

Analiza odchodzenia klientów

Bartosz Chądryński & Michał Turek

2023-05-05

Wstęp

Preprocessing

Wykresy

Zacznijmy od analizy wykresów. Na początek zmienne ciągłe. Na wykresie 1 i 3 widzimy, że zmienne te są w znacząco różnych skalach, więc prawdopodobnie potrzebna będzie normalizacja. Natomiast na wykresie 2 i 4 widać, że każda ze zmiennych ma istotnie różny rozkład, gdy pogrupujemy ją ze względu na Churn.

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Fri May 5 21:15:05 2023

	V1	V2	V3	V4	V5	V6
tenure	Min. : 1.00	1st Qu.: 9.00	Median :29.00	Mean :32.42	3rd Qu.:55.00	Max. :72.00
MonthlyCharges	Min. : 18.25	1st Qu.: 35.59	Median : 70.35	Mean : 64.80	3rd Qu.: 89.86	Max. :118.75
TotalCharges	Min. : 18.8	1st Qu.: 401.4	Median :1397.5	Mean :2283.3	3rd Qu.:3794.7	Max. :8684.8

Z kolei na 5 widzimy, że w niektórych przypadkach są duże różnice w ilości obserwacji z każdej kategorii, jeśli chodzi o daną zmienną. W szczególności takimi zmiennymi są *PhoneService*, czy *MultipleLines*.

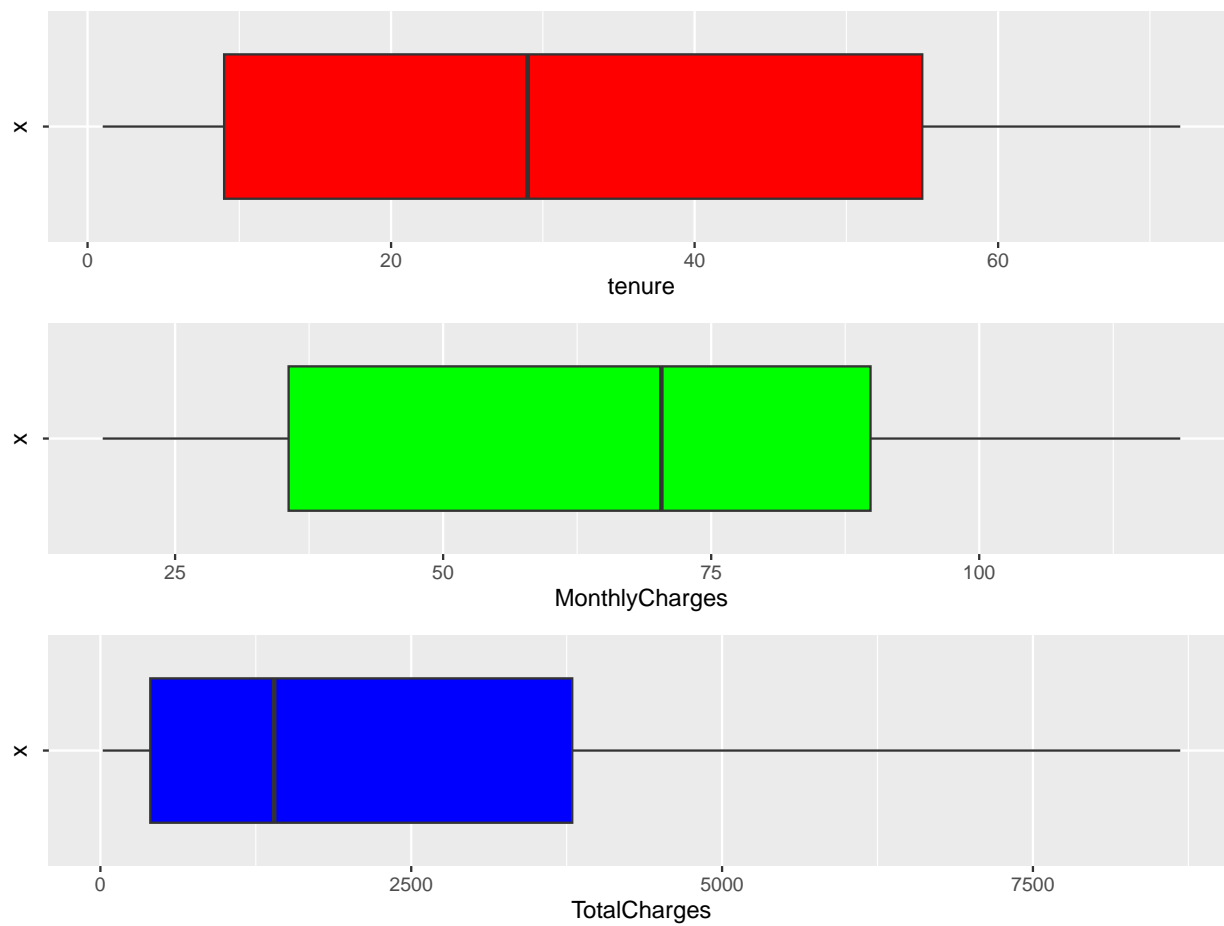


Figure 1: Boxploty zmiennych ciągłych

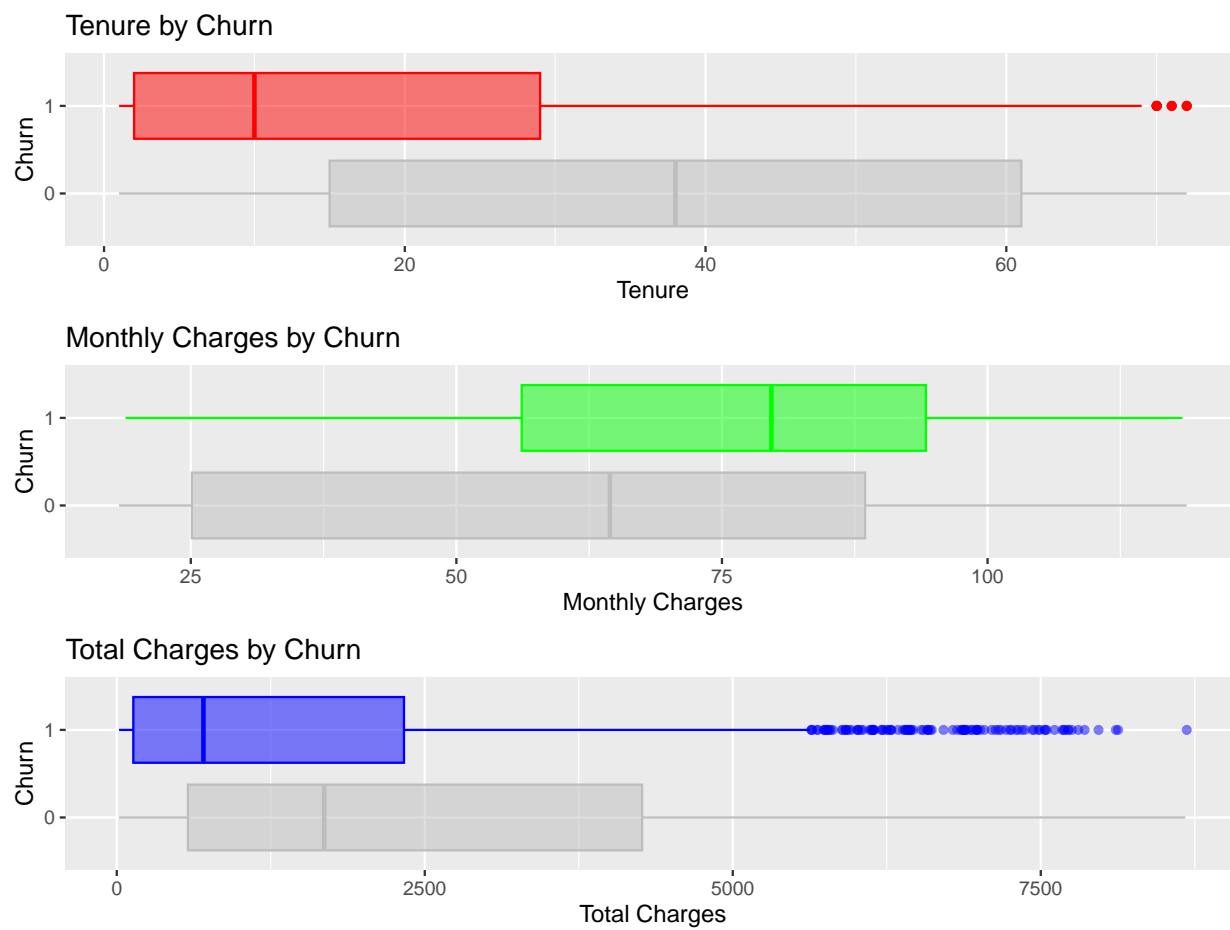


Figure 2: Boxploty zmiennych ciągłych z podziałem ze względu na Churn

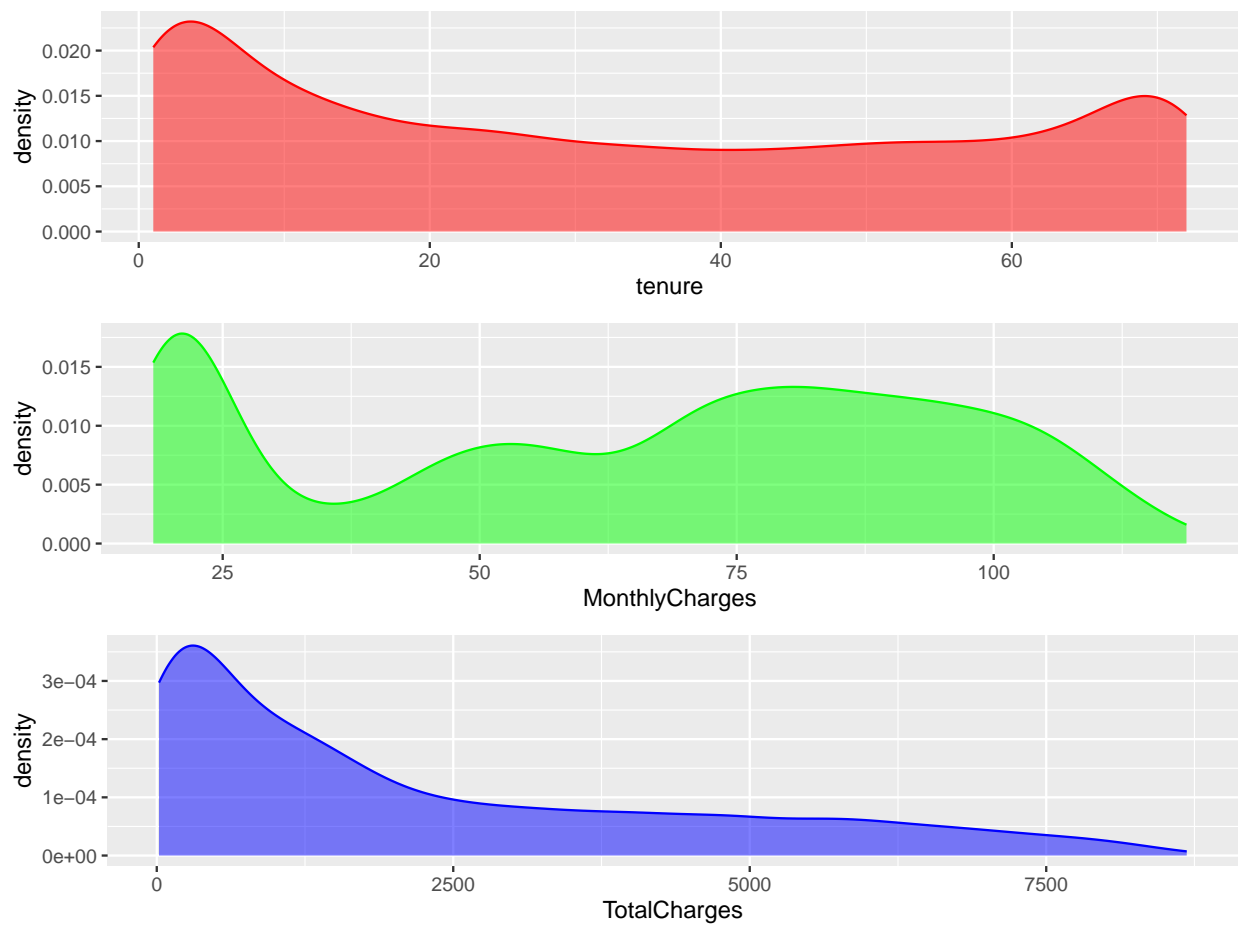


Figure 3: Estymator jądrowy gęstości

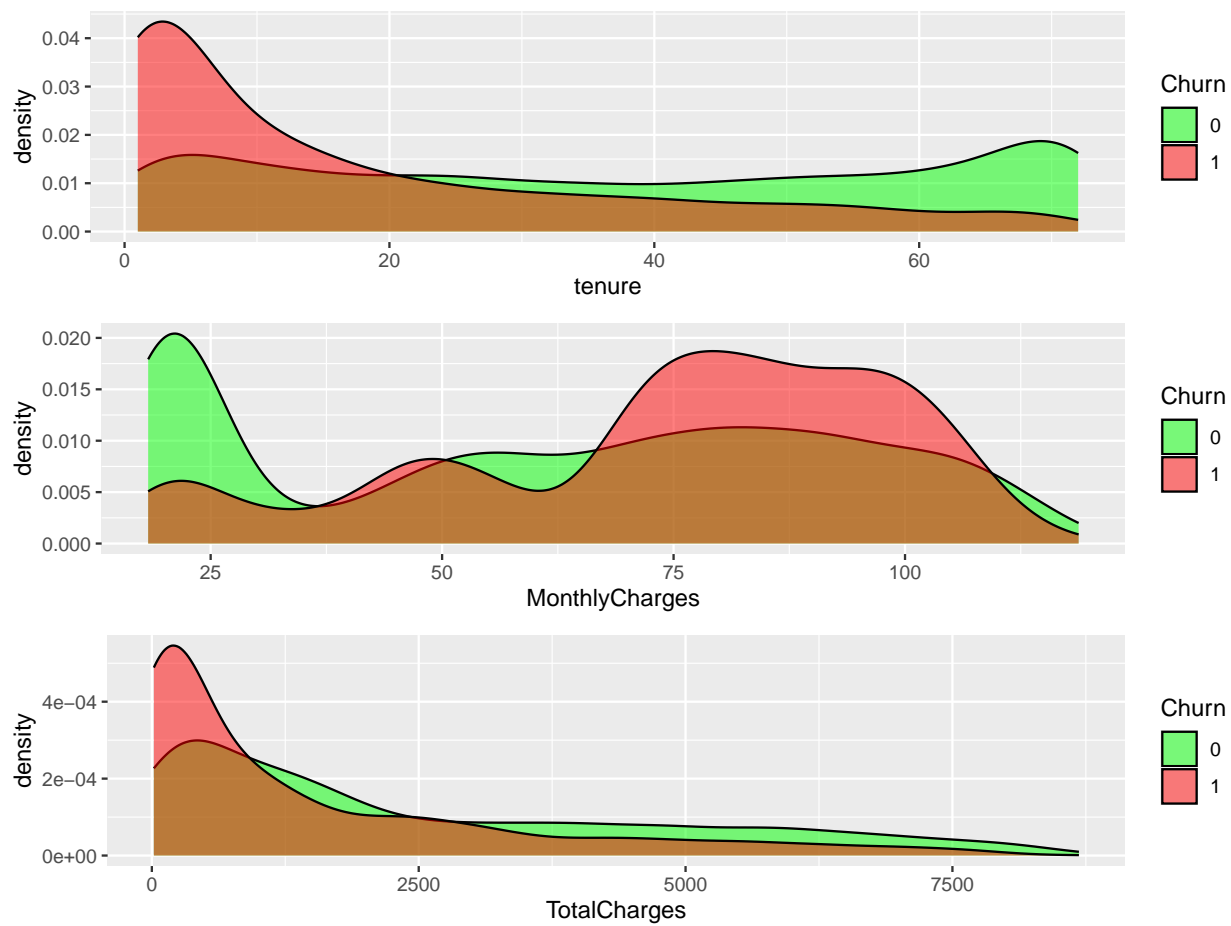


Figure 4: Estymator jądrowy gęstości z podziałem ze względu na Churn

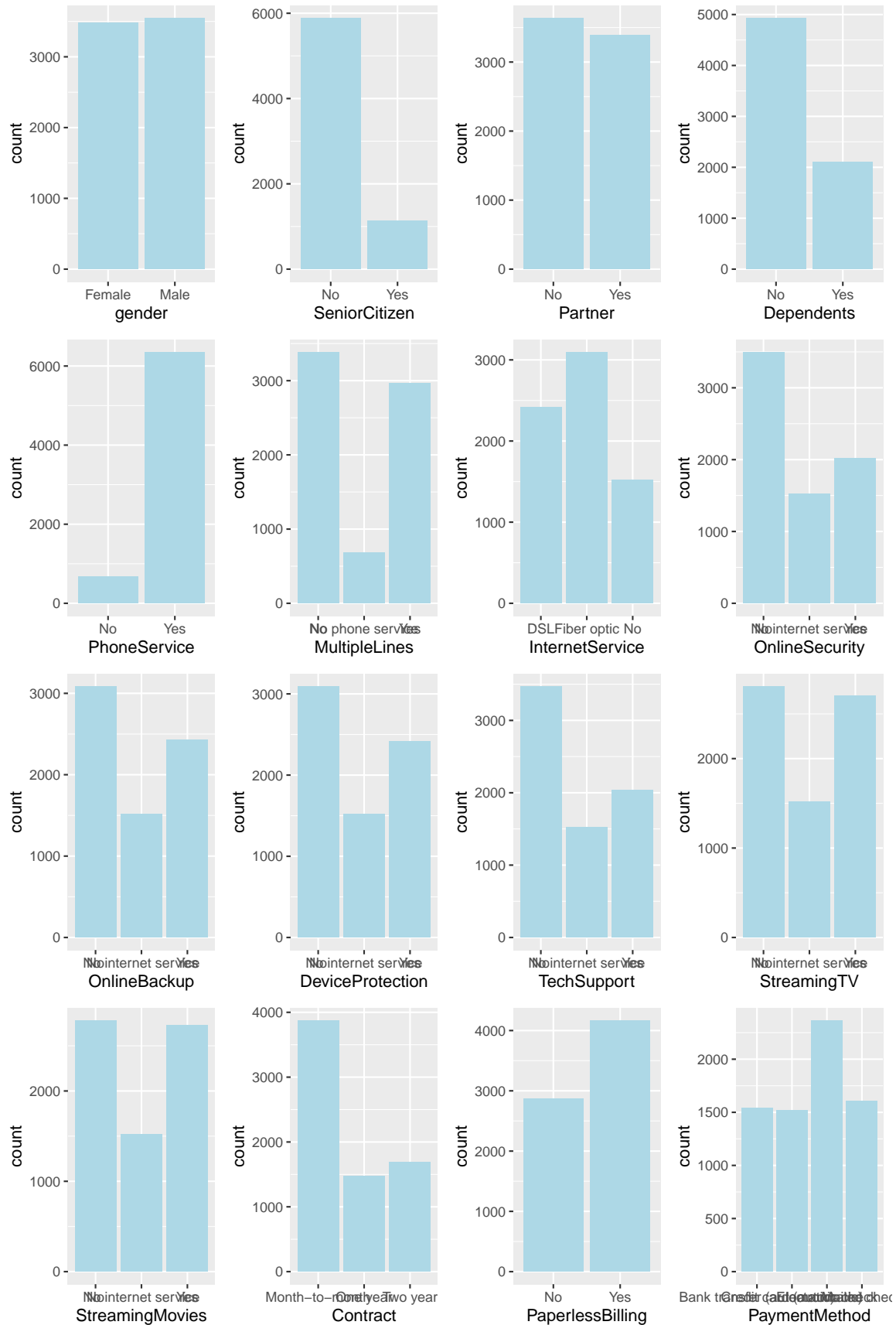


Figure 5: Wykres ilości obserwacji z podziałem na kategorie zmiennych

Interpretacja Wyników

W naszych danych jest zaledwie 11 obserwacji z brakującymi danymi (na 7033 łącznie). Zatem zasadne jest pominięcie ich w trakcie analizy danych. Nie stosujemy żadnej imputacji. Ilość danych może być obciążająca dla niektórych modeli. Jeśli będą występowały problemy ze złożonością obliczeniową, to dla konkretnego modelu będziemy decydować o przeprowadzeniu analizy dla ewentualnego podzbioru danych.

W tabeli poniżej mamy macierz korelacji zmiennych ciągłych. Jak widać istnieje mocna korelacja pomiędzy tym jak długo klient korzysta/korzystał z usług, a kwotą jaką zapłacił za usługi. Nie powinno to dziwić. Na razie jednak nie decydujemy się na wyrzucenie którejs z zmiennych, ponieważ zarówno czas jak i koszt może być istotny w kontekście odchodzenia klientów. Te dwie rzeczy nie muszą być ze sobą powiązane w pełni. Może być tak, że odchodzą głównie nowi klienci, niezależnie od tego ile płacą. Albo może być tak, że odchodzą klienci, którzy zapłacili rachunki powyżej pewnej sumy, niekoniecznie będący długo/krótco stażem.

	tenure	MonthlyCharges	TotalCharges
tenure	1.00	0.25	0.83
MonthlyCharges	0.25	1.00	0.65
TotalCharges	0.83	0.65	1.00

Najprawdopodobniej potrzebne będzie wykonanie transformacji danych, w szczególności normalizacji. Natomiast jeśli chodzi o obserwacje odstające, to nie ma ich za dużo. Pojawiają się licznie w przypadku zmiennej *TotalCharges* pogrupowanej ze względu na *Churn*. Widać, że jest tendencja, aby odchodzący klienci należeć do jednej z dwóch grup. Są albo nowymi klientami, albo klientami z dużym stażem. Ta druga grupa jest na wykresie pudełkowym interpretowana jako obserwacje odstające. W rzeczywistości należy to interpretować tak, że rozkład tej zmiennej jest dwumodalny, nie będziemy stosować technik mających na celu ignorowanie lub zmniejszenie wpływu tych obserwacji, znacząco odbiegających od reszty.

Klasyfikacja

Regresja Liniowa

Zacznijmy od metod, w których bierzemy pod uwagę jedynie zmienne ciągłe. Na początek regresja liniowa.

	0	1
0	968	220
1	64	153

Table 1: Confusion matrix at threshold = 1.52

Regresja Logistyczna

	0	1
0	942	163
1	90	210

Table 2: Confusion matrix at threshold = 0.52

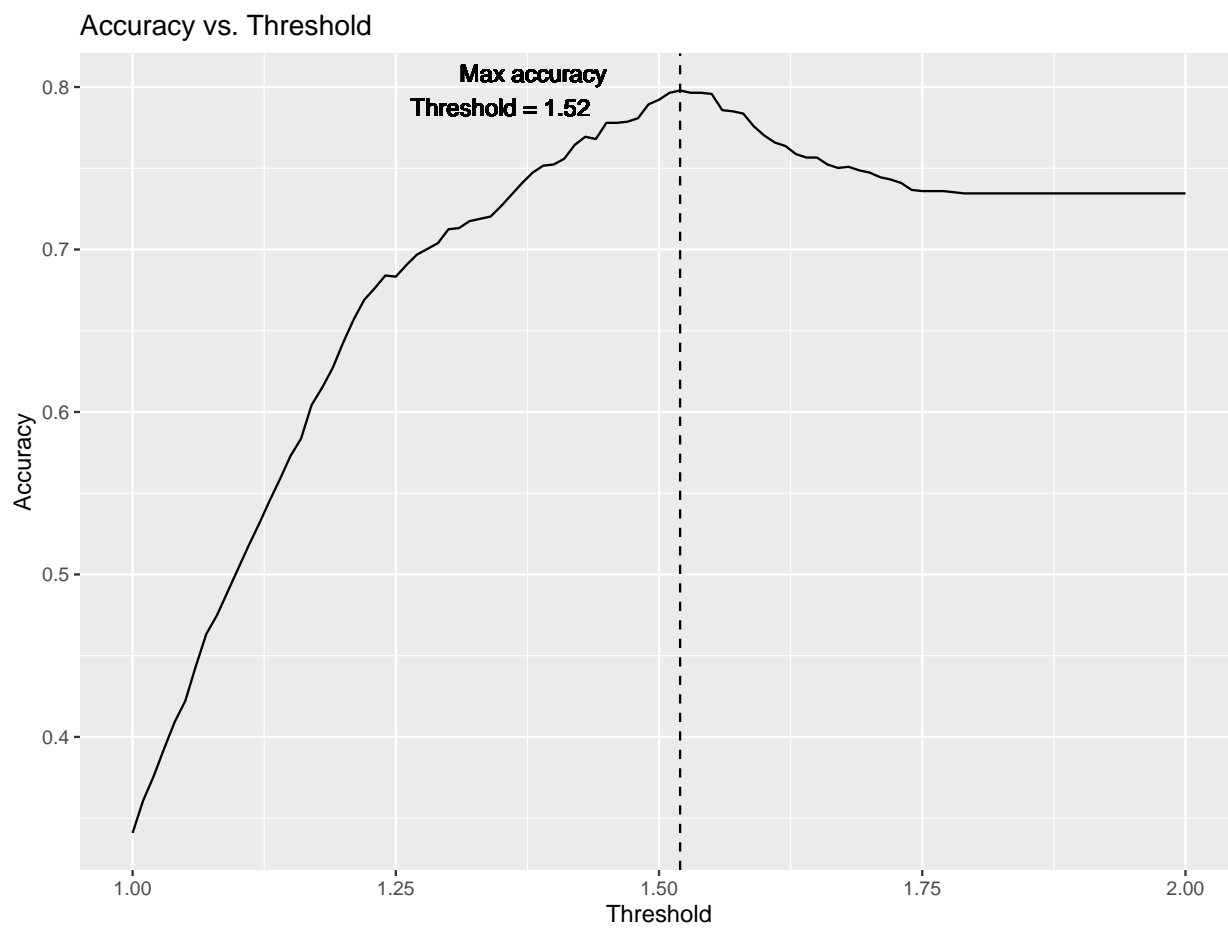


Figure 6: Skuteczność predykcji dla poszczególnych punktów odcięcia

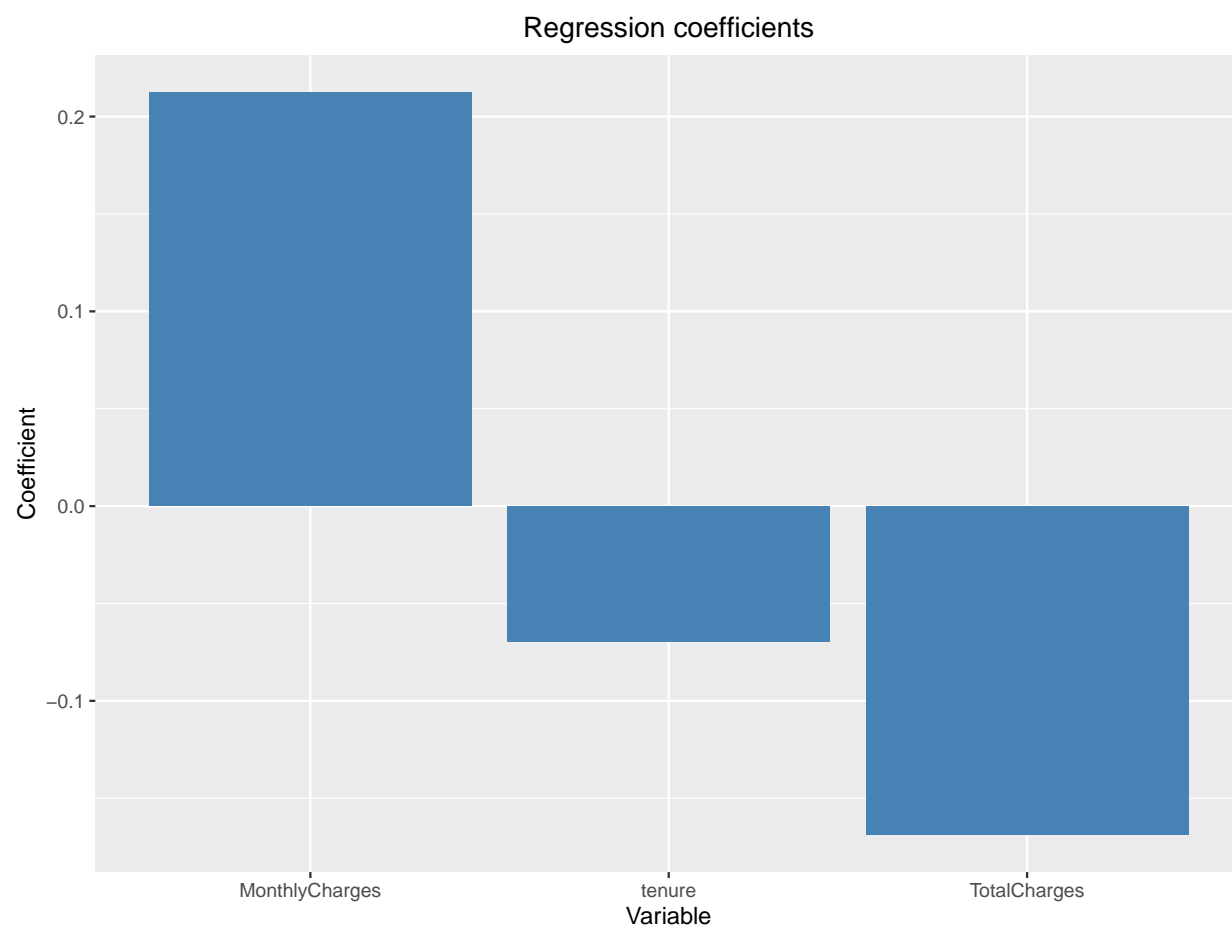


Figure 7: wartości współczynników w modelu regresji logistycznej

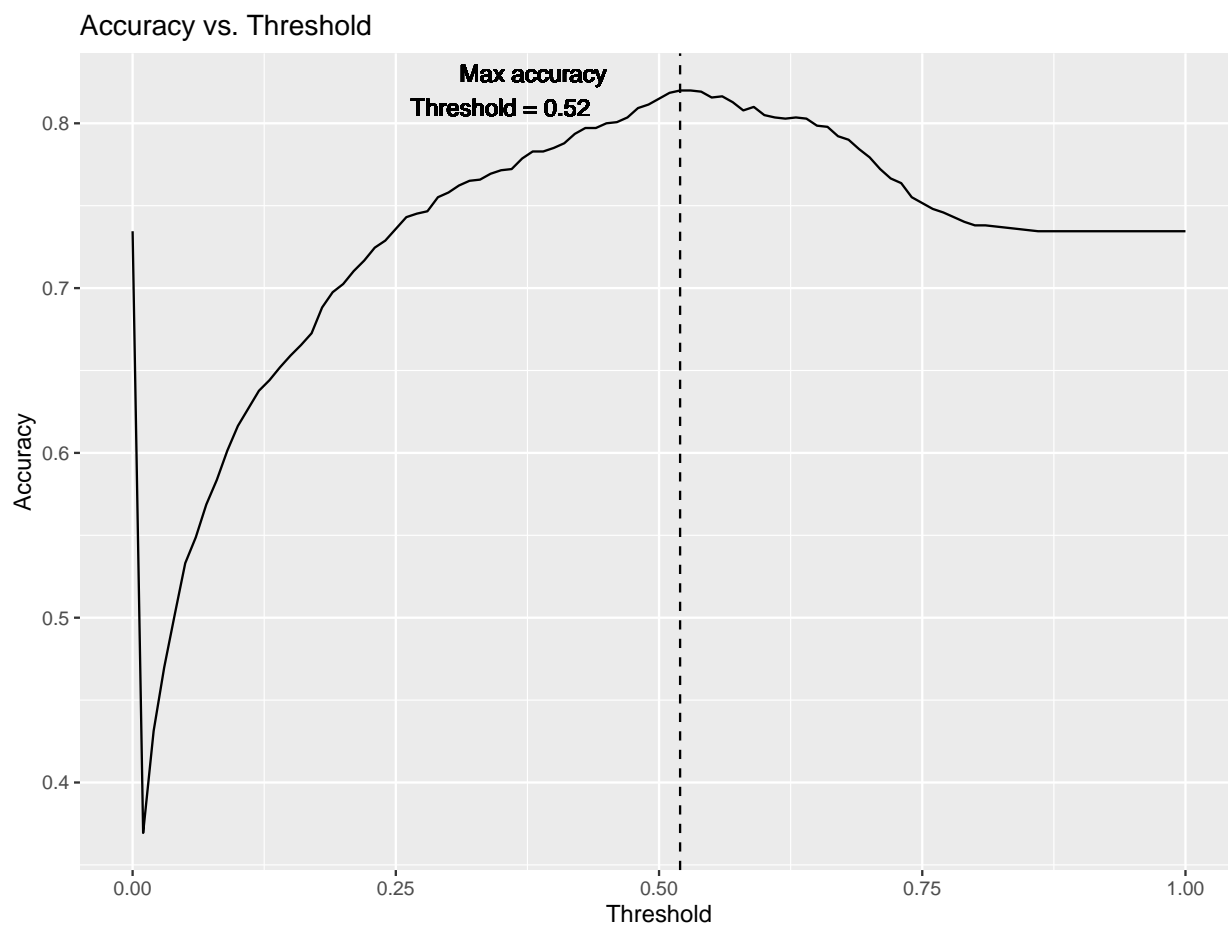


Figure 8: Skuteczność predykcji dla poszczególnych punktów odcięcia

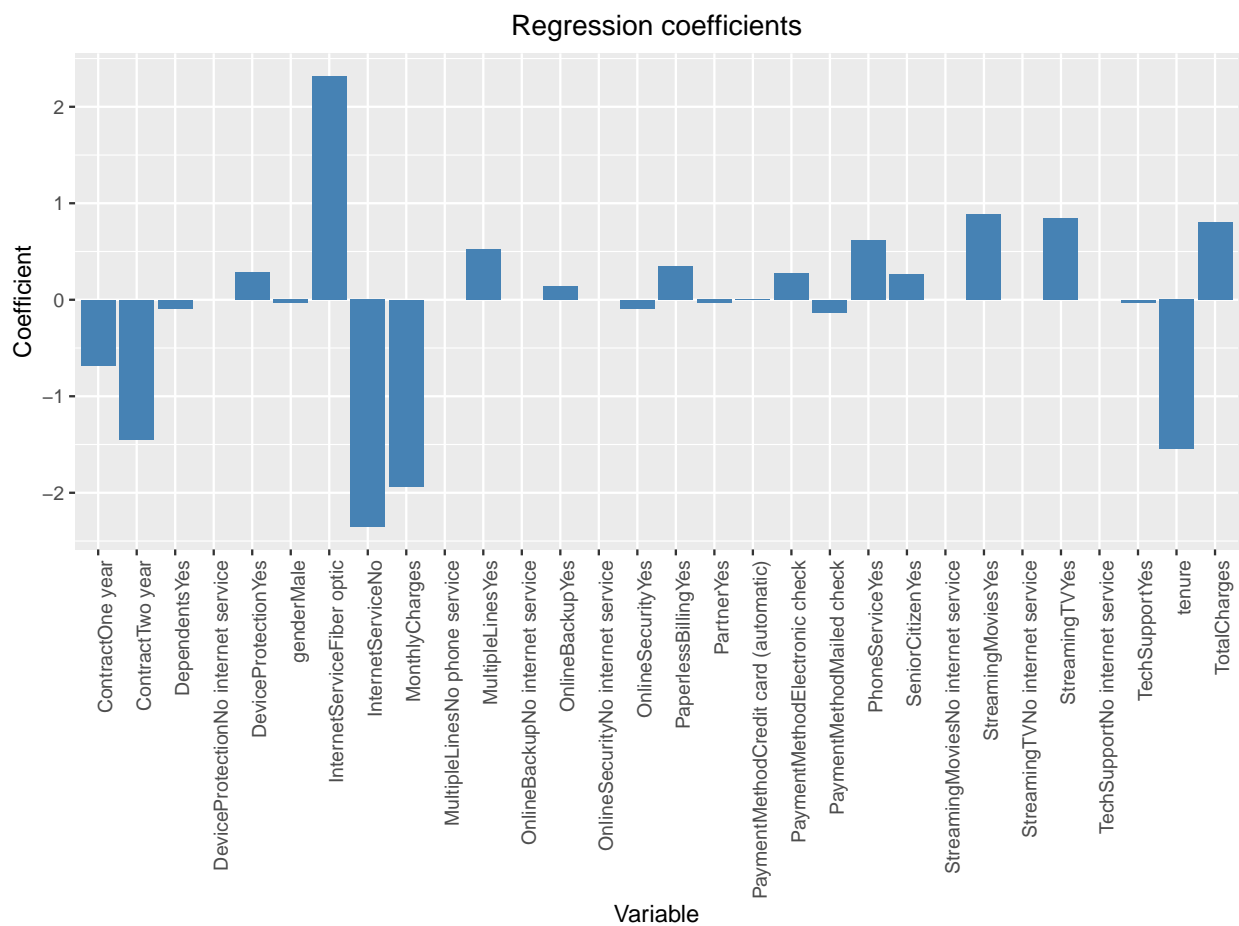


Figure 9: wartości współczynników w modelu regresji logistycznej

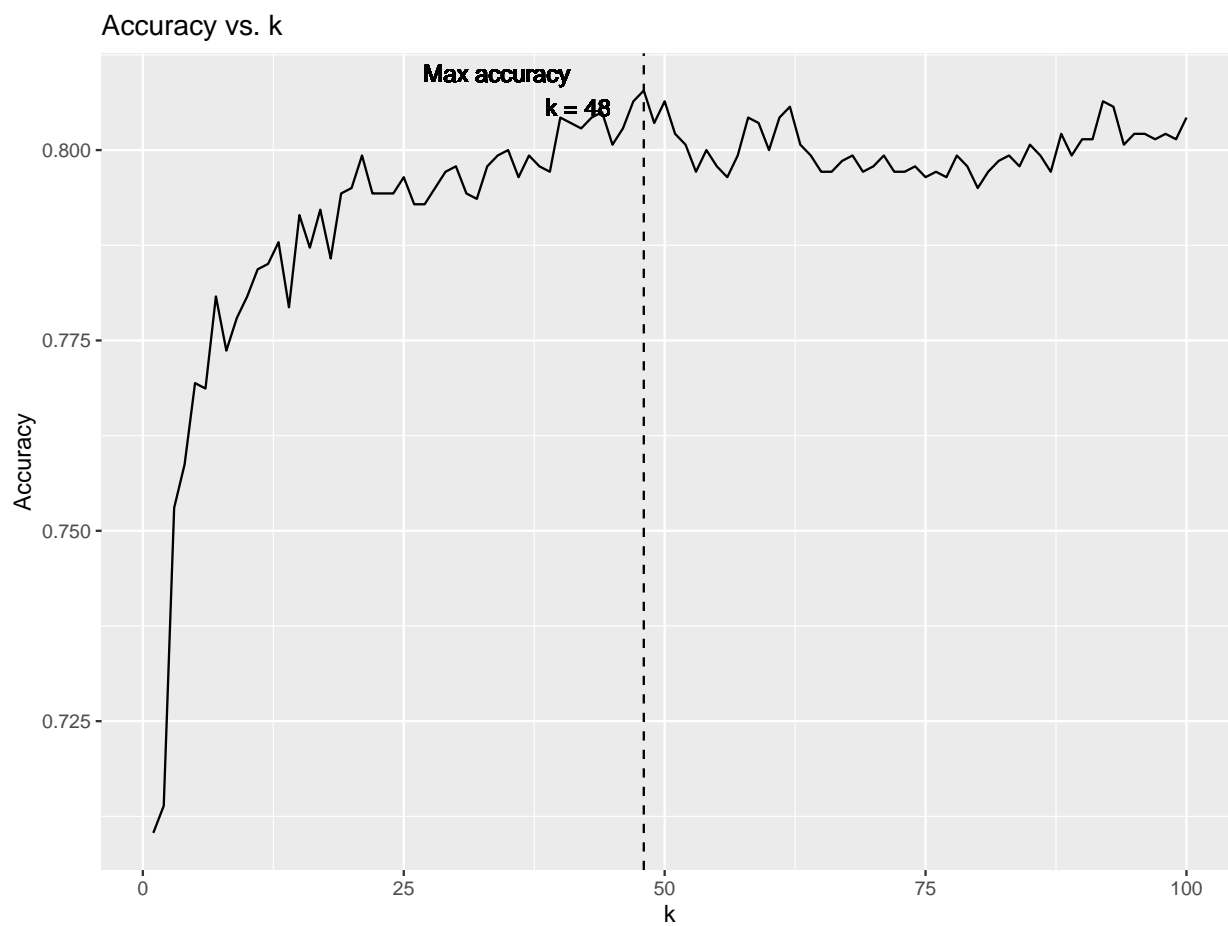


Figure 10: Skuteczność predykcji dla poszczególnych wartości k

	0	1
0	778	254
1	90	283

Table 3: Confusion matrix

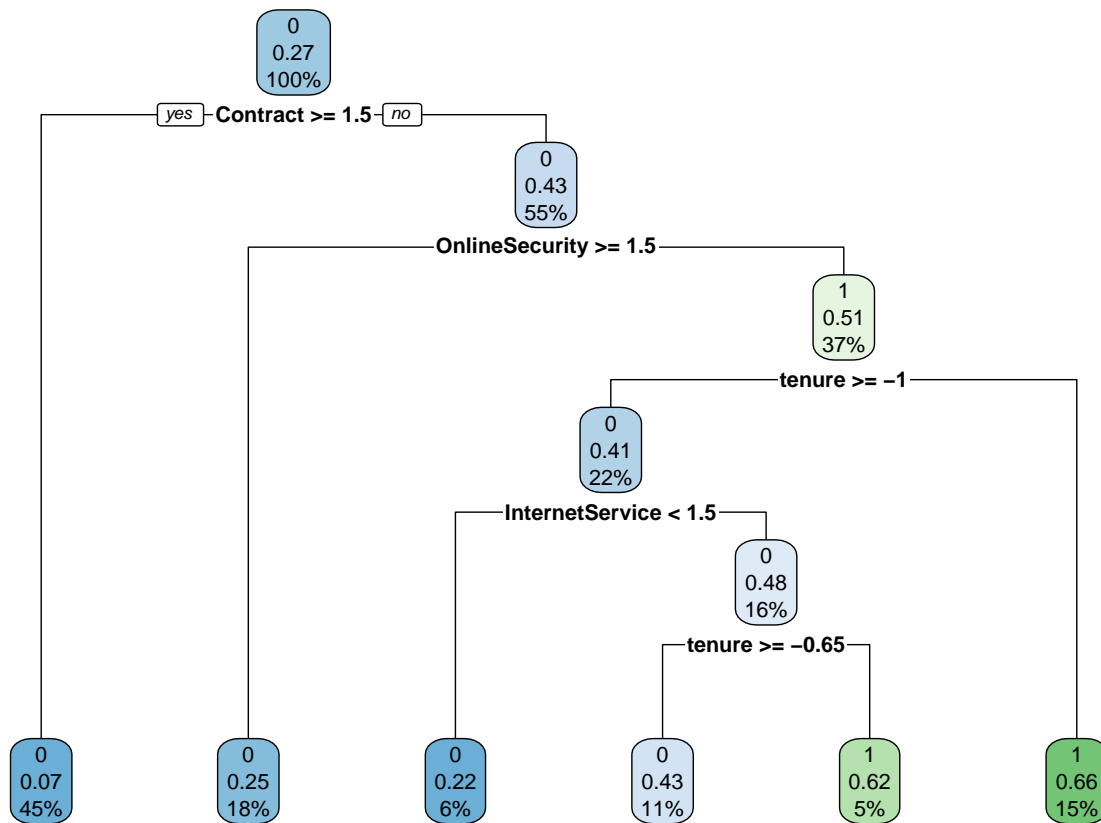
	0	1
0	917	115
1	158	215

Table 4: Confusion matrix for k = 48

Algorytm Naiwnego Bayesa

Algorytm k sąsiadów

Drzewo decyzyjne



	0	1
0	934	98
1	201	172

Table 5: Confusion matrix