Credit Risk

Filip Mordarski & Mateusz Wasielewski

5 12 2020

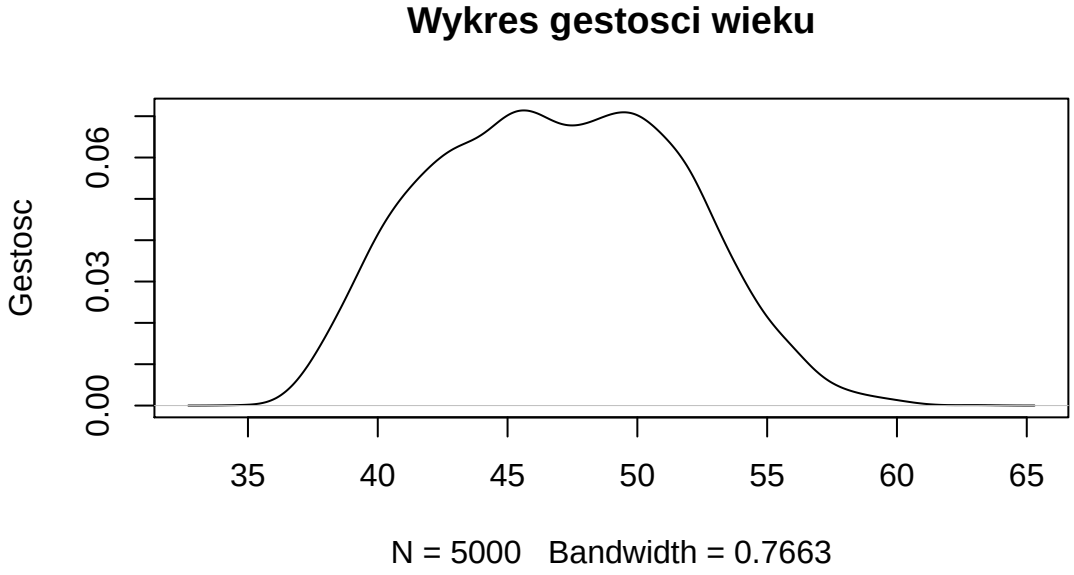
Spis treści

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 modelów, krzywa ROC oraz bootstrapowe przedziały ufności.

Tworzenie zmiennych objaśniających

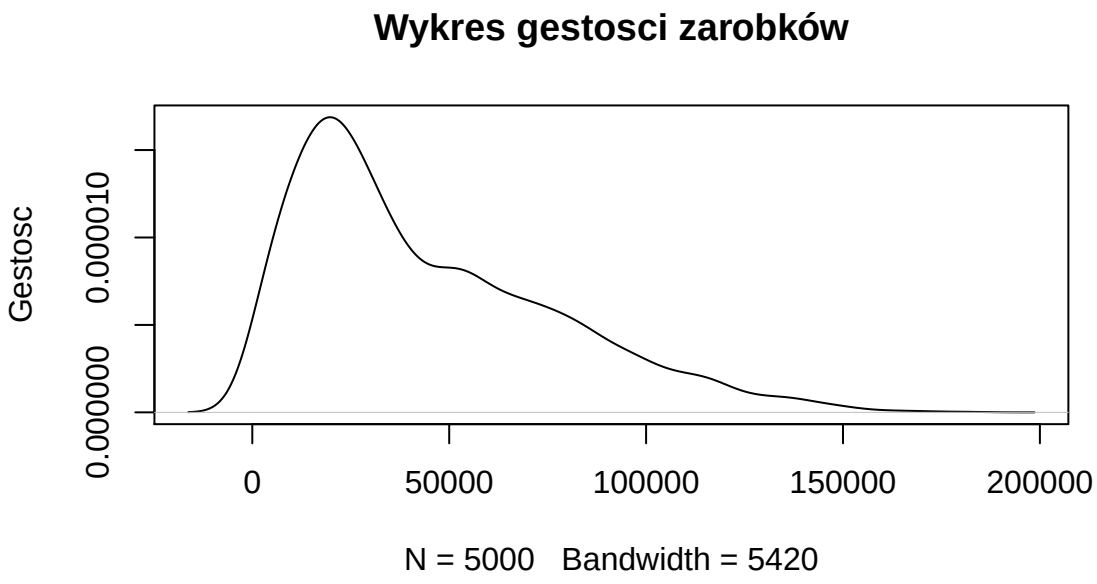
W pierwszej kolejności zostały utworzone zmienne objaśniające potrzebne do modelu PD. Pierwszą zmienną, która została wygenerowana na podstawie utworzonych wcześniej zmiennych jest wiek. Został wyliczony okres trwania umowy w latach na podstawie różnicy między obecną datą a wartością w zmiennej *agreement_start*. 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.

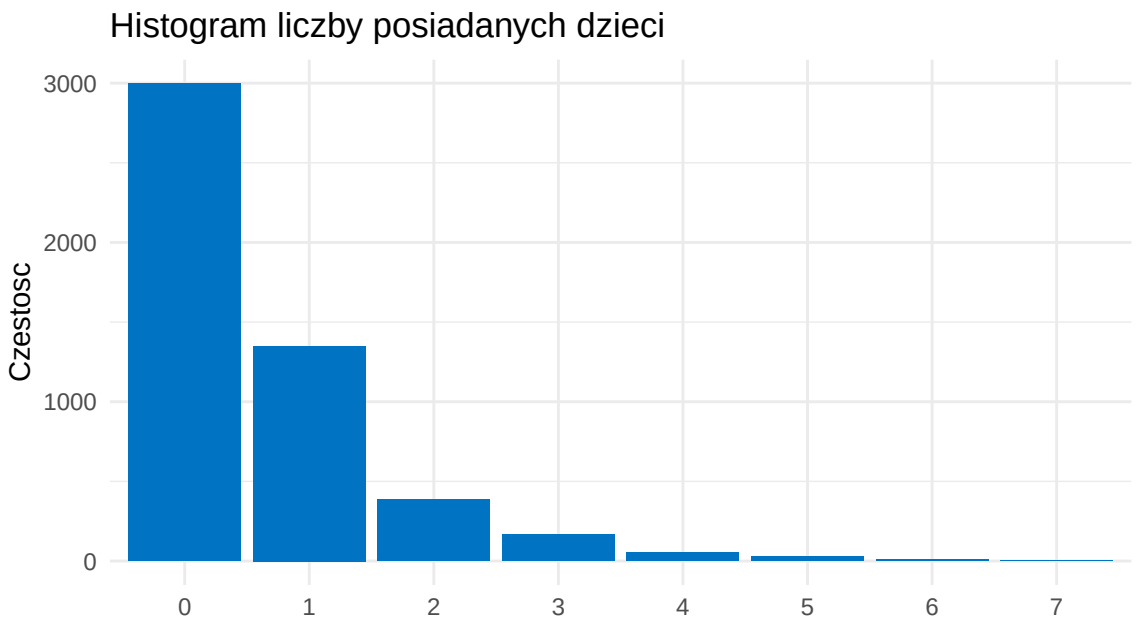
Na podstawie zmiennej, określającej czy dana osoba jest zatrudniona, wygenerowano zmienną czy dana osoba jest zatrudniona na pełny etat. Prawdopodobieństwo tego wynosi 80 %. Poniższy wykres przedstawia histogram tej zmiennej.

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.

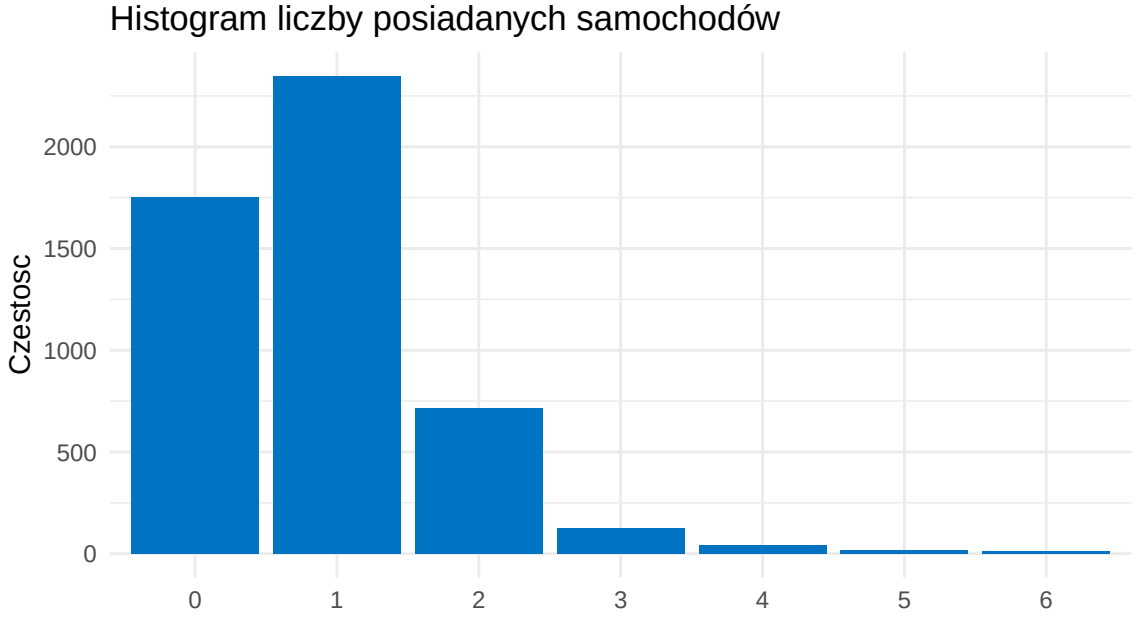


Kolejno, została wygenerowana zmienna, mówiąca o tym czy dana osoba jest singlem, czy żyje w związku z inną osobą. Prawdopodobieństwo, że ktoś jest singlem w Stanach Zjednoczoncyh wynosi 30 %. Przedstawiono histogram tej zmiennej w zbiorze.

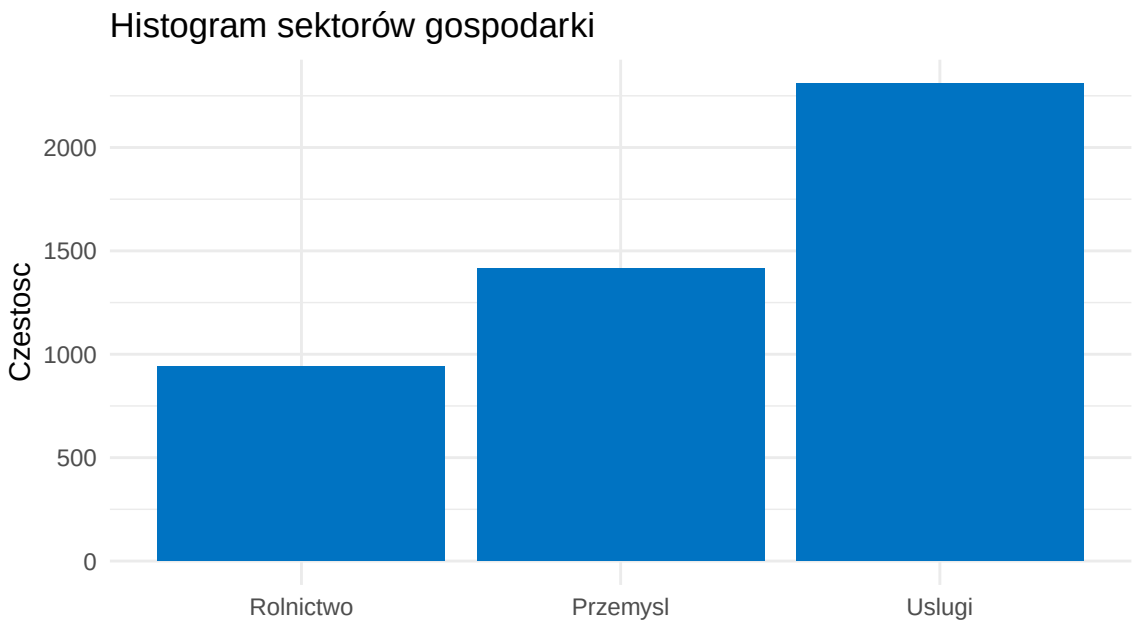
Następnie, została wygenerowana liczba posiadanych dzieci przez daną osobą. Zmienna ta została wygenerowana na podstawie zmiennej, określającej czy dana osoba jest singlem czy nie jest. Poniżej przedstawiono histogram tej wygenerowanej zmiennej.



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: liczba_dzieci - [wartość z losowania liczb [1,2] z 50% prawdopodobieństwami] + 1. Poniżej przedstawiono histogram liczby posiadanych samochodów przez klientów.



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



Badanie rozkładów zmiennych

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

Tablica 1: Zestawienie wartosci kurtozy zmiennych objasniajacych

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 wartosciami

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: Wartosci kurtozy zmiennych po winsoryzacji

value_mortgage_win	value_nonmortgage_win	annual_income_win
3.290116	2.645227	2.36553

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

```
##       X.x                     name                ID          agreement_start
##  Min.   :   1   Bryant Schinner:   2   100-27-1954:   1   1990-01-02:   1
##  1st Qu.:1259   Catalina Nolan :   2   100-41-2606:   1   1990-01-05:   1
##  Median :2458   Haywood Dare   :   2   101-26-9697:   1   1990-01-06:   1
##  Mean   :2486   Kirk Bayer     :   2   101-38-5899:   1   1990-01-07:   1
##  3rd Qu.:3718   Landon Bailey  :   2   102-26-4623:   1   1990-01-08:   1
##  Max.   :4999   Noah Ferry     :   2   102-62-8395:   1   1990-01-09:   1
##                 (Other)        :3514   (Other)    :3520   (Other)   :3520
##                creditcardnr     default        date_of_default value_mortgage
##  1007-6197-9830-4323:   1   Min.   :0.000   2004-01-04:   1   Min.   :      0
##  1007-8232-2779-7267:   1   1st Qu.:0.000   2004-01-08:   1   1st Qu.:      0
##  1010-2603-7123-6102:   1   Median :0.000   2004-01-15:   1   Median :      0
##  1012-8482-6593-3607:   1   Mean   :0.198   2004-01-22:   1   Mean   : 146155
##  1013-5570-9510-2830:   1   3rd Qu.:0.000   2004-01-29:   1   3rd Qu.: 279654
##  1016-3352-6046-9824:   1   Max.   :1.000   (Other)   : 693   Max.   :1830299
##  (Other)            :3520                   NA's      :2828
##  value_nonmortgage      age       employed full_time annual_income    single
##  Min.   :   78     Min.   :35.0   0: 246   0: 897    Min.   :     1   0:1784
##  1st Qu.: 4721     1st Qu.:43.0   1:3280   1:2629    1st Qu.: 18703   1:1742
##  Median : 8177     Median :47.0                      Median : 34929
##  Mean   : 9813     Mean   :46.9                      Mean   : 44614
##  3rd Qu.:13108     3rd Qu.:50.0                      3rd Qu.: 65420
##  Max.   :54748     Max.   :63.0                      Max.   :182379
##
##       kids            car         sector   value_mortgage_win
##  Min.   :0.000   Min.   :0.0000   0: 246   Min.   :     0
##  1st Qu.:0.000   1st Qu.:0.0000   1: 668   1st Qu.:     0
##  Median :0.000   Median :1.0000   2:1001   Median :     0
##  Mean   :0.618   Mean   :0.8976   3:1611   Mean   :137912
##  3rd Qu.:1.000   3rd Qu.:1.0000            3rd Qu.:279654
##  Max.   :7.000   Max.   :6.0000            Max.   :702425
##
##  value_nonmortgage_win annual_income_win
##  Min.   : 1715         Min.   :  5042
##  1st Qu.: 4721         1st Qu.: 18703
##  Median : 8177         Median : 34929
##  Mean   : 9576         Mean   : 43868
##  3rd Qu.:13108         3rd Qu.: 65420
##  Max.   :23398         Max.   :109824
##
```

```
##       X.x                       name                ID          agreement_start
##  Min.   :   2   Oliver Wintheiser:   2   100-10-9506:   1   1990-01-03:   1
##  1st Qu.:1230   Scottie D'Amore  :   2   100-48-4032:   1   1990-01-04:   1
##  Median :2594   Aaron Swift      :   1   101-31-7035:   1   1990-01-10:   1
```

```
##  Mean   :2534   Abdul Krajcik    :   1   102-42-3662:   1   1990-01-11:   1
##  3rd Qu.:3808   Abel Stark       :   1   103-21-3473:   1   1990-01-17:   1
##  Max.   :5000   Abraham Nikolaus :   1   103-51-9594:   1   1990-01-18:   1
##                 (Other)          :1466   (Other)    :1468   (Other)   :1468
##               creditcardnr     default         date_of_default
##  1007-6595-4169-2500:   1   Min.   :0.0000   2004-02-11:   1
##  1019-3206-5994-1992:   1   1st Qu.:0.0000   2004-03-09:   1
##  1024-5119-7938-6705:   1   Median :0.0000   2004-04-08:   1
##  1032-9219-6414-9085:   1   Mean   :0.2157   2004-05-12:   1
##  1050-7668-9868-7825:   1   3rd Qu.:0.0000   2004-05-14:   1
##  1063-9566-2338-8440:   1   Max.   :1.0000   (Other)   : 313
##  (Other)            :1468                    NA's      :1156
##  value_mortgage    value_nonmortgage      age        employed full_time
##  Min.   :      0   Min.   :  104     Min.   :37.00   0:  88   0: 347
##  1st Qu.:      0   1st Qu.: 4690     1st Qu.:43.00   1:1386   1:1127
##  Median :      0   Median : 8121     Median :47.00
##  Mean   : 147757   Mean   : 9829     Mean   :46.95
##  3rd Qu.: 276476   3rd Qu.:13348     3rd Qu.:50.00
##  Max.   :1455568   Max.   :57542     Max.   :60.00
##
##  annual_income    single       kids             car         sector
##  Min.   :    12   0:717   Min.   :0.0000   Min.   :0.0000   0: 88
##  1st Qu.: 18526   1:757   1st Qu.:0.0000   1st Qu.:0.0000   1:273
##  Median : 34843           Median :0.0000   Median :1.0000   2:415
##  Mean   : 44133           Mean   :0.6065   Mean   :0.8657   3:698
##  3rd Qu.: 64755           3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :178529           Max.   :6.0000   Max.   :6.0000
##
##  value_mortgage_win value_nonmortgage_win annual_income_win
##  Min.   :     0     Min.   : 1715         Min.   :  5042
##  1st Qu.:     0     1st Qu.: 4690         1st Qu.: 18526
##  Median :     0     Median : 8121         Median : 34843
##  Mean   :138868     Mean   : 9601         Mean   : 43410
##  3rd Qu.:276476     3rd Qu.:13348         3rd Qu.: 64755
##  Max.   :702425     Max.   :23398         Max.   :109824
##
```

```
##  [1] "X.x"                   "name"                  "ID"
##  [4] "agreement_start"       "creditcardnr"          "default"
##  [7] "date_of_default"       "value_mortgage"        "value_nonmortgage"
## [10] "age"                   "employed"              "full_time"
## [13] "annual_income"         "single"                "kids"
## [16] "car"                   "sector"                "value_mortgage_win"
## [19] "value_nonmortgage_win" "annual_income_win"
```

```
##
## 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
```

Oszacowano również drugi model, w którym zmieniliśmy założenie co do rozkładu składnika losowego. 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
```

Ewaluacja modelu

Po oszacowaniu obu modelów, 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
##
```

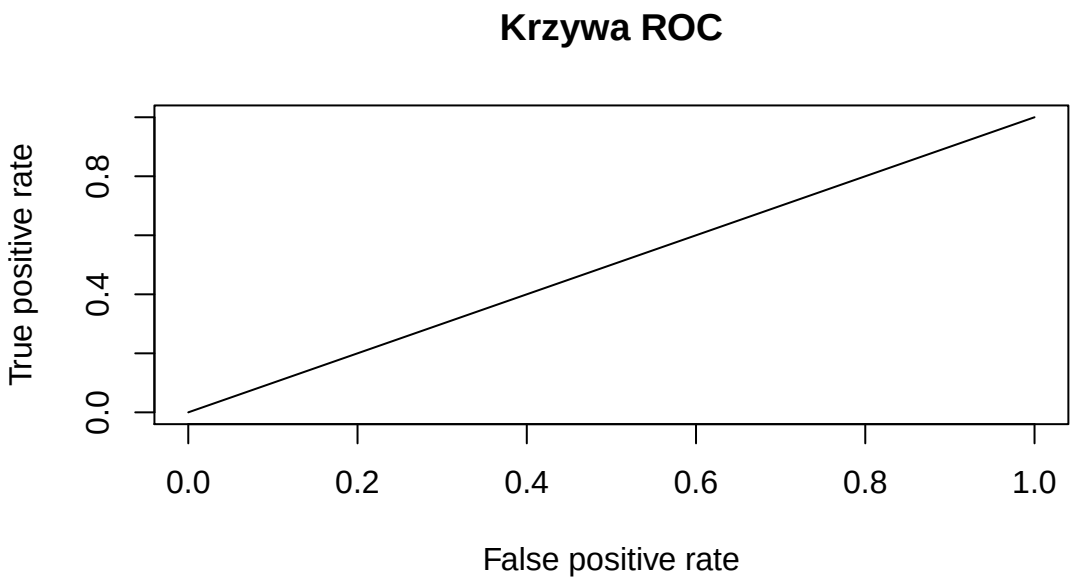
Następnie wygenerowano podobną tabelę kontyngencji dla drugiego modelu logitowego.

```
## 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
##
```

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.



Powyższa krzywa obrazuje zdolność predykcyjną modelu dla różnych progów odcięcia. Wygląd powyższej krzywej pokrywa się z powyższą oceną modelów 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
```