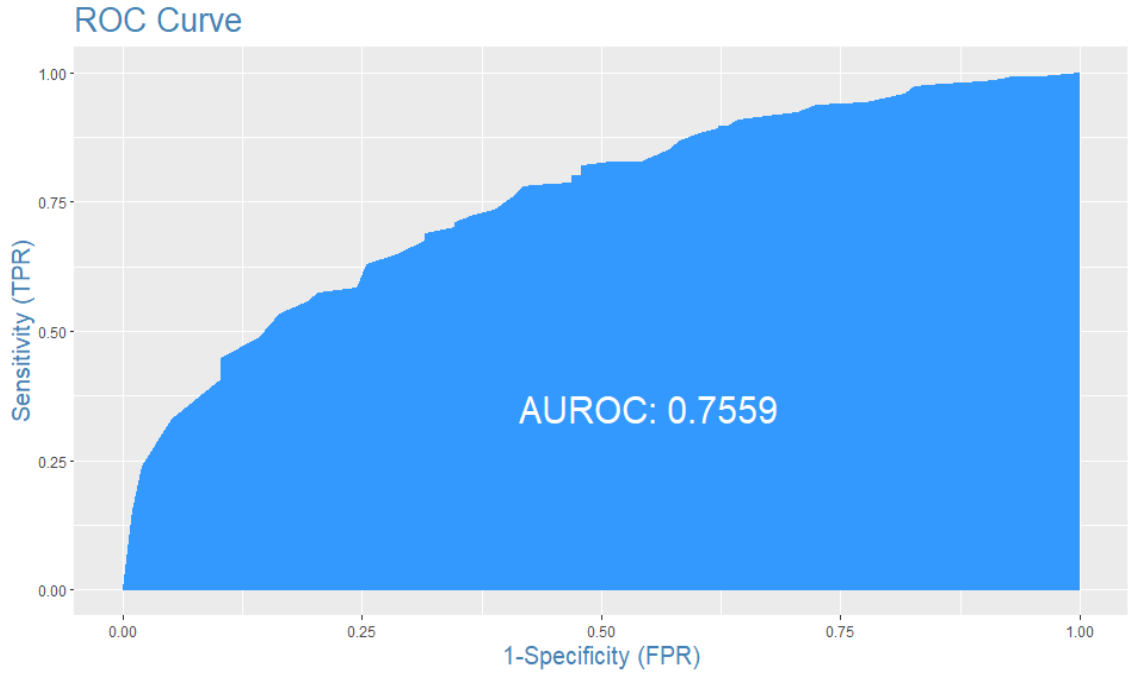
Za każdym razem w rozważanych u nas przypadkach, oś pionowa będzie czułością, zaś oś pozioma 1 - specyficzność. Klasyfikator idealny zatem wynosiłby stale 0 na osi pionowej oraz 1 na osi poziomej. Stwierdzamy zatem, że lepszym modelem będzie zawsze ten położony powyżej innych modeli.

Poniżej dokonamy porównania krzywych ROC oraz ich wartości AUC dla wszystkich modeli $M1$, $M2$, $M3$ i $M4$.

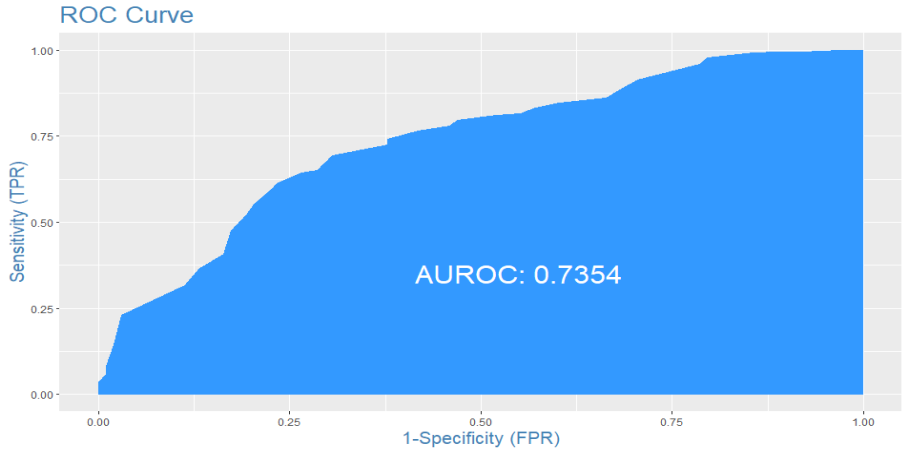


Rysunek 34: Krzywe ROC wszystkich modeli wraz z wyznaczonymi wartościami AUC- opracowanie własne

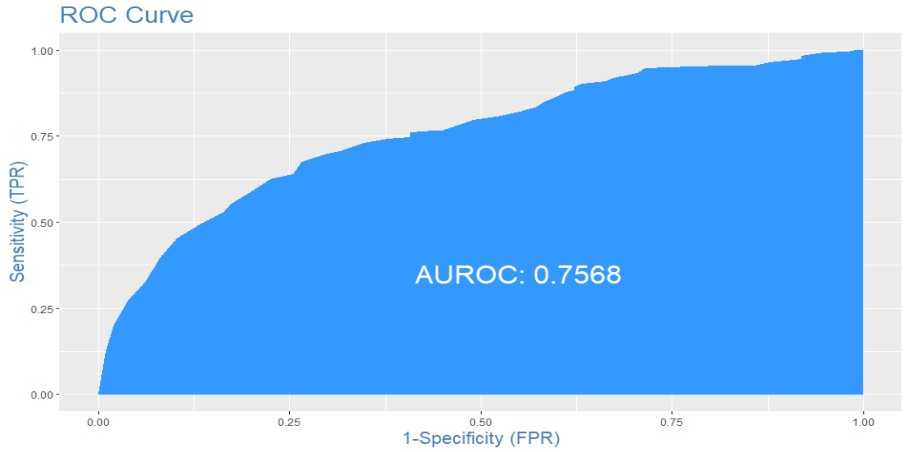
Powinnyśmy wybrać model położony powyżej innych. Przyglądając się powyższemu rysunkowi widzimy, że nie jesteśmy w stanie wybrać takowego, ponieważ (zgodnie ze wcześniejszymi wnioskami) modele są do siebie bardzo zbliżone, zatem podają bardzo zbliżone krzywe. Dalej prezentujemy cztery wykresy krzywych ROC poszczególnych modeli wraz z wyznaczonymi wartościami AUC na nich.



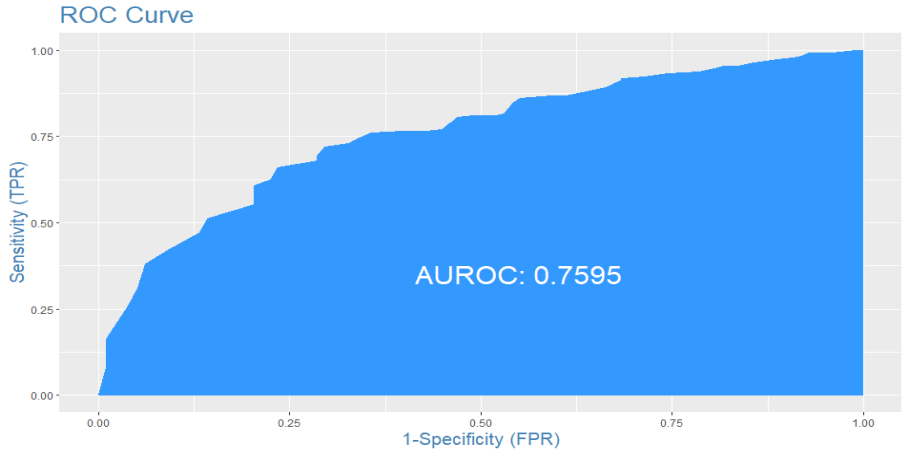
Rysunek 35: Wykres krzywej ROC modelu M1 - opracowanie własne



Rysunek 36: Wykres krzywej ROC modelu M2 - opracowanie własne



Rysunek 37: Wykres krzywej ROC modelu M3 - opracowanie własne



Rysunek 38: Wykres krzywej ROC modelu M4 - opracowanie własne

	M1_AUC	M2_AUC	M3_AUC	M4_AUC
AUC	0.7559101	0.7354326	0.7567873	0.7595346

Tabela 24: Wartości AUC poszczególnych modeli

Po analizie powyższych rysunków (34) – (38) oraz tabeli (24) dochodzimy do wniosku, że najlepszym modelem jest model $M4$, ponieważ wartość jego pola AUC jest największa. Jednakże, wartości AUC bardzo nieznacznie różnią się dla wszystkich modeli, dlatego wybierając obojętnie który model, nie popełnimy dużego błędu.

Jak już wspomnieliśmy wcześniej, statystyka Kołmogorowa-Smirnowa jest testem nieparametryczny używanym do porównywania rozkładów jednowymiarowych cech statystycznych. Większa wartość tego testu wskazuje na lepsze rozróżnianie klientów dobrych od złych. Największą wartość KS ma model $M4$, co pokrywa się z naszymi wcześniejszymi wnioskami z wyznaczonych wartości AUC.

	AUC	KS
M1	0.7559	0.3583
M2	0.7354	0.3852
M3	0.7568	0.3999
M4	0.7595	0.4122

Tabela 25: Wartości AUC oraz statystyki Kołmogorowa-Smirnowa.

9 ZADANIE 8

Dla każdego z wybranych modeli wyznaczmy optymalne punkty odcięcia.
Punkt odcięcia najprościej możemy zdefiniować jako granicę w prawdopodobieństwie klasyfikacji, pomiędzy klasą dobra i złą.

W celu wyznaczenia optymalnego punktu odcięcia skorzystamy z biblioteki `InformationValue` oraz funkcji `optimalCutoff` [10]. Funkcja ta określa optymalny próg wyniku prawdopodobieństwa przewidywania na podstawie konkretnych celów problemu.

Wyróżniamy cztery rodzaje optymalnego odcięcia:

- *Ones* - maksymalizuje wykrywanie zdarzeń lub jedynek
- *Zeros* - maksymalizuje wykrywanie zdarzeń niebędących zdarzeniami lub zer
- *Both* - kontroluje odsetek wyników fałszywie dodatnich i fałszywie ujemnych stopy poprzez maksymalizację wskaźnika J. Youdena
- *Misclasserror* - minimalizuje błąd błędnej klasyfikacji

Tabela (26) wskazuje nam, jakie punkty odcięcia zostały wybrane dla danej metody w konkretnych modelach.

[10] Więcej informacji dostępnych pod `http://r-statistics.co/Information-Value-With-R.html`

	Ones	Zeros	Both	missclasserror
M1	0.1097	0.9897	0.8097	0.3197
M2	0.2548	0.9748	0.6748	0.3848
M3	0.0466	0.9748	0.7366	0.2666
M4	0.0416	0.9748	0.6916	0.3316

Tabela 26: Punkty odcięcia dla każdego z modeli

	accuracy	error	sensitivity	specificity	pp	np
M1	0.7272727	0.2727273	0.8190045	0.5204082	0.7938596	0.5604396
M1.cut.off.both	0.6458	0.3542	0.5791855	0.7959184	0.8648649	0.4561404
M2	0.7084639	0.2915361	0.8461538	0.3979592	0.7601626	0.5342466
M2.cut.off.both	0.7022	0.2978	0.7058824	0.6938776	0.8387097	0.5112782
M3	0.7084639	0.2915361	0.8280543	0.4387755	0.7689076	0.5308642
M3.cut.off.both	0.6928	0.3072	0.6742081	0.7346939	0.8514286	0.5
M4	0.7018634	0.2981366	0.8144796	0.4693878	0.7758621	0.5287356
M4.cut.off.both	0.7179	0.2821	0.7239819	0.7040816	0.8465608	0.5307692

Tabela 27: Tabela z wyznaczonymi wartościami metryk wszystkich modeli i porównanie z metodą punktów odcięcia

10 PODSUMOWANIE

W sprawozdaniu dokonaliśmy dokładnej analizy zbioru `germancredit`. Nauczyliśmy się dzielić dane w odpowiednich proporcjach na zbiory testowe oraz treningowe, a także, na podstawie zbioru uczącego dokonywaliśmy wyboru modelu regresji logistycznej. Analizowaliśmy kryteria AIC oraz BIC, sprawdzaliśmy, kiedy model zwraca najlepsze wartości. Nauczyliśmy się wyznaczać macierze pomyłek oraz analizować wynikające z nich własności - dokładność, błąd, precyzję przewidywania pozytywnego, precyzję przewidywania negatywnego, czułość oraz specyficzność. Ponadto dowiedzieliśmy się, na czym polega kryterium IV oraz WOE, a także zdefiniowaliśmy i określiliśmy różne rodzaje punktów odcięcia.

Można dojść do ogólnego wniosku, że przy badaniu jakichkolwiek danych nie powinno polegać się jedynie na jednej metodzie, a sprawdzać kilka, nawet jeśli wartości wychodzą podobne.

11 BIBLIOGRAFIA

1. **Alicja Jokiel - Rokita** *Wykłady ze Statystyki w Finansach i Ubezpieczeniach*
2. **Kamil Bogus** *Rzeczy udostępnianie po laboratoriach*