Zastosowanie modelowania matematycznego w bankowości - lista 3

Łukasz Rębisz

2023-05-27

Zadanie 1

Oznaczmy portfel niepracujący zgodnie z definicją niewykonania zobowiązania (default):

Produkt zostaje przeklasyfikowany do portfela niepracującego, jeżeli spełniony jest co najmniej jeden z poniższych warunków:

- rachunek znajduje się w windykacji,
- liczba dni z przeterminowaniem > 400 PLN i $> 1\%$ wartości ekspozycji przekracza 90 dni (DMA),
- nastąpiła śmierć wszystkich kredytobiorców.

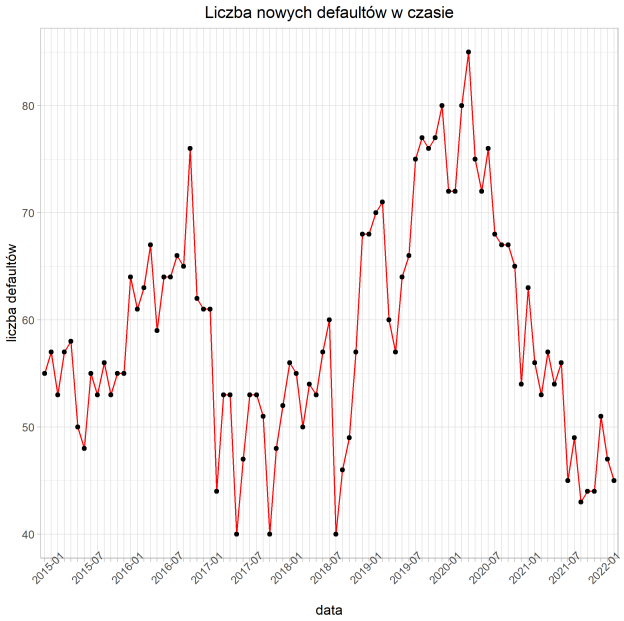
Rachunek oznaczany jest jako *cure* (powrót do porfela pracującego), jeśli przestał spełniać definicję *defaulta* i przez 92 kolejne dni posiadał prawidłową obsługę, tj. nie posiadał więcej niż dni przeterminowania. (DMA).

Zadanie 2

a) Liczba defaultów

Spośród 4920 kredytów zawartych w badanych danych w statusie default znalazło się 409 produktów.

Zbadajmy rozkład pojawiania się nowych defaultów w czasie.



Powyższy wykres wskazuje na dwa okresy, w których pojawiało się więcej nowych *defaultów*:

- rok 2016 ,
- lata 2019-2020.

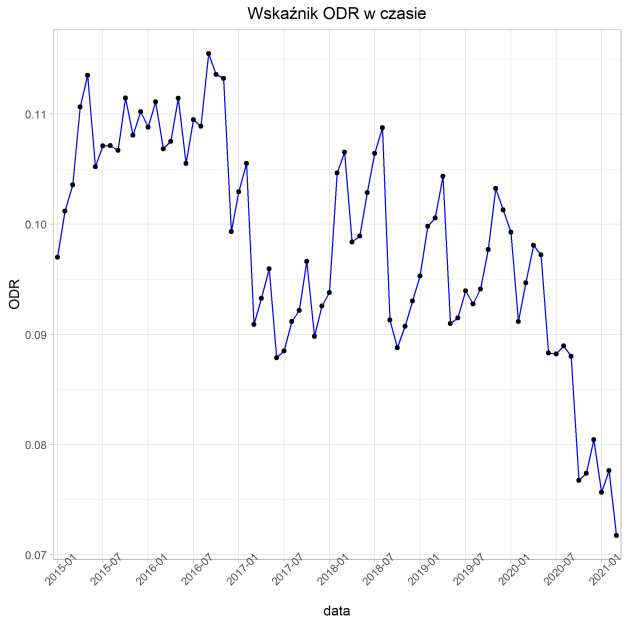
Pomiędzy powyższymi okresami następowały spadki liczby nowych defaultów.

b) ODR

Oznaczmy obserwowany odsetek niewykonania zobowiązania

$$ODR = \frac{\text{liczba ekspozycji, które zdefaultują w ciągu roku od momentu } t}{\text{liczba wszystkich ekspozycji w portfelu pracującym w okresie } t},$$

gdzie momenty t to poszczególne miesiące.



Powyższy wykres wskazuje na stosunkowo wysoki wskaźnik ODR w okresach:

- lata 2015-2016,
- okres styczeń 2018 - kwiecień 2020.

Pomiędzy okresami (tzn. w 2017 roku) oraz od maja 2020 roku nastąpił spadek wskaźnika ODR.

Na powyższym wykresie ograniczyliśmy zakres na osi OX tak, aby dany kredyt był notowany przez kolejny rok (dane obejmują okres do marca 2022).

Zadanie 3

Dla otrzymanych defaultów wyznaczmy obserwowaną stratę LGD:

$$LGD_m = 1 - \frac{\sum_{i=1}^{m} \left(\frac{CF_i}{(1+d/12)^i} \right)}{EAD},$$

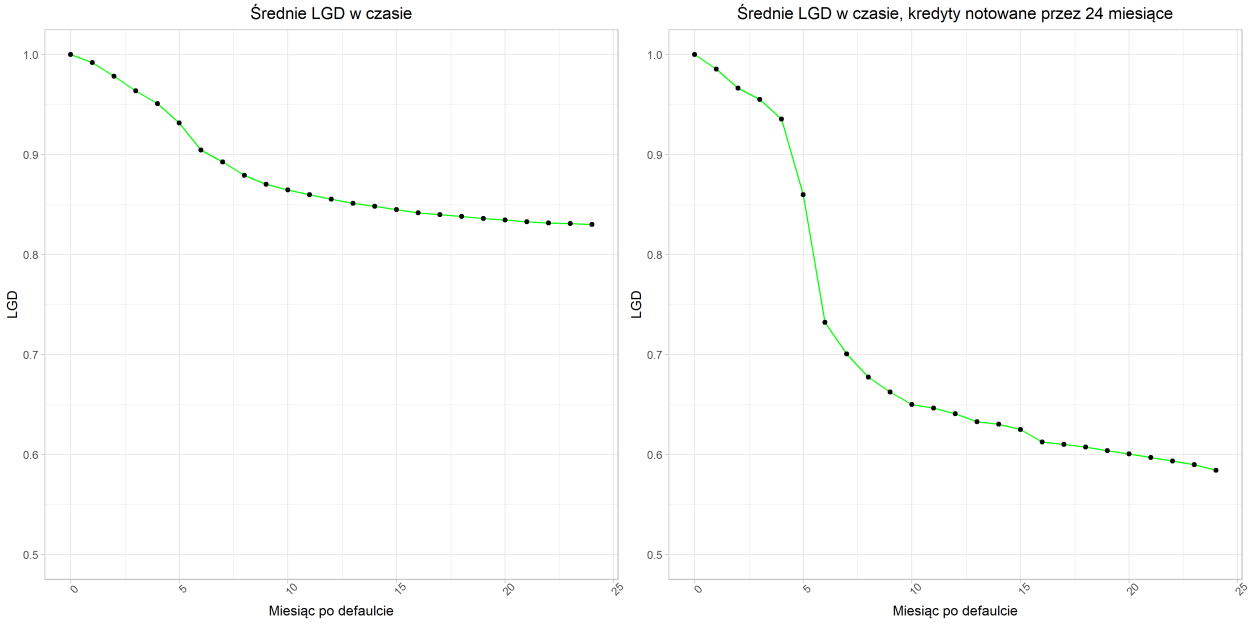
gdzie:

- $d = 9\%$ - stopa dyskonta,
- CF_i - przepływ w i-tym miesiącu po defaulcie,
- EAD - ekspozycja na moment defaulta.

CF to różnica sald w kolejnych miesiącach pomniejszona o ewentualne umorzenia.

W przypadku powrotu do portfela pracującego uznajemy, że CF jest równe zaangażowaniu (kwocie pozostałej do spłaty).

Poniższe wykresy obrazują wielkość straty LGD w poszczególnych miesiącach.



- Wykres po prawej stronie obejmuje kredyty, które były notowane przez pełen okres 24 miesięcy po defaulcie. Wartości LGD dla takich kredytów są mniejsze, ponieważ dla kredytów będących na defaulcie przez okres mniejszy niż 24 miesiące zakładaliśmy brak dalszej spłaty w nienotowanym okresie.
- Analiza wartości LGD pokazuje, że wskaźnik ten spada wraz z upływem czasu (zjawisko naturalne spowodowane częściową spłatą kredytów, które znalazły się w statusie deafult). Ponadto, obserwujemy gwałtowny spadek wartości LGD pomiedzy piątym a ósmym miesiącem po defaulcie. Oznacza to, że w tym okresie najwięcej kredytów zaczyna sie ponownie spłacać. Po 15-20 miesiącach wartość LGD stabilizuje się na poziomie ok. 60%.

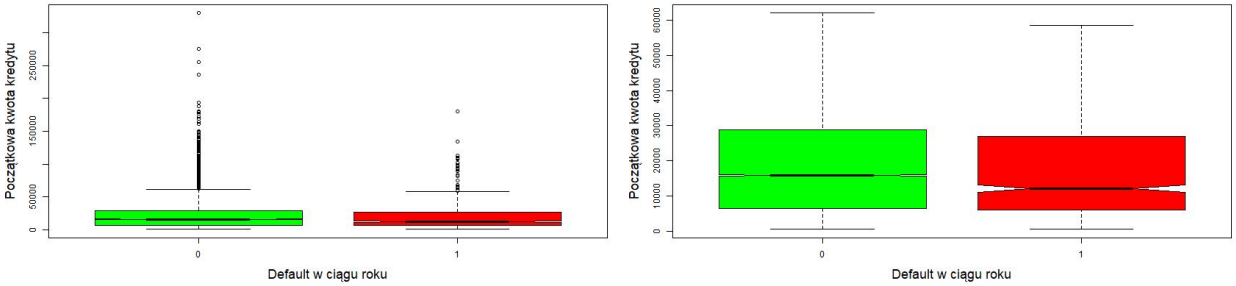
Zadanie 4

Celem zadania jest **zamodelowanie PD** na podstawie dostępnych danych. Dane dla kredytów sprawdzamy co kwartał w okresie styczeń 2015 - marzec 2021, ponieważ będziemy badać prawdodopodobieństwo przejścia kredyty w stan default w ciągu roku (dysponujemy zaś danymi do marca 2022). Dla każdego z kredytów w danym miesiącu wyznaczamy następujące dane:

- czy kredyt znalazł się w stanie default w ciągu kolejnego roku,
- początkowa kwota kredytu,
- zaangażowanie w stosunku do początkowej kwoty kredytu w momencie: na miesiąc przed defaultem lub po roku dla kredytów, które nie znalazły się w stanie default,
- liczba lat do końca kredytu na miesiąc przed defaultem lub po roku dla kredytów, które nie znalazły się w stanie default,
- średnie miesięczne opoźnienie przed znalezieniem się w stanie default.

Analiza danych

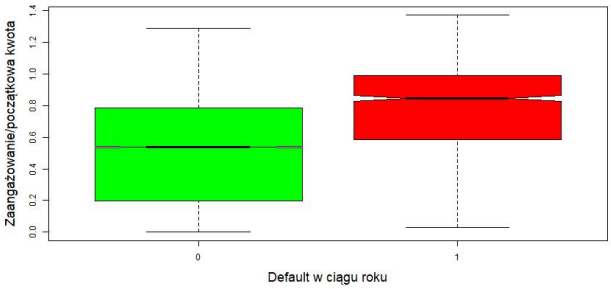
Zbadajmy, jak przejście w stan default zależy od każdej z wyznaczonych danych osobno. W tym celu przeanalizujmy wykresy pudełkowe zmiennej *default* w zależności od danej zmiennej objaśniającej.

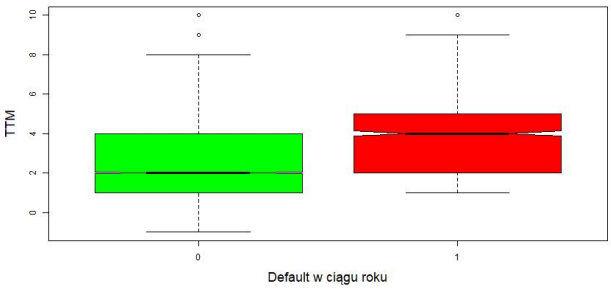


Analiza powyższych histogramów dla zmiennej *początkowa kwota kredytu* pokazuje, że:

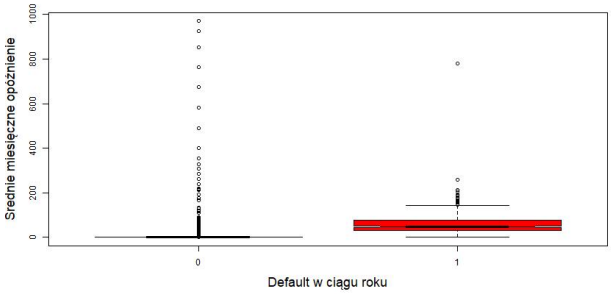
- NIE ma wyraźnej różnicy w rozkładzie tej zmiennej dla kredytów z oraz bez defaulta.
- Kredyty, które nie znalazły się w stanie default, charakteryzują się większą liczbą dużych obserwacji odstających.
- Wykres pudełkowy bez obserwacji odstających (wykres po prawej stronie) pokazuje, że mediany dla obu grup (wcięcie wykresu) znajdują się na podobnym poziomie. Wykres dla kredytów, które nie znalazły się w stanie default jest nieznacznie przesunięty w stronę większych *początkowych wartości kredytu*.

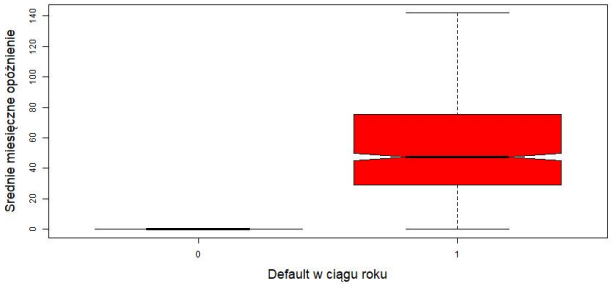
Wniosek: *początkowa kwota kredytu* nie ma wyraźnego wpływu na przejście kredytu w stan default. Niemiej jednak rozkład tej zmiennej dla obu grup nie jest identyczny. Z tego powdou nie usuwamy tej zmiennej z modelu.





- Wykres pudełkowy *zaangażowania sprzed defaulta w stosunku do początkowej kwoty kredytu* (wykres po lewej stronie) pokazuje, że kredyty, które znalazły się w stanie deafult charakteryzują się znaczącą większą wartością tej zmiennej niż kredyty, które nie były w ciągu kolejnego roku w stanie default. Świadczy o tym porównanie median dla obu grup (wcięcia wykresów). Ponadto, większość kredytów, które znalazły się w stanie default, posiadało w momencie zaprzestania spłaty stosunkowo wysokie zaangażownie (na poziomie 60-100 %). Wniosek: zmienna ma istotny wpływ na zmienną objaśnianą.
- Wykres liczby lat pozostałych do końca kredytu *TTM* w momencie defaulta pokazuje, że kredyty, które znalazły się w stanie deafult posiadają średnio większą liczbę lat do końca kredytów (porównanie median). Wniosek: zmienne *TTM* ma istotny wpływ na zmienną *default*.



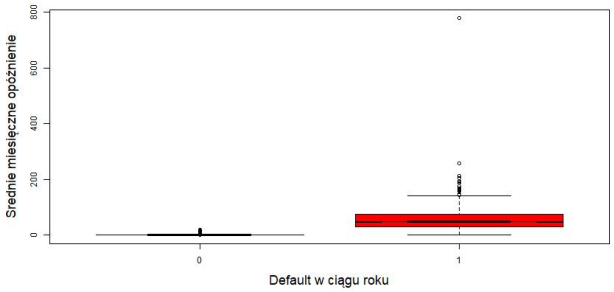


Powyższe wykresy pudełkowe dla zmiennej *średnie miesięczne opóźnienie* pokazują, że:

- Kredyty, które nie znalazły się w stanie default charakteryzują się dużą liczbą obserwacji odstających (wykres po lewej stronie). Wartości odstające są tak duże, że kredyty te znalazłyby się w stanie default. Wynikają one najprawdopodobniej z błędu danych, np. naliczania opóźnień dla zapłaconych rat lub w pełni spłaconych kredytów. Z tego powodu dane mogą zaburzyć model.
- Wykres po prawej stronie (bez obserwacji odstających) wyraźnie pokazuje, że kredyty, które nie przeszły w stan default charakteryzują się *średnim miesięcznym opóźnieniem* bliskim zero w przeciwieństwie do kredytów, które przeszły w stan default (mediana dla tej grupy wynosi ok. 50 *dni opóźnienia*).

Wniosek: *średnie miesięczne opóźnienie* ma istotny wpływ na zmienną objaśnianą *default*. Obserwacje odstające znaczącą odbiegają od wartości w grupie kredytów bez defaulta. W celu uniknięcia błędów modelu wynikających z błędu danych zmieńmy wartości obserwacji odstających w tej grupie na wartość mediany dla tej grupy (równą zero).

Po zmianie wartości odstających wykres pudełkowy dla tej zmiennej przedstawia się następująco:



Porównanie wykresów dla obu grup pozwala wysunąć analogiczne wnioski do otrzymanych w poprzedniej analiznie tej zmiennej.

Model regresji logistycznej

Skonstruujmy **model regresji logistycznej** dla powyższych danych. Zakładana w modelu zależność pomiędzy zmiennymi:

$$\mu = \frac{1}{1 + exp(-\eta)},$$

gdzie μ oznacza średnie prawdodobieństwo osiągnięcia statusu *default* w ciągu roku, natomiast η to kombinacja liniowa zmiennych objaśniających.

Otrzymujemy następujące wyniki dla poszczególnych zmiennych:

zmienna	parametr	p-wartość
Początkowa kwota kredytu	-0.0000389	<0.001
Zaangażowanie/początkowa kwota	0.3555	0.101
TTM	-2.0924	<0.001
Śr. mies. opóźnienie	0.336	<0.001

Analiza p-wartości dla poszczególnych zmiennych pokazuje, że jedynie zmienna *Zaangażowanie/początkowa kwota* NIE ma istotnego wpływu na zmienną objaśnianą *default*. Sprawdźmy zatem, czy model skonstruowany bez tej zmiennej nie jest istotnie gorszy od pełnego modelu.

Porównanie modelu logistycznego z oraz bez zmiennej *Zaangażowanie/początkowa kwota* metodą *ANOVA* (test Deviance):

otrzymana p-wartość dla testu Deviance wynosi 0.1041, zatem nie mamy podstaw do odrzucenia hipotezy mówiącej, że model BEZ zmiennej *Zaangażowanie/początkowa kwota* jest lepszy od pełnego modelu. Zatem zmienna ta NIE ma istotnego wpływu na predykcję *defaulta*.

Porównajmy następnie model skonstruowany bez odrzuconej zmiennej z analogicznym modelem, w którym dodatkowo badamy również interakcje pomiędzy zmiennymi.

Porównanie modelu logistycznego bez zmiennej *Zaangażowanie/początkowa kwota* w wersji z oraz bez interakcji pomiędzy zmiennymi metodą *ANOVA* (test Deviance):

otrzymana p-wartość dla testu Deviance jest mniejsza niż 0.001, zatem odrzucamy hipotezę mówiącą, że model bez interakcji jest lepszy od modelu z interakcjami. Interakcje pomiędzy zmiennymi mają zatem istotny wpływ na zmienną objaśnianą *default*.

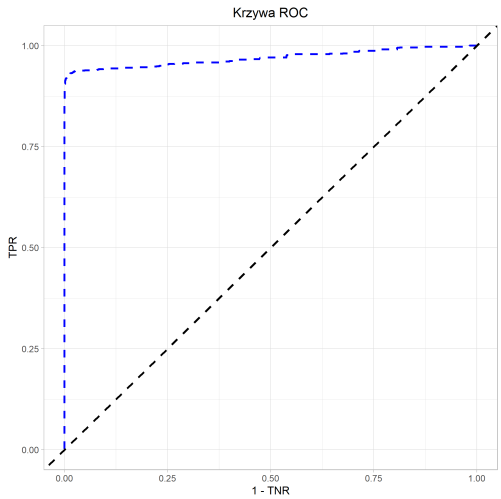
Wniosek

Za optymalny model uznajemy **model regresji logistycznej** oparty o zmienne:

Początkowa kwota kredytu, TTM, Średnie miesięczne opóźnienie,

w którym rozważamy ponadto interakcje pomiędzy zmiennymi.

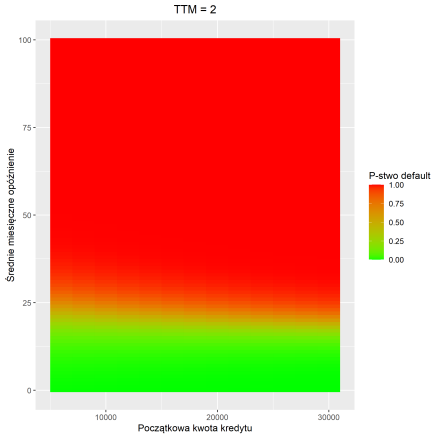
Krzywa ROC



Powyższa krzywa ukazuje, jak predykcja uzyskana dzięki skonstruowanemu modelu różni się od sytuacji bez modelu (czarna prosta). Pole pod otrzymaną krzywą ROC równe 0.9694 świadczy o bardzo dobrej predykcji (pole równe 1 - predykcja idealna). Oś *OY* odpowiada czułości testu (*TPR - True Positive Rate*), natomiast oś *OX* swoistości testu (*TNR - True Negative Rate*).

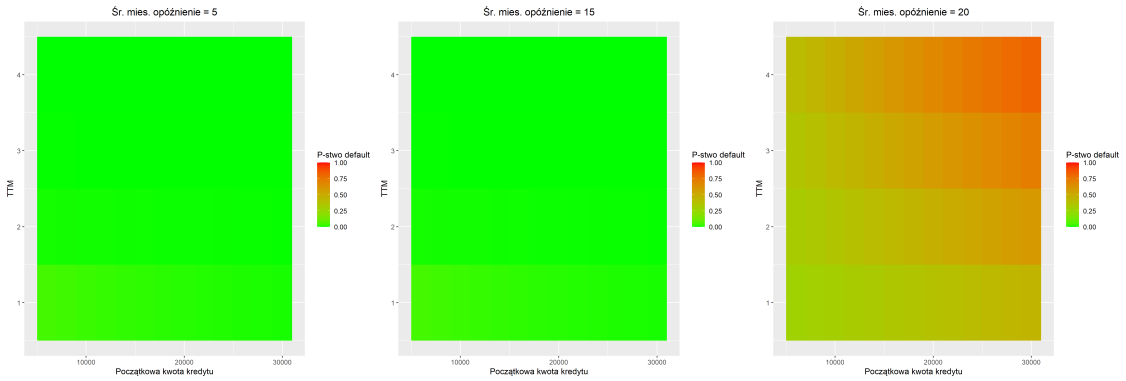
Predykcja

Wyznaczmy przewidywane prawdopodobieństwo zdarzenia default w zależności od wartości zmiennych objaśniających *Początkowa kwota kredytu, TTM, Średnie miesięczne opóźnienie.*



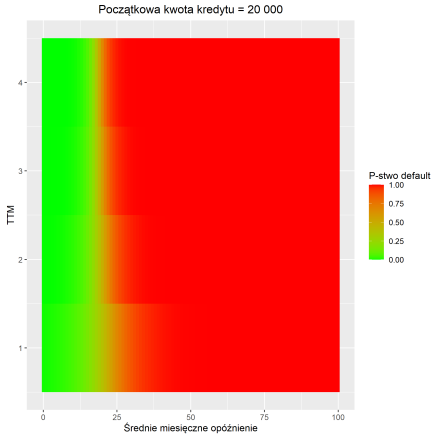
Analiza powyższego wykresu (dla innych wartości *TTM* wykresy analogiczne) pokazuje, że:

- *Początkowa kwota kredytu* NIE wpływa istotnie na prawdopodobieństwo wystąpienia zdarzenia default.
- *Średnie miesięczne opóźnienie* ma z kolei bardzo duży wpływ na zdarzenie default.
- Dla średniego opóźnienia poniżej 20 dni p-stwo jest znikome, następnie gwałtownie rośnie, osiągając wartość 1 dla ok. 30 dni opóźnienia.



Analiza powyższych wykresów pozwala wysunąć następujące wnioski:

- *Średnie miesięczne opóźnienie* jest kluczowym parametrem - dla 5 oraz 15 dni opóźnienia p-stwo zdarzenia default jest w przybliżeniu równe 0 bez względu na TTM oraz *początkową kwotę kredytu*.
- W przypadku 20 dni opóźnienia (i więcej) obserwujemy następującą zależność: im większa wartość kredytu oraz liczba lat do końca kredytu (TTM), tym większe jest p-stwo zdarzenia default.



Powyższy wykres (dla innych kwot kredytów wykres analogiczny) pokazuje, że:

- *Średnie miesięczne opóźnienia* powyżej 25 dni stanowi grupę bardzo wysokiego ryzyka pojawienia się zdarzenia default.
- Liczba lat pozostałych do końca kredytu TTM również wpływa na p-stwo pojawienia się zdarzenia default. Im mniej pozostałych do końca kredytu lat, tym mniejsze jest ryzyko defaulta.

Wnioski

- Za najważniejszą zmienną odpowiadającą za ryzyko wystąpnienia zdarzenia default uznajemy *średnie miesięczne opóźnienie*. Jeśli średnia wartość przekracza 25 dni, mamy do czynienia z grupą bardzo wysokiego ryzyka zdarzenia default (p-stwo zdarzenia default na poziomie 75-100%). Kredyty posiadające średnio powyżej 20 dni opóźnienia również stanowią grupę ryzyka (25-50%).
- Ryzyko zależy także od *początkowej kwoty kredytu* oraz liczby lat pozostałych do końca kredytu (TTM). Kredyty dla których wartości tych zmiennych są duże ($TTM > 3$, *kwota kredytu* $> 20\,000$) również należy zaliczyć do grupy ryzyka, o ile *średnie miesięczne opóźnienie* przekracza 20 dni (p-stwo ok 50% dla kredytów bez większego średniego opóźnienia).
- W szczególności nowe kredyty (największa wartość TTM) są zagrożone wystąpieniem zdarzenia *default* w większym stopniu niż kredyty w finałowej fazie spłaty.

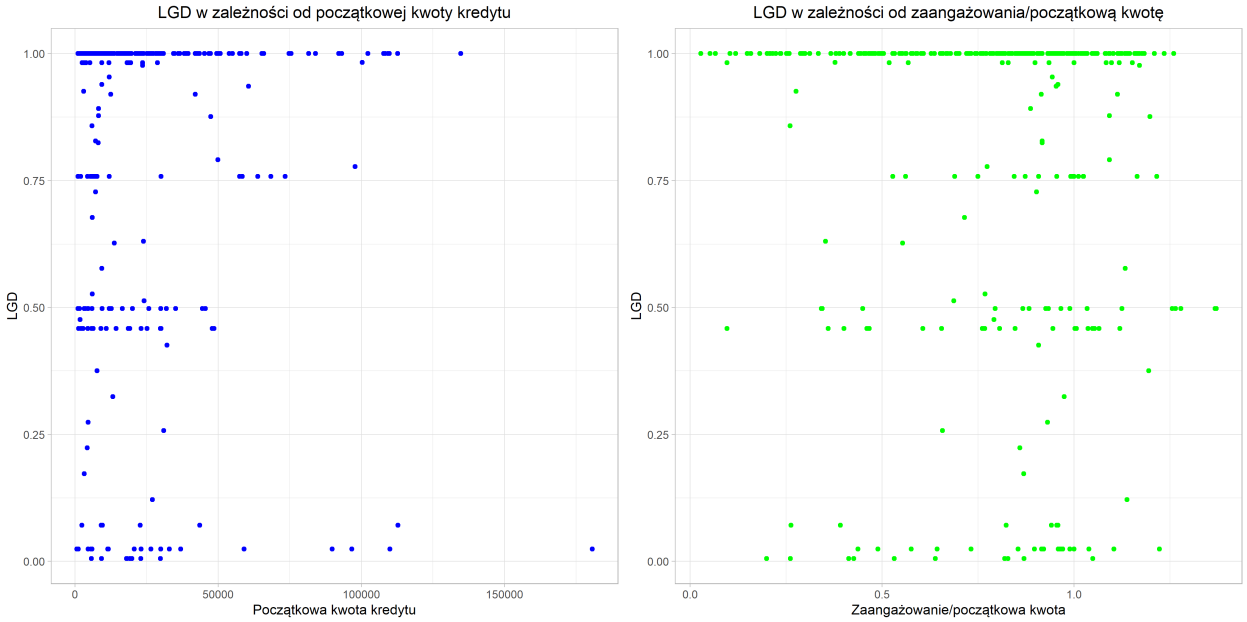
Zadanie 5

Celem zadania jest **zamodelowanie LGD** po 24 miesiącach od przejścia kredytu w stan default na podstawie dostępnych danych. Dla każdego z kredytów, które znalazł się w stanie default i był notowany w okresie 24 miesięcy od przejścia w stan default wyznaczamy następujące dane:

- początkowa kwota kredytu,
- zaangażowanie w stosunku do początkowej kwoty kredytu w momencie na miesiąc przed defaultem,
- liczba lat do końca kredytu na miesiąc przed defaultem,
- średnie miesięczne opoźnienie przed znalezieniem się w stanie default.

Analiza danych

Zbadajmy, jak wartość LGD zależy od każdej z wyznaczonych danych osobno. W tym celu przeanalizujmy wykresy punktowe zmiennej LGD w zależności od danej zmiennej objaśniającej.

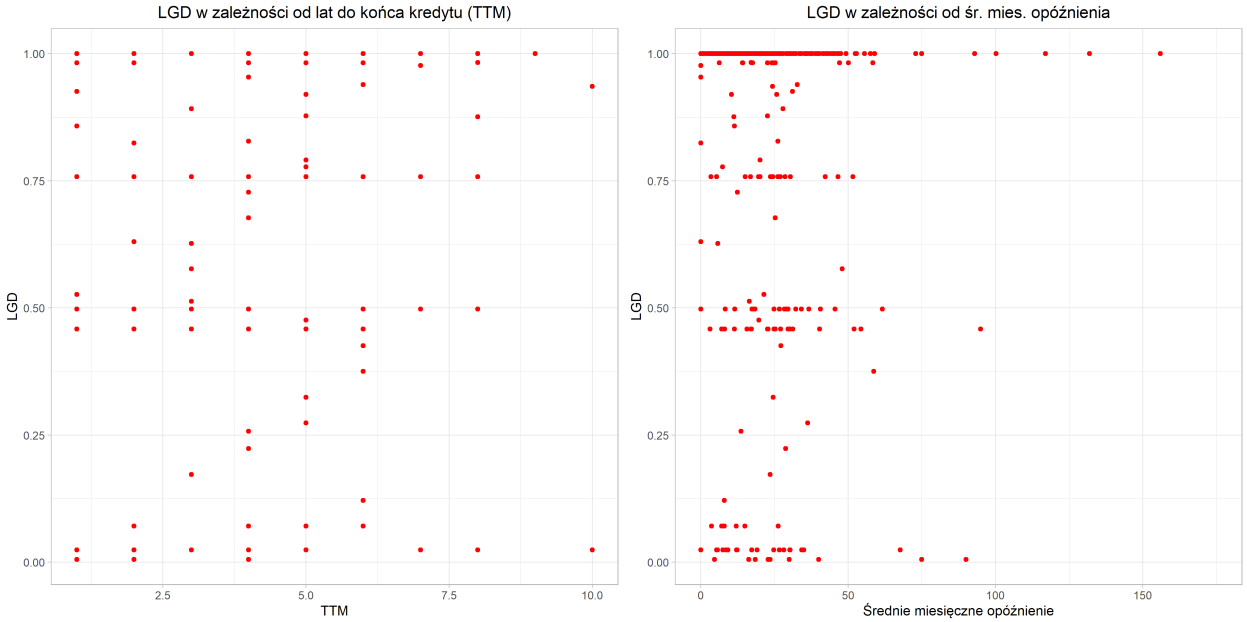


Analiza wykresu LGD w zależności od *początkowej kwoty kredytu* (wykres po lewej stronie) pokazuje, że:

- Rozkład wartości LGD kumuluje się w okolicach wartości: 0, 0.5 oraz 1.
- Większość kredytów opiewa na kwotę poniżej 50 000 - stąd dla tej grupy obserwujemy najwięszkszą liczbę zarówno małych jak i dużych wartości LGD.
- Kredyty powyżej 75 000 stanowią nieliczną grupę i dzielą się zasadniczo na takie, które w ogóle nie były spłacane po przejściu w stan default (LGD=1) oraz kredyty, które powróciły do porfela pracującego (LGD bliskie zera).

Analiza wykresu LGD w zależności od *zaangażowania w momencie default w stosunku do początkowej kwoty kredytu* (wykres po prawej stronie) pokazuje, że:

- Wartość LGD równa 1 występuje dla każdej z wartości badanej zmiennej objaśniającej. Wartości bliskie zero nie występują z kolei dla *zaangażowania* poniżej 0.25.
- Większości obserwacji obejmuje kredyty, które w momencie wejścia w stan default posiadały *zaangażowanie* powyżej 50%.



Analiza wykresu *LGD* w zależności od *lat pozostałych do końca kredytu (TTM)* (wykres po lewej stronie) pokazuje, że:

- Rozkład wartości *LGD* w zależności od tej zmiennej objaśniającej jest równomierny.
- Większość obserwacji dotyczy kredytów, dla których wartość *TTM* jest mniejsze niż 6 lat.

Analiza wykresu *LGD* w zależności od *średniego miesięcznego opóźnienia* (wykres po prawej stronie) pokazuje, że:

- Większość obserwacji obejmuje kredyty posiadające poniżej 50 *dni opóźnienia*.
- Kredyty posiadające powyżej 50 *dni opóźnienia* wykazują się w większości całkowitym zaprzestaniem spłaty ($LGD = 1$).
- W przypadku tej zależności również obserwujemy wyraźną kumulację rozkładu *LGD* wokół wartości: 0, 0.5 oraz 1.

Model regresji liniowej

Skonstruujmy **model regresji liniowej** dla powyższych danych (zakładamy w modelu zależność liniową pomiędzy zmienną *LGD* a zmiennymi objaśniającymi).

Otrzymujemy następujące wyniki dla poszczególnych zmiennych:

zmienna	parametr	p-wartość
Początkowa kwota kredytu	0.00000278	0.00353
Zaangażowanie/początkowa kwota	1.219	< 0.001
TTM	-0.07596	<0.001
Śr. mies. opóźnienie	-0.000947	0.0729

Powyższe dane (zwłaszcza analiza p-wartości) świadczą o tym, że jedyną zmienną, którą możemy uznać za nieistotną jest zmienna *średnie miesięczne opóźnienie*. Sprawdźmy zatem, czy model skonstruowany bez tej zmiennej nie jest istotnie gorszy od pełnego modelu.

Porównanie modelu liniowego z oraz bez zmiennej *średnie miesięczne opóźnienie* metodą *ANOVA* (test Deviance):

otrzymana p-wartość dla testu Deviance wynosi 0.072, zatem nie mamy podstaw do odrzucenia hipotezy mówiącej, że model BEZ zmiennej *średnie miesięczne opóźnienie* jest lepszy od pełnego modelu. Zatem zmienna ta NIE ma istotnego wpływu na predykcję *LGD*.

Porównajmy następnie model skonstruowany bez odrzuconej zmiennej z analogicznym modelem, w którym dodatkowo badamy również interakcje pomiędzy zmiennymi.

Porównanie modelu liniowego bez zmiennej *średnie miesięczne opóźnienie* w wersji z oraz bez interakcji pomiędzy zmiennymi metodą *ANOVA* (test Deviance):

otrzymana p-wartość dla testu Deviance jest mniejsza niż 0.001, zatem odrzucamy hipotezę mówiącą, że model bez interakcji jest lepszy od modelu z interakcjami. Interakcje pomiędzy zmiennymi mają zatem istotny wpływ na zmienną objaśnianą *LGD*.

Wniosek

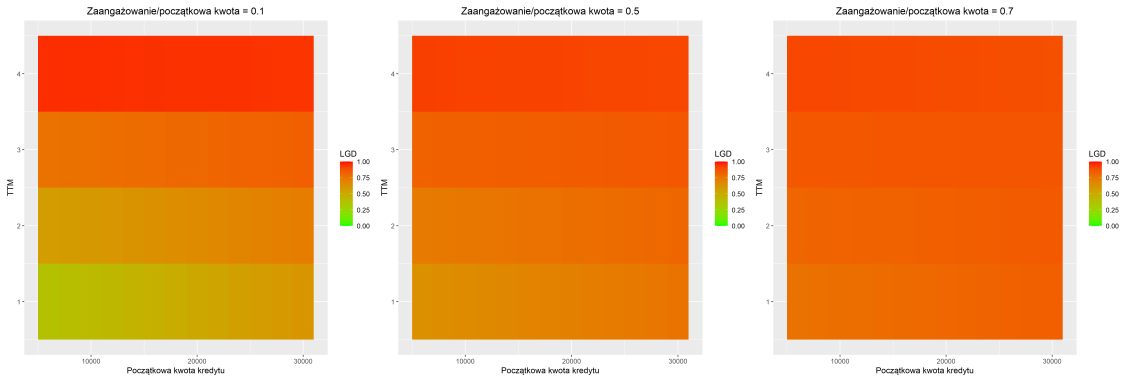
Za optymalny model uznajemy **model regresji liniowej** oparty o zmienne:

Początkowa kwota kredytu, Zaangażowanie/początkowa kwota, TTM,

w którym rozważamy ponadto interakcje pomiędzy zmiennymi.

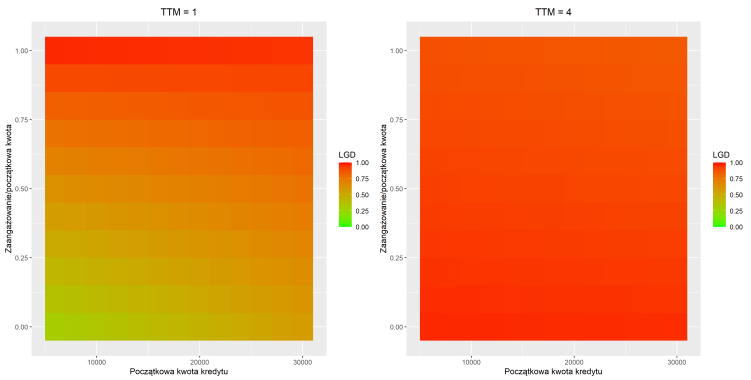
Predykcja

Wyznaczmy przewidywane wartość *LGD* po 24 miesiącach po deafulcie w zależności od wartości zmiennych objaśniających *Początkowa kwota kredytu, Zaangażowanie/początkowa kwota, TTM*.



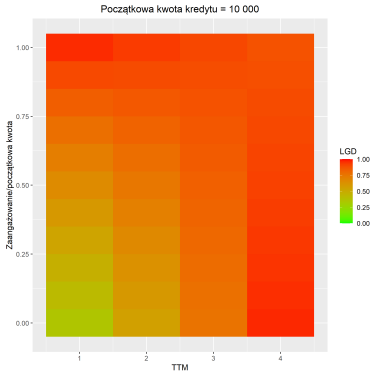
Analiza powyższych wykresów pokazuje, że:

- Zmienna *zaangażowanie/początkowa kwota* ma decydujący wpływ na wartość *LGD*: jeśli *zaangażowanie* w momencie default jest małe, to wysoką grupę ryzyka stanowią jedynie kredyty, których termin spłaty *TTM* jest wysoki. W przypadku dużego *zaangażowania* bez względu na początkową kwotę oraz *TTM* wartość *LGD* osiąga wartości rzędu 0.75.



Analiza powyższych wykresów pokazuje, że:

- Kredyty bliskie końca okresu spłaty ($TTM = 1$) stanowią niską grupę ryzyka - jedynie w przypadku wysokiego zaangażównaia (powyżej 0.75) obserwujemy wysoki wskaźnik LGD.
- Dla kredytów z większą liczbą lat do końca spłaty obserwujemy wysoki wskaźnik LGD bez względu na *zaangażowanie* oraz *początkową kwotę kredytu*.



Analiza powyższego wykresu (dla innych *początkowych kwot kredytu* - wykresy analogiczne), że:

- Wartość LGD rośnie wraz ze wzrostem wartości TTM oraz *zaangażowania/początkowej kwoty*.
- Kredyty bliskie końca okresu spłaty z małym zangażowaniem stanowią grupę niskiego ryzyka: dla $TTM = 1$ oraz *zaangażowania* < 0.25 wartość $LGD < 0.25$.
- Kredyty o *zaangażowaniu* > 0.75 stanowią grupę wysokiego ryzyka zwłaszcza w przypadku małej wartości lat do koca kredytu TTM.

Wnioski:

- Grupa kredytów bliskich końca okresu spłaty ($TTM = 1$) stanowi grupę małego ryzyka: prognozowana wartość LGD po 24 miesiącach po defaulcie jest dla tej grupy mniejsza niż 0.25. Jedynie w przypadku wysokiego *zaangażowania* (powyżej 0.75) takie kredyty stanowią grupę wysokiego ryzyka ($LGD > 0.75$).
- Kredyty z *zaangażowaniem* > 0.75 stanowią grupę wysokiego ryzyka. Bez względu na *początkową kwotę kredytu* oraz TTM wartość LGD dla tej grupy zwykle przekracza poziom 0.75.
- Sama wartość *początkowej kwoty kredytu* nie wpływa istotnie na wartość LGD.