

Credit Risk

Filip Mordarski & Mateusz Wasielewski

5 12 2020

Spis treści

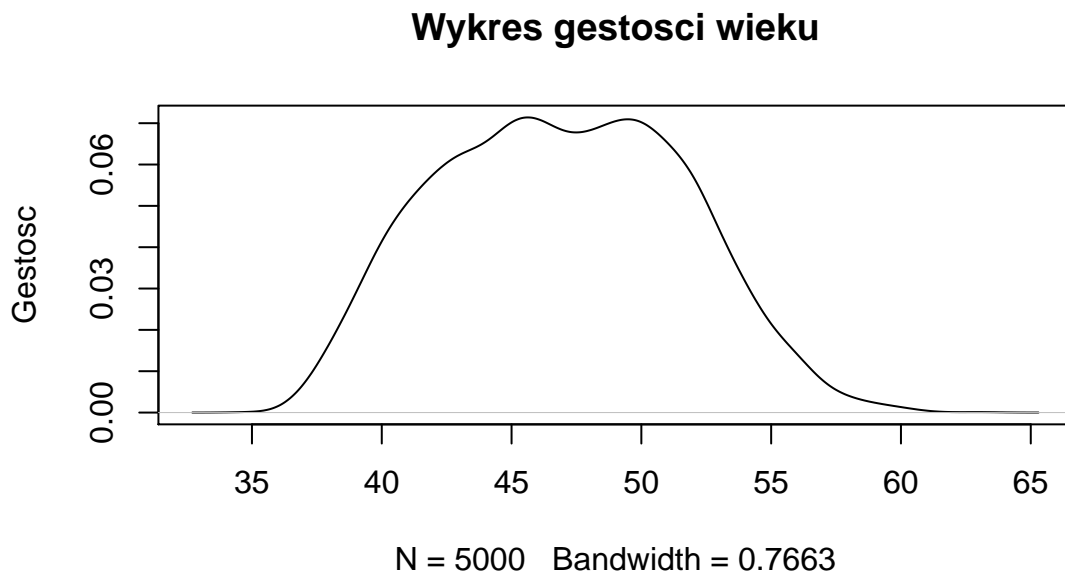
| | |
|--|----|
| Wstęp | 1 |
| Tworzenie zmiennych objaśniających | 2 |
| Badanie rozkładów zmiennych | 6 |
| Winsoryzacja zmiennych zaburzonych nietypowymi wartościami | 6 |
| Szacowanie parametrów modelu | 7 |
| Ewaluacja modelu | 9 |
| Bootstrapping | 11 |

Wstęp

Poniższy raport będzie zawierał analizę modelu ryzyka kredytowego, oszacowanego na podstawie losowo wygenerowanego zbioru danych. Pierwsza część raportu będzie opisywała sposób losowania oraz rozkłady zmiennych objaśniających. Została przedstawiona również wizualizacja tych zmiennych. Następnie, została przeprowadzona ‘winsoryzacja’ w celu poprawy wyników naszej analizy. W kolejnej części raportu, zostały oszacowane dwa modele, ze zmienną objaśnianą, mówiącą czy dany klient doświadczy ‘defaultu’ czy nie. W ostatniej części raportu została przeprowadzona ewaluacja obu modeli, krzywa ROC oraz bootstrapowe przedziały ufności.

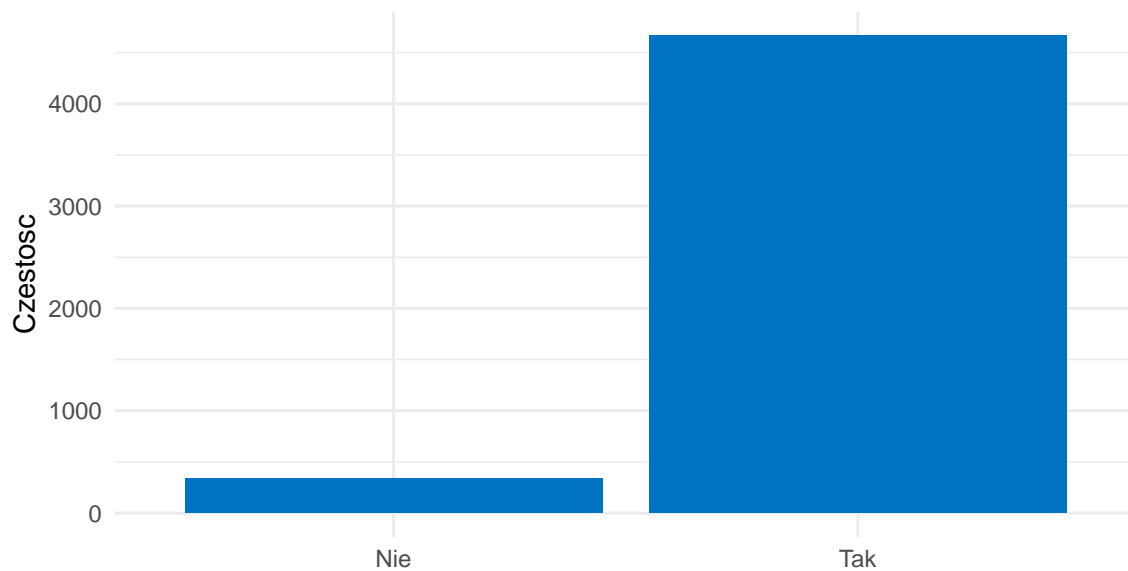
Tworzenie zmiennych objaśniających

W pierwszej kolejności zostały utworzone zmienne objaśniające potrzebne do modelu PD. Pierwszą zmienną, która została wygenerowana na podstawie utworzonych wcześniej zmiennych jest wiek. Został wyliczony okres trwania umowy w latach na podstawie różnicy między obecną datą a wartością w zmiennej *agreement_start*. Następnie został wygenerowany wektor wartości z rozkładu gamma z parametrem kształtu równym 3 oraz parametrem skali równym 2. Wiek został wyznaczony dodając do siebie: czas trwania umowy, liczbę 18 (wiek kiedy człowiek może podpisać wiążącą umowę kredytową) oraz wylosowaną wartość z rozkładu gamma, oznaczającą różnicę w latach pomiędzy datą podpisania umowy a osiągnięciem pełnoletności. Poniższy wykres przedstawia gęstość tej utworzonej zmiennej w naszym zbiorze.



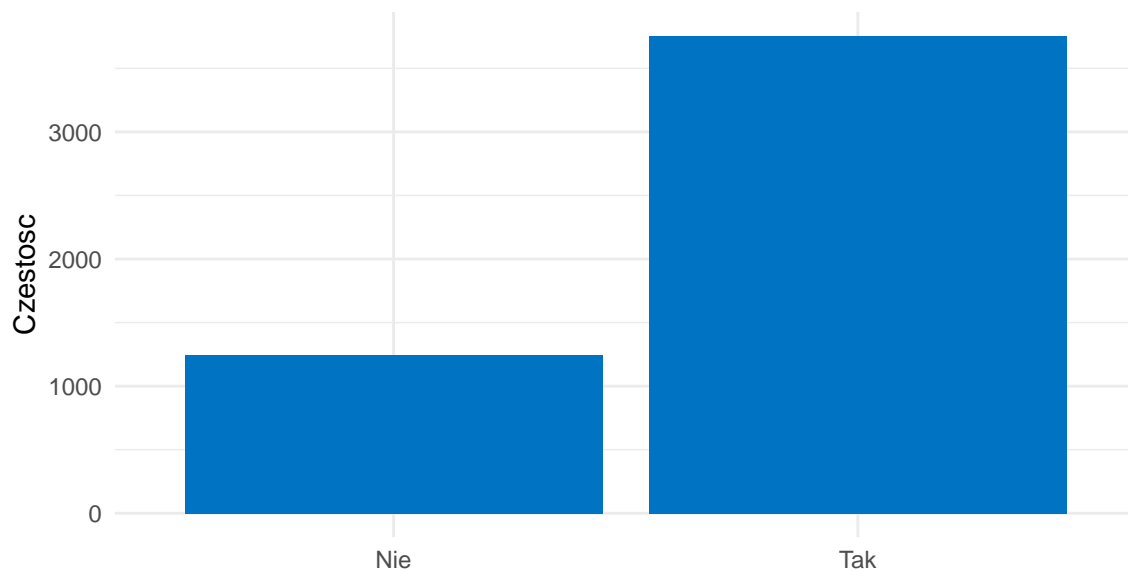
Następnie została wygenerowana zmienna, która określa czy dany pracownik jest zatrudniony, czy też nie. Prawdopodobieństwo bezrobocia zostało ustalone na poziomie 6.7 %. Wartość ta odzwierciedla średnią stopę bezrobocia w 2019 roku w Stanach Zjednoczonych. Poniższy wykres przedstawia histogram tej zmiennej w zbiorze.

Histogram zmiennej, określająca czy klient jest zatrudniony



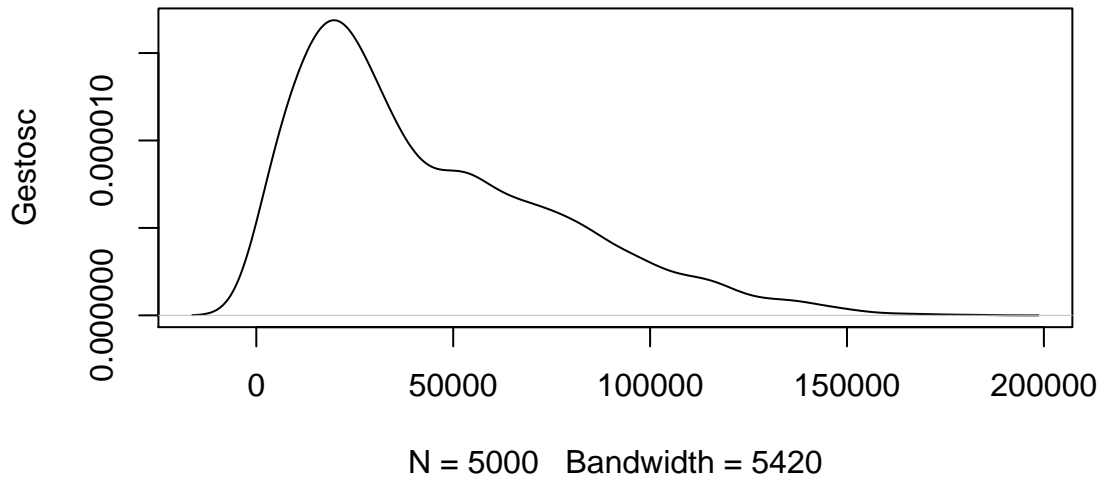
Na podstawie zmiennej, określającej czy dana osoba jest zatrudniona, wygenerowano zmienną czy dana osoba jest zatrudniona na pełny etat. Prawdopodobieństwo tego wynosi 80 %. Poniższy wykres przedstawia histogram tej zmiennej.

Histogram zmiennej, określająca czy klient pracuje na pełen etat



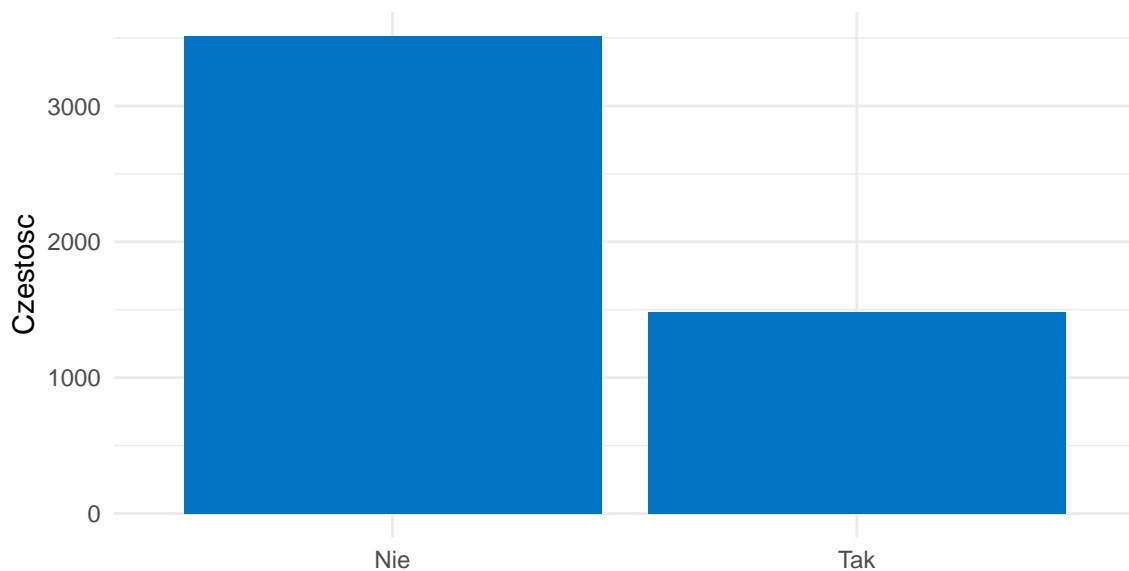
Zmienną, która z pewnością może okazać się istotna w tworzeniu modelu PD jest dochód roczny danej osoby. Wartości te zostały wylosowane z rozkładu normalnego. Średnia dla osób zatrudnionych na pełen etat została ustalona na poziomie 48000 USD z odchyleniem standardowym na poziomie 15000 USD. Dla osób niezatrudnionych na pełen etat wartość średnia została ustalona na poziomie 20000 USD, natomiast odchylenie 10000 USD. Poniżej zaprezentowano wykres gęstości tej zmiennej.

Wykres gestosci zarobków



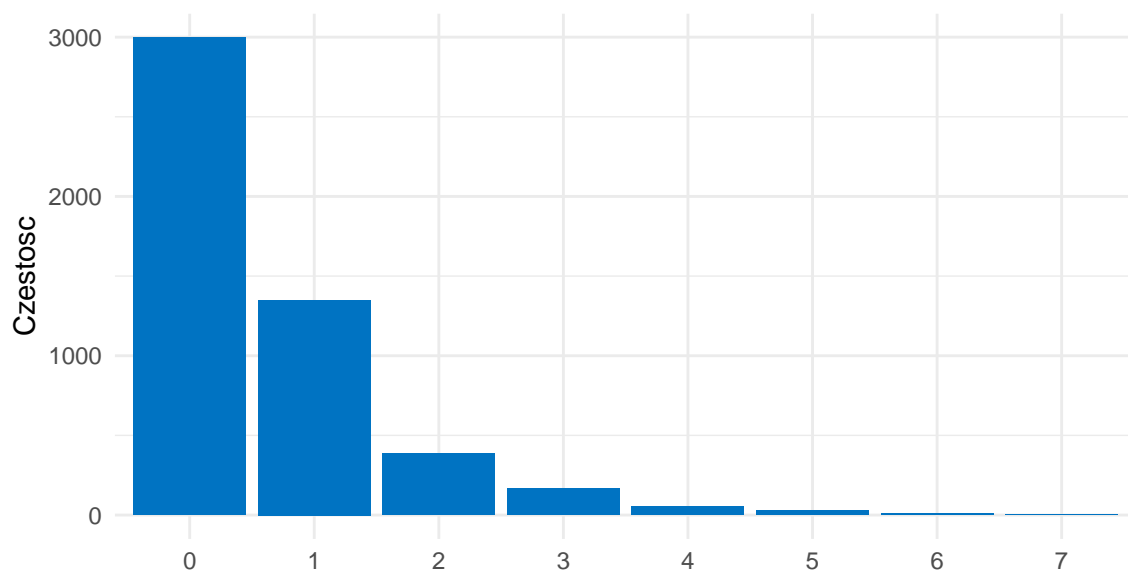
Kolejno, została wygenerowana zmienna, mówiąca o tym czy dana osoba jest singlem, czy żyje w związku z inną osobą. Prawdopodobieństwo, że ktoś jest singlem w Stanach Zjednoczonych wynosi 30 %. Przedstawiono histogram tej zmiennej w zbiorze.

Histogram zmiennej, określająca czy klient jest singlem



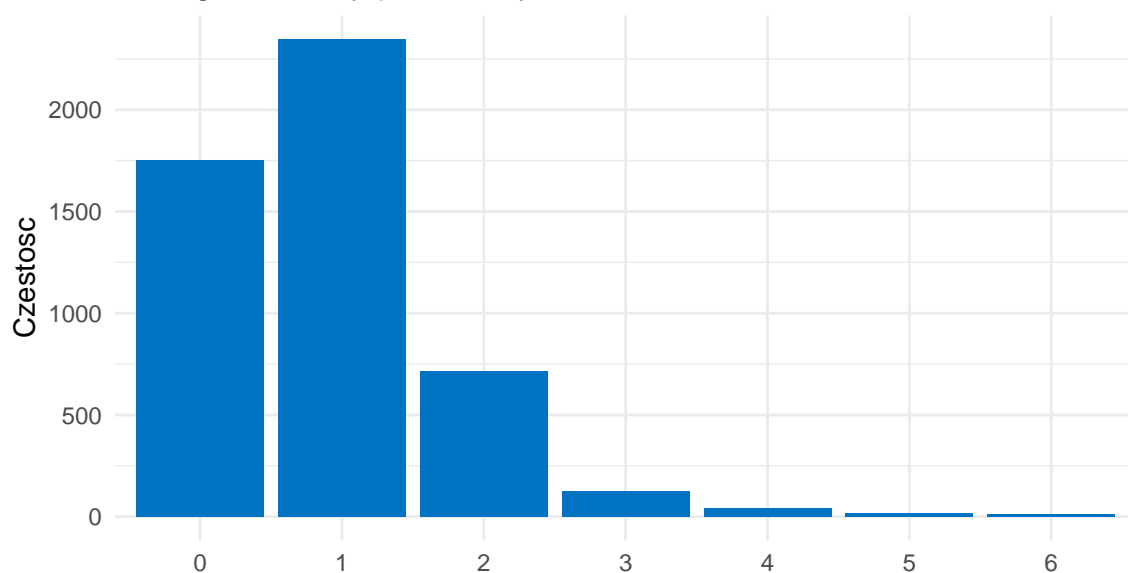
Następnie, została wygenerowana liczba posiadanych dzieci przez daną osobą. Zmienna ta została wygenerowana na podstawie zmiennej, określającej czy dana osoba jest singlem czy nie jest. Poniżej przedstawiono histogram tej wygenerowanej zmiennej.

Histogram liczby posiadanych dzieci

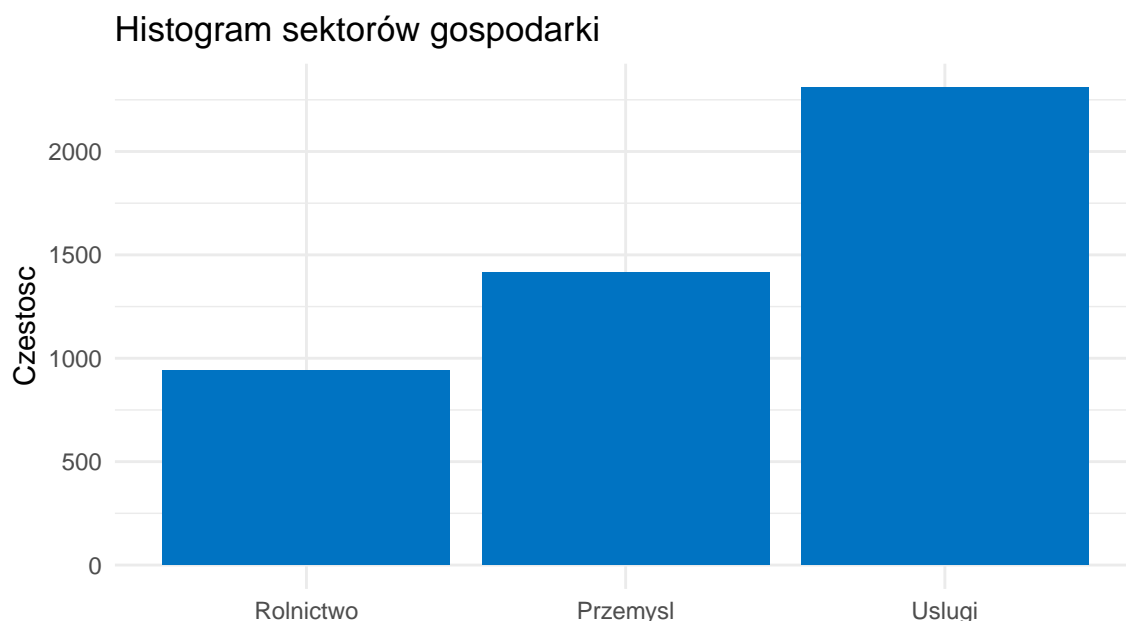


Kolejno, została wygenerowana zmienna, mówiąca o liczbie posiadanych samochodów przez klienta. Dla osób nieposiadających dzieci lub mających jedno dziecko, liczba ta została wylosowana z następującego zakresu: $[0, 1, 2]$ z prawdopodobieństwami równymi kolejno: $[40\%, 50\%, 10\%]$. Dla klientów mających więcej niż jedno dziecko, liczba samochodów zależy od liczby posiadanych dzieci i jest obliczona za pomocą następującej formuły: $\text{liczba_dzieci} - [\text{wartość z losowania liczb } [1,2] \text{ z } 50\% \text{ prawdopodobieństwami}] + 1$. Poniżej przedstawiono histogram liczby posiadanych samochodów przez klientów.

Histogram liczby posiadanych samochodów



Następnie, została wygenerowana zmienna porządkowa, określająca sektor gospodarki, w którym pracują klienci. Prawdopodobieństwo wystąpienia następujących sektorów $[\text{rolnictwo}, \text{przemysł}, \text{usługi}]$ wśród zatrudnionych wynosi kolejno: $[20\%, 30\%, 50\%]$. Poniżej przedstawiono histogram sektorów gospodarki.



Badanie rozkładów zmiennych

W celu sprawdzenia czy zmienne objaśniające mają wiele wartości odstających, zbadano kurtozę każdej zmiennej ciągłej. Przyjęto, że dla tej miary spłaszczenia tolerowaną wartością będzie 3.

Tablica 1: Zestawienie wartości kurtozy zmiennych objaśniających

| value_mortgage | value_nonmortgage | age | annual_income |
|----------------|-------------------|----------|---------------|
| 5.601227 | 5.420217 | 2.310545 | 3.329684 |

Z powyższej tabeli wynika, że dla wartości kredytu hipotecznego oraz wartości kredytu bez hipoteki kurtozy wyniosły ok. 5.5, a co za tym idzie, przekraczają ustaloną wartość graniczną. Jest to spowodowane licznymi wartościami odstającymi, których wpływ na analizę trzeba zminimalizować. Niepokojąca wartość kurtozy występuje zarówno przy zmiennej roczny przychód, ponieważ wynosi powyżej 3. Istnieją różne metody radzenia sobie z tzw. “outliersami”, jedna z nich zostanie przedstawiona w następnym akapicie. Dla zmiennej wiek wartość kurtozy jest mniejsza od 3, co pozwala zakładać normalność rozkładu tej zmiennej.

Winsoryzacja zmiennych zaburzonych nietypowymi wartościami

Podczas procesu winsoryzacji wartości odstające nie są usuwane, a jedynie podmieniane na ostatnie wartości znajdujące się w nieobciętym obszarze, dzięki czemu nie tracimy liczby obserwacji. W tym przypadku odcięte zostały skrajne obszary 5-procentowe. Poniższa tabela prezentuje wartości kurtozy po winsoryzacji tych trzech zmiennych ciągłych, których wspomniana miara spłaszczenia była większa od 3.

Tablica 2: Wartości kurtozy zmiennych po winsoryzacji

| value_mortgage_win | value_nonmortgage_win | annual_income_win |
|--------------------|-----------------------|-------------------|
| 3.290116 | 2.645227 | 2.36553 |

Można zauważyć znaczną poprawę rozkładów tych dwóch zmiennych, co pozytywnie wpłynie na dalszą analizę.

Szacowanie parametrów modelu

Pierwszym krokiem podczas szacowaniu modelu było podzielenie zbioru na dwa podzbiory: treningowy (3504 obserwacje) i testowy (1496 obserwacje).

Ze względu na fakt, że zmienna objaśniana jest binarna, zdecydowano się na oszacowanie modelu logitowego.

```
##
## Call:
## glm(formula = default ~ value_mortgage_win + value_nonmortgage_win +
##      age + employed + full_time + annual_income_win + single +
##      kids + car + sector, family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7558  -0.6788  -0.6532  -0.6177   1.9079
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.59632526684   0.46437721584  -3.438 0.000587 ***
## value_mortgage_win    0.00000006073   0.00000018179   0.334 0.738310
## value_nonmortgage_win  0.00000437555   0.00000689982   0.634 0.525980
## age              0.00370857496   0.00908012157   0.408 0.682960
## employed1       -0.12203374224   0.19371197621  -0.630 0.528711
## full_time1       0.17025274439   0.12137940106   1.403 0.160721
## annual_income_win -0.00000216603   0.00000153150  -1.414 0.157270
## single1          0.00265541768   0.08469359886   0.031 0.974988
## kids            -0.02170332446   0.05406887797  -0.401 0.688125
## car              0.06963291349   0.06005722235   1.159 0.246276
## sector1         -0.05293487243   0.11720198555  -0.452 0.651517
## sector2          0.03919600644   0.10056017130   0.390 0.696702
## sector3                  NA                  NA      NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 3508.8  on 3525  degrees of freedom
## Residual deviance: 3503.0  on 3514  degrees of freedom
## AIC: 3527
##
## Number of Fisher Scoring iterations: 4
```

Niestety żaden z regresorów nie ma istotnego wpływu na zmienną objaśnianą, co może być spowodowane losowymi wartościami zmiennych. W przypadku zmiennej `sector=3` można zaobserwować wartości NA dla współczynników równania. Powodem tego jest fakt, że zmienna `sector` jest zmienną porządkową i przyjmuje 4 wartości: 0,1,2,3. Model estymując współczynniki zmiennych porządkowych najpierw rozбивa każdą wartość na zmienną binarną, następnie pomija jedną z nich w modelu - w tym wypadku pominął `sector=0`. Istotna w tym wyjaśnieniu jest pewna zależność występująca w naszym zbiorze danych: zmienna `sector` przyjmuje wartość 0 wtedy i tylko wtedy gdy zmienna `employed`, opisująca czy dany pracownik jest zatrudniony, również przyjmuje wartość 0. Z tego powodu model nie tylko pominął zmienną `sector0`, ale również zagregował ją w zmiennej `employed=0`. W wyniku tego współczynniki zmiennej `sector=3` nie mogą zostać wyestymowane, gdyż teraz to ona jest wykluczoną wartością zmiennej porządkowej.

Oszacowano również drugi model, w którym zmieniliśmy założenie co do rozkładu składnika losowego. Zamiast transformacji logistycznej zastosowano odwrotność dystrybucyjny rozkładu normalnego. W wyniku tego, został oszacowany model probitowy.

```
##
## Call:
## glm(formula = default ~ value_mortgage_win + value_nonmortgage_win +
##      age + employed + full_time + annual_income_win + single +
##      kids + car + sector, family = binomial(link = "probit"),
##      data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7552  -0.6790  -0.6534  -0.6172   1.9095
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.96947681101  0.26501925687  -3.658 0.000254 ***
## value_mortgage_win  0.00000003432  0.00000010399   0.330 0.741377
## value_nonmortgage_win 0.00000258572  0.00000394658   0.655 0.512353
## age              0.00227087192  0.00518114158   0.438 0.661172
## employed1        -0.06935488691  0.11051063558  -0.628 0.530275
## full_time1        0.09727880196  0.06892840220   1.411 0.158156
## annual_income_win -0.00000123123  0.00000087292  -1.410 0.158399
## single1           0.00192308893  0.04832555636   0.040 0.968257
## kids             -0.01212469384  0.03096661776  -0.392 0.695398
## car              0.03927093234  0.03433777034   1.144 0.252762
## sector1          -0.03026457772  0.06652167311  -0.455 0.649139
## sector2           0.02164358760  0.05752642937   0.376 0.706740
## sector3              NA              NA         NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 3508.8  on 3525  degrees of freedom
## Residual deviance: 3503.0  on 3514  degrees of freedom
## AIC: 3527
##
## Number of Fisher Scoring iterations: 4
```

Tutaj podobnie jak w przypadku estymacji modelu logitowego pojawiają się wartości NA przy współczynnikach zmiennej `sector=3`.

Ewaluacja modelu

Po oszacowaniu obu modeli, dokonano ich ewaluacji. Poniżej przedstawiono tabelę kontyngencji dla pierwszego modelu logitowego.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 1156  318
##           1    0    0
##
##           Accuracy : 0.7843
##           95% CI : (0.7624, 0.805)
##       No Information Rate : 0.7843
##       P-Value [Acc > NIR] : 0.515
##
##           Kappa : 0
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##           Sensitivity : 1.0000
##           Specificity : 0.0000
##       Pos Pred Value : 0.7843
##       Neg Pred Value :    NaN
##           Prevalence : 0.7843
##       Detection Rate : 0.7843
##   Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##       'Positive' Class : 0
##
```

Następnie wygenerowano podobną tabelę kontyngencji dla drugiego modelu - probitowego.

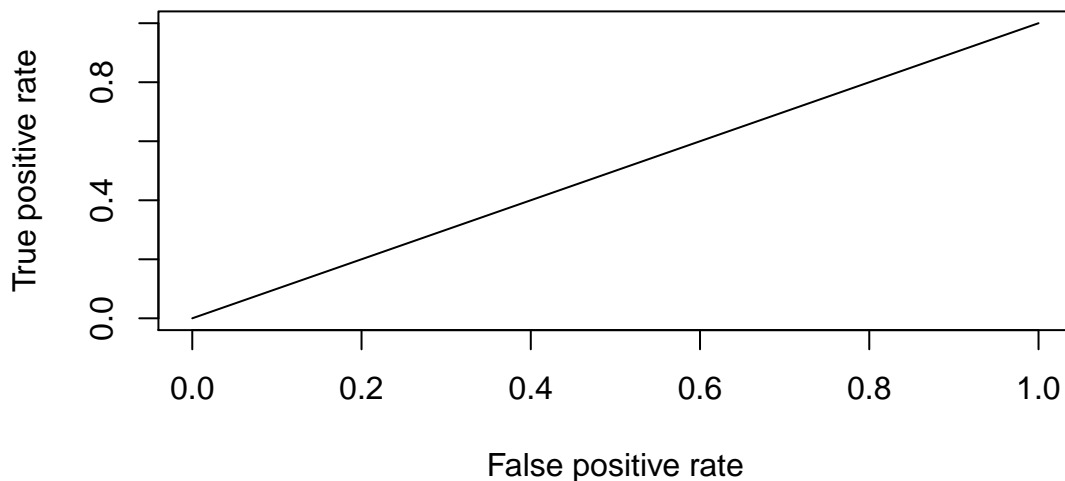
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 1156  318
##           1    0    0
##
##           Accuracy : 0.7843
##           95% CI : (0.7624, 0.805)
##       No Information Rate : 0.7843
##       P-Value [Acc > NIR] : 0.515
##
##           Kappa : 0
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##           Sensitivity : 1.0000
##           Specificity : 0.0000
```

```
##          Pos Pred Value : 0.7843
##          Neg Pred Value :   NaN
##          Prevalence : 0.7843
##          Detection Rate : 0.7843
##          Detection Prevalence : 1.0000
##          Balanced Accuracy : 0.5000
##
##          'Positive' Class : 0
##
```

Model ten dostarcza takich samych predykcji jak model pierwszy. Z tego względu dalsza analiza będzie opierać się na oszacowaniach jednego z powyższych modeli.

Na podstawie powyższych tabel możemy stwierdzić, że nasze modele nie są efektywne w predykcji 'defaultu'. Wszystkie obserwacje zostały dopasowane do kategorii 0, czyli brak 'defaultu'. Mamy relatywnie wysoki poziom trafności modeli, ponieważ jest on na poziomie 78 %. Przyczyną takiego stanu rzeczy, jest to że 78% obserwacji ze zbioru testowego nie miało 'defaultu'. Ewaluacja tych modeli wykazała, że nie powinniśmy wyciągać żadnych dalekoidących wniosków na podstawie ich oszacowania. Następnie została wygenerowana krzywa ROC.

Krzywa ROC



Powyższa krzywa obrazuje zdolność predykcyjną modelu dla różnych progów odcięcia. Wygląd powyższej krzywej pokrywa się z powyższą oceną modeli za pomocą tabel kontyngencji. Oszacowane modele są tak efektywne jak klasyfikator losowy. Z uwagi na to, że modele nie przewidują dla żadnego klienta wartości defaultu równego 1, czułość naszego modelu wynosi 100%, natomiast swoistość 0%. Z tego względu pole pod krzywą ROC wynosi 0.5, co potwierdza poniższy wydruk z R.

```
auc_t <- performance(ROCRpred_t, measure = "auc")
auc_t <- auc_t@y.values[[1]]
auc_t
```

```
## [1] 0.5
```

Bootstrapping

Ostatnim elementem pracy jest wyznaczenie bootstrapowych przedziałów ufności. Do obliczeń przyjęto 90 - procentowy przedział ufności i przeprowadzono 1000 obserwacji.

```
N <- length(data$default) # how long will be our new sample
M <- 1000 # how many times we will generate new sample
auc_vector <- c(1:M) # starting values of AUC vector
alpha <- 0.1

data$predict_boots <- predict(model, type = 'response', newdata = data)
data$predict_boots <- ifelse(data$predict_boots<0.5, 0, 1)

tic()

for (i in 1:M){

  # 1 step - drawing rows from the sample with replacement
  resample <- sample_n(data, N, replace = TRUE)

  # 2 step - compute AUC for resample
  # ROC Curve
  ROCRpred <- prediction(resample$predict_boots, resample$default)
  ROCRperf <- performance(ROCRpred, 'tpr', 'fpr')
  # AUC value
  auc <- performance(ROCRpred, measure = "auc")
  auc <- auc@y.values[[1]]
  auc_vector[i] <- auc_t # here are stored AUC values for each subsample

}

toc()
```

```
## 8.18 sec elapsed
```

```
lower <- quantile(auc_vector, alpha / 2)

upper <- quantile(auc_vector, 1 - (alpha / 2) )

paste0('AUC = ', auc_t, ' a ', (1 - alpha) * 100,
       '% przedział ufności jest pomiędzy ', lower, ' a ', upper )
```

```
## [1] "AUC = 0.5 a 90% przedział ufności jest pomiędzy 0.5 a 0.5"
```

Taki wynik nie daje niestety pełnej satysfakcji z wyznaczenia przedziałów bootstrapowych. Można podejrzewać, że jest to spowodowane losowymi wartościami zmiennych.