Credit Risk

Filip Mordarski & Mateusz Wasielewski

5 12 2020

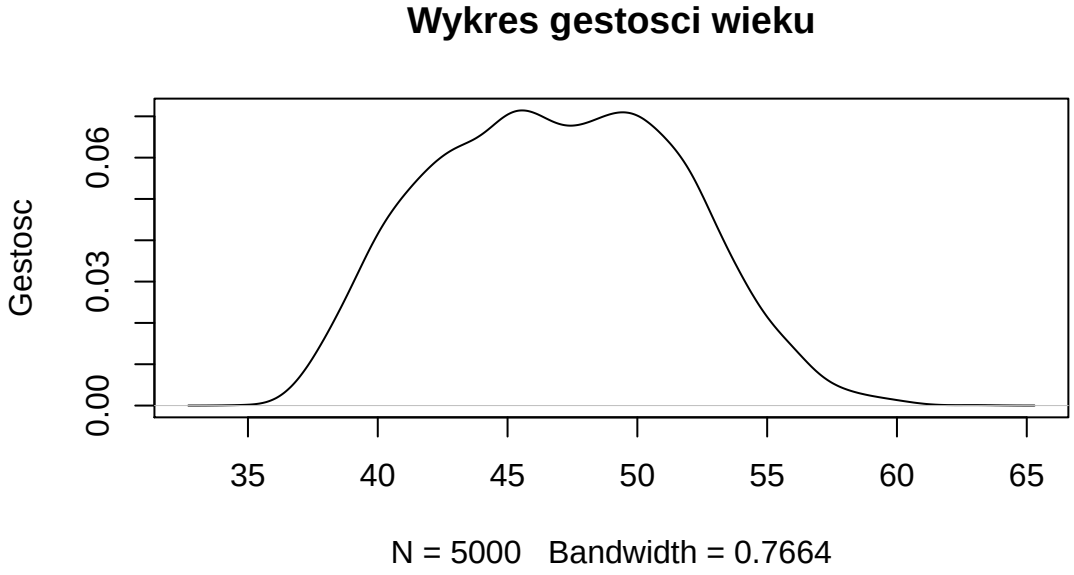
Spis treści

Wstęp

Poniższy raport będzie zawierał analizę modelu ryzyka kredytowego.

Tworzenie zmiennych objaśniających

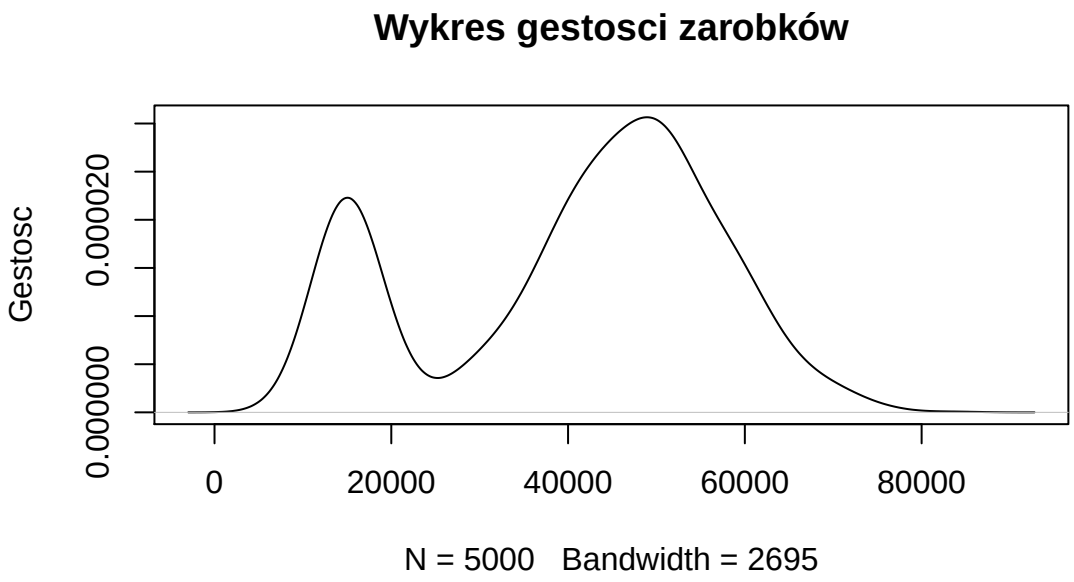
W pierwszej kolejności zostały utworzone zmienne objaśniające potrzebne do modelu PD. Pierwszą zmienną, która została wygenerowana na podstawie utworzonych wcześniej zmiennych jest wiek. Został wyliczony okres trwania umowy w latach na podstawie różnicy między obecną datą a wartością w zmiennej *agreement_start*. 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 3.6 %. 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.

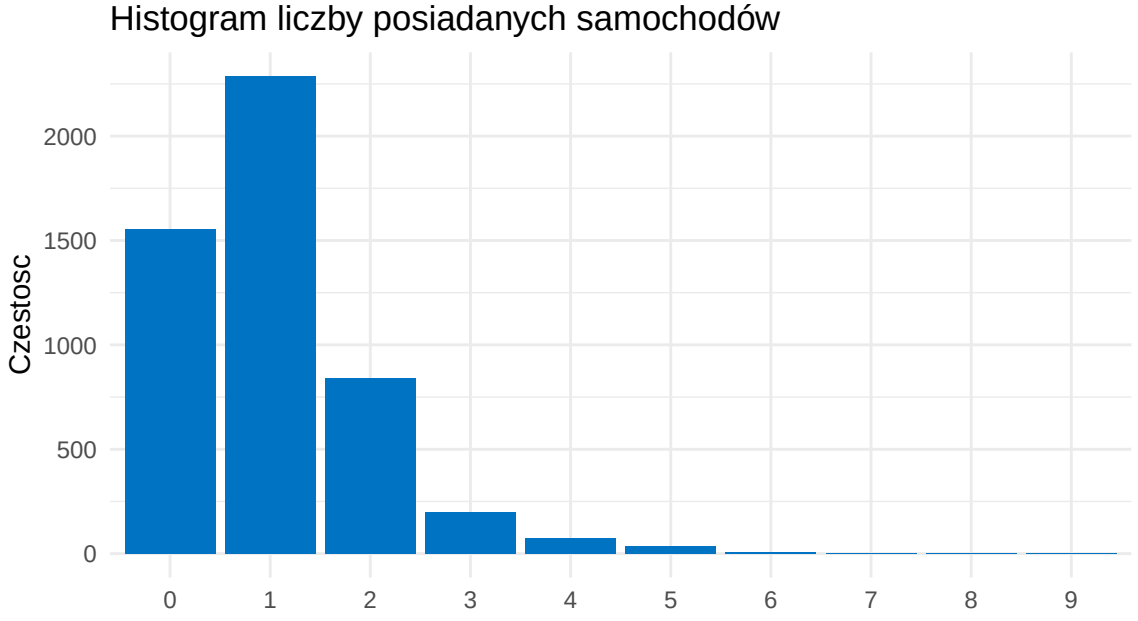
Zmienną, która z pewnością może okazać się istotna w tworzeniu modelu PD jest dochód roczny danej osoby. Wartości te zostały wylosowane z rozkładu normalnego. Średnia dla osób zatrudnionych na pełen etat została ustalona na poziomie 48000 USD z odchyleniem standardowym na poziomie 10000 USD. Dla osób niezatrudnionych na pełen etat wartość średnia została ustalona na poziomie 15000 USD, natomiast odchylenie 3000 USD. Poniżej zaprezentowano wykres gęstości tej zmiennej.



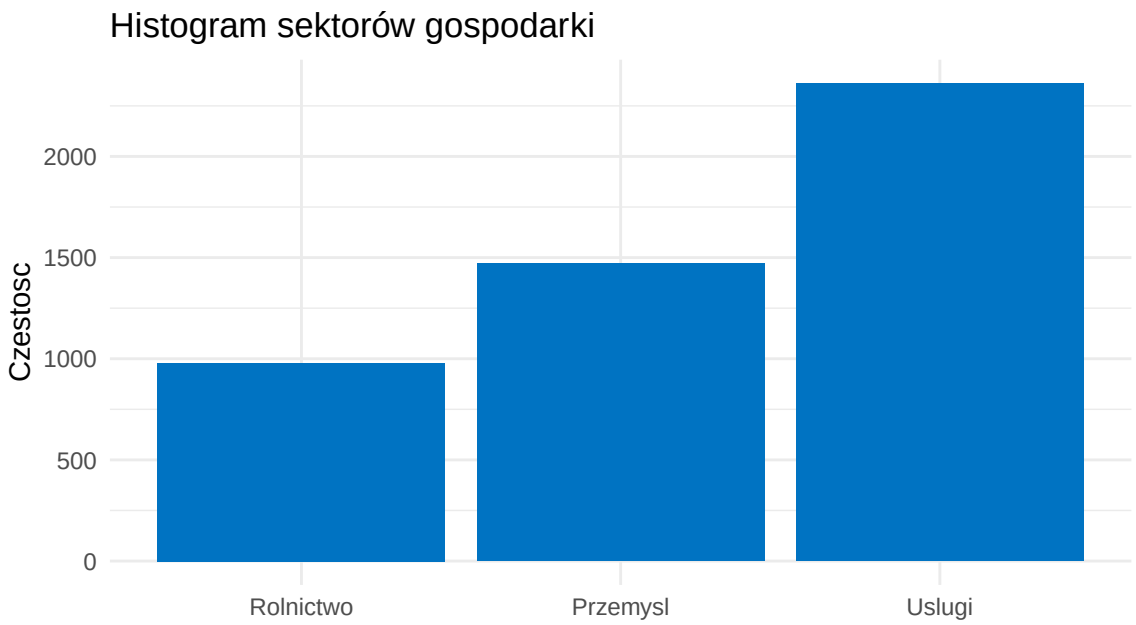
Kolejno, została wygenerowana zmienna, mówiąca o tym czy dana osoba jest singlem, czy żyje w związku z inną osobą. Prawdopodobieństwo, że ktoś jest singlem w Stanach Zjednoczoncyh wynosi 50.2 %. Przedstawiono histogram tej zmiennej w zbiorze.

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 wartości kurtozy zmiennych objaśniających

value_mortgage	value_nonmortgage	age	annual_income
5.601227	5.420217	2.310862	2.141716

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. Istnieją różne metody radzenia sobie z tzw. “outliersami”, jedna z nich zostanie przedstawiona w następnym akapicie. Dla zmiennych wiek oraz roczny przychód wartość kurtozy jest mniejsza od 3, co pozwala zakładać normalność rozkładów tych zmiennych.

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 czterech zmiennych ciągłych, których wspomniana miara spłaszczenia była większa od 3.

Tablica 2: Wartości kurtozy zmiennych po winsoryzacji

value_mortgage_win	value_nonmortgage_win
3.290116	2.645227

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

```
##       X.x                     name                ID
##  Min.   :   1   Brent Hane      :   2   100-10-9506:   1
##  1st Qu.:1290   Clarice Jacobson:   2   100-27-1954:   1
##  Median :2560   Haywood Dare    :   2   100-41-2606:   1
##  Mean   :2545   Noah Ferry      :   2   100-48-4032:   1
##  3rd Qu.:3811   Owen Keebler    :   2   101-26-9697:   1
##  Max.   :5000   Abbey Gleichner :   1   101-31-7035:   1
##                 (Other)         :3493   (Other)    :3498
##    agreement_start             creditcardnr     default
##  1990-01-02:   1   1007-6197-9830-4323:   1   Min.   :0.0000
##  1990-01-03:   1   1007-6595-4169-2500:   1   1st Qu.:0.0000
##  1990-01-04:   1   1010-2603-7123-6102:   1   Median :0.0000
##  1990-01-05:   1   1012-8482-6593-3607:   1   Mean   :0.2029
##  1990-01-08:   1   1013-5570-9510-2830:   1   3rd Qu.:0.0000
##  1990-01-09:   1   1016-3352-6046-9824:   1   Max.   :1.0000
##  (Other)   :3498   (Other)            :3498
##    date_of_default value_mortgage    value_nonmortgage      age
##  2004-01-04:   1   Min.   :      0   Min.   :   78     Min.   :35.00
##  2004-01-08:   1   1st Qu.:      0   1st Qu.: 4741     1st Qu.:43.00
##  2004-01-15:   1   Median :      0   Median : 8258     Median :47.00
##  2004-01-22:   1   Mean   : 151831   Mean   : 9832     Mean   :46.92
##  2004-02-03:   1   3rd Qu.: 291464   3rd Qu.:13277     3rd Qu.:50.00
##  (Other)   : 706   Max.   :1830299   Max.   :43596     Max.   :63.00
##  NA's      :2793
##     employed        full_time      annual_income       single
##  Min.   :0.0000   Min.   :0.0000   Min.   : 6106   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:30229   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :44520   Median :1.0000
##  Mean   :0.9623   Mean   :0.7803   Mean   :40726   Mean   :0.5026
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:52509   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :77699   Max.   :1.0000
##
##       kids             car            sector      value_mortgage_win
##  Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :     0
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:     0
##  Median :1.0000   Median :1.000   Median :2.000   Median :     0
##  Mean   :0.9013   Mean   :1.023   Mean   :2.278   Mean   :142806
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:291464
##  Max.   :9.0000   Max.   :9.000   Max.   :3.000   Max.   :702425
##                                   NA's   :132
##  value_nonmortgage_win
##  Min.   : 1715
##  1st Qu.: 4741
##  Median : 8258
##  Mean   : 9622
##  3rd Qu.:13277
```

```
##  Max.   :23398
##
```

```
##       X.x                     name                ID
##  Min.   :   3   Bryant Schinner:   2   102-42-3662:   1
##  1st Qu.:1170   Catalina Nolan :   2   102-62-8395:   1
##  Median :2360   Ian Purdy      :   2   102-81-2600:   1
##  Mean   :2397   Scottie D'Amore:   2   103-47-2162:   1
##  3rd Qu.:3600   Aaron Stokes   :   1   104-13-2239:   1
##  Max.   :4996   Aaron Swift    :   1   105-54-9220:   1
##                 (Other)        :1486   (Other)    :1490
##    agreement_start              creditcardnr     default
##  1990-01-06:   1   1007-8232-2779-7267:   1   Min.   :0.0000
##  1990-01-07:   1   1021-9495-5857-5755:   1   1st Qu.:0.0000
##  1990-01-12:   1   1031-9098-4136-8313:   1   Median :0.0000
##  1990-01-15:   1   1032-9219-6414-9085:   1   Mean   :0.2039
##  1990-01-16:   1   1034-2090-9030-2458:   1   3rd Qu.:0.0000
##  1990-01-19:   1   1035-4768-5162-5799:   1   Max.   :1.0000
##  (Other)   :1490   (Other)            :1490
##    date_of_default value_mortgage    value_nonmortgage      age
##  2004-01-29:   1   Min.   :      0   Min.   :  365     Min.   :37.00
##  2004-02-11:   1   1st Qu.:      0   1st Qu.: 4616     1st Qu.:43.00
##  2004-03-03:   1   Median :      0   Median : 7992     Median :47.00
##  2004-03-09:   1   Mean   : 134440   Mean   : 9782     Mean   :46.91
##  2004-03-18:   1   3rd Qu.: 224365   3rd Qu.:13042     3rd Qu.:50.00
##  (Other)   : 300   Max.   :1376702   Max.   :57542     Max.   :60.00
##  NA's      :1191
##     employed        full_time      annual_income       single
##  Min.   :0.0000   Min.   :0.0000   Min.   : 5118   Min.   :0.000
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:26266   1st Qu.:0.000
##  Median :1.0000   Median :1.0000   Median :44080   Median :0.000
##  Mean   :0.9619   Mean   :0.7574   Mean   :40267   Mean   :0.488
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:52194   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :84687   Max.   :1.000
##
##       kids             car            sector      value_mortgage_win
##  Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :     0
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:     0
##  Median :1.0000   Median :1.000   Median :3.000   Median :     0
##  Mean   :0.8817   Mean   :1.014   Mean   :2.308   Mean   :127393
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:3.000   3rd Qu.:224365
##  Max.   :7.0000   Max.   :7.000   Max.   :3.000   Max.   :702425
##                                   NA's   :57
##  value_nonmortgage_win
##  Min.   : 1715
##  1st Qu.: 4616
##  Median : 7992
##  Mean   : 9493
##  3rd Qu.:13042
##  Max.   :23398
##
```

```
##  [1] "X.x"                   "name"
##  [3] "ID"                    "agreement_start"
```

```
##  [5] "creditcardnr"          "default"
##  [7] "date_of_default"       "value_mortgage"
##  [9] "value_nonmortgage"     "age"
## [11] "employed"              "full_time"
## [13] "annual_income"         "single"
## [15] "kids"                  "car"
## [17] "sector"                "value_mortgage_win"
## [19] "value_nonmortgage_win"
```

```
##
## Call:
## glm(formula = default ~ value_mortgage_win + value_nonmortgage_win +
##     age + employed + full_time + annual_income + single + kids +
##     car, family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8162  -0.6919  -0.6552  -0.6106   1.9624
##
## Coefficients:
##                            Estimate    Std. Error z value   Pr(>|z|)
## (Intercept)           -2.5623015201  0.4909968129  -5.219 0.00000018 ***
## value_mortgage_win     0.0000002001  0.0000001765   1.133     0.2570
## value_nonmortgage_win  0.0000063111  0.0000068238   0.925     0.3550
## age                    0.0227001290  0.0089992649   2.522     0.0117 *
## employed              -0.0705851262  0.2417107566  -0.292     0.7703
## full_time              0.1130257968  0.1904375557   0.594     0.5528
## annual_income          0.0000002856  0.0000046686   0.061     0.9512
## single                -0.0114309582  0.0842098227  -0.136     0.8920
## kids                  -0.0499135760  0.0526508584  -0.948     0.3431
## car                    0.0518883933  0.0612924009   0.847     0.3972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3534.9  on 3503  degrees of freedom
## Residual deviance: 3524.3  on 3494  degrees of freedom
## AIC: 3544.3
##
## Number of Fisher Scoring iterations: 4
```

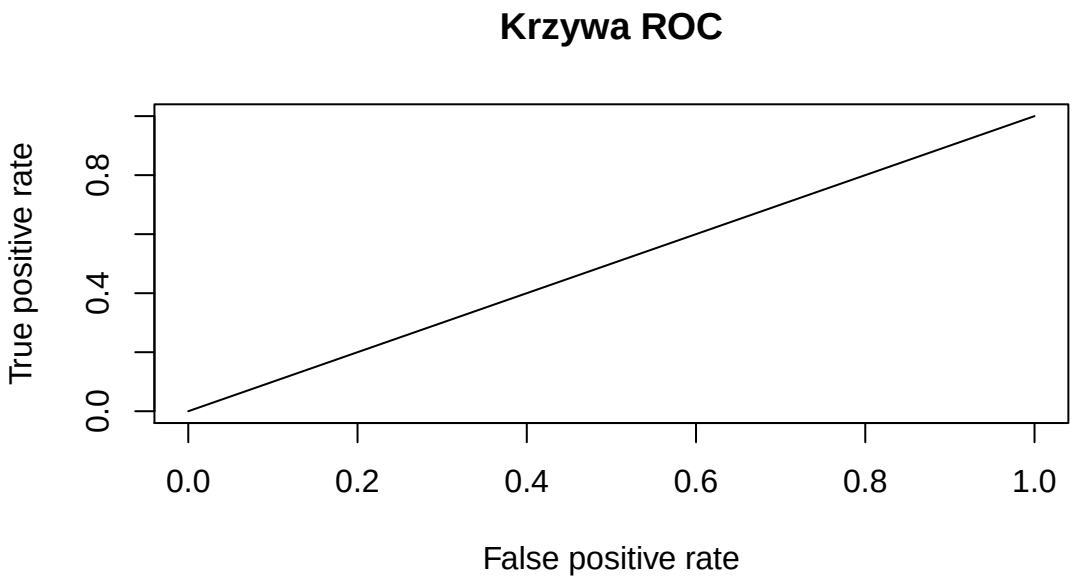
Ewaluacja modelu

Po oszacowaniu modelo, dokonano jego ewaluacji. Poniżej przedstawiono tabelę kontyngencji dla oszacowanego powyżej modelu.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1191  305
```

```
##          1    0    0
##
##                Accuracy : 0.7961
##                  95% CI : (0.7748, 0.8163)
##     No Information Rate : 0.7961
##     P-Value [Acc > NIR] : 0.5153
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.7961
##          Neg Pred Value :    NaN
##              Prevalence : 0.7961
##          Detection Rate : 0.7961
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

Na podstawie powyższej tabeli możemy stwierdzić, że nasz model nie jest efektywny w predykcji 'defaultu'. Wszystkie obserwacje zostały dopasowane do kategorii 0, czyli brak 'defaultu'. Mamy relatywnie wysoki poziom trafności modelu, ponieważ jest on na poziomie 79 %. Przyczyną takiego stanu rzeczy, jest to że 78% obserwacji ze zbioru testowego nie miało 'defaultu'. Ewaluacja tego modelu wykazała, że nie powinniśmy wyciągać żadnych dalekoidących wniosków na podstawie oszacowań tego modelu. Następnie została wygenerowana krzywa ROC dla oszacowanego modelu.



Powyższa krzywa obrazuje zdolność predykcyjną modelu dla różnych progów odcięcia. Wygląd powyższej krzywej pokrywa się z powyższą oceną modelu za pomocą tabeli kontyngencji. Oszacowany model jest tak

efektywny jak klasyfikator losowy. Z uwagi na to, że model nie przewiduje 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
```