Analiza odchodzenia klientów

Bartosz Chądzyński 255680 & Michał Turek 246993

2023-05-06

Wstęp

Nasz projekt będzie dotyczył analizy danych dotyczących odchodzenia klientów firmy telekomunikacyjnej. Naszym celem jest zrozumienie, jakie czynniki wpływają na decyzję klientów o pozostaniu lub odejściu od firmy oraz jak te czynniki wpływają na skuteczność działań związanych z retencją klientów. W ramach projektu przeprowadzamy analizę danych, w tym eksploracyjną analizę, w której badamy rozkłady zmiennych oraz korelacje między nimi. Wprowadzamy również preprocessing danych, w tym normalizację oraz kodowanie zmiennych kategorycznych. Następnie tworzymy modele predykcyjne, które pozwalają na przewidywanie odchodzenia klientów. Przetestujemy różne algorytmy klasyfikacji, dobierając ostatecznie najlepszy.W efekcie naszej analizy otrzymujemy narzędzie predykcyjne.

Preprocessing

Analiza opisowa

Zbiór danych Telco Customer Churn składa się z 7043 obserwacji (klientów) i 21 zmiennych.

-customerID - unikalny identyfikator klienta

-gender - płeć klienta

-SeniorCitizen - czy klient jest emerytem (1) czy nie (0)

-Partner - czy klient ma partnera (Tak/Nie)

-Dependents - czy klient ma na utrzymaniu innych członków rodziny (Tak/Nie)

-tenure - okres w miesiącach, przez który klient był klientem firmy

-PhoneService - czy klient korzysta z usług telefonicznych (Tak/Nie)

-MultipleLines - czy klient ma więcej niż jedną linię telefoniczną (Tak/Nie/Brak usługi)

-InternetService - typ łącza internetowego (DSL, Fiber optic, Brak usługi)

-OnlineSecurity - czy klient korzysta z usług zabezpieczeń internetowych (Tak/Nie/Brak usługi)

-OnlineBackup - czy klient korzysta z usług kopii zapasowych danych online (Tak/Nie/Brak usługi)

-DeviceProtection - czy klient korzysta z usług zabezpieczeń urządzeń (Tak/Nie/Brak usługi)

-TechSupport - czy klient korzysta z usług technicznej pomocy (Tak/Nie/Brak usługi)

-StreamingTV - czy klient korzysta z usług strumieniowego przesyłania telewizji (Tak/Nie/Brak usługi)

-StreamingMovies - czy klient korzysta z usług strumieniowego przesyłania filmów (Tak/Nie/Brak usługi)

-Contract - typ umowy (Month-to-month, One year, Two year)

-PaperlessBilling - czy klient otrzymuje faktury w formie papierowej (Tak/Nie)

-PaymentMethod - metoda płatności (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

-MonthlyCharges - miesięczny rachunek klienta

-TotalCharges - łączny rachunek klienta

-Churn - czy klient zrezygnował z usług firmy (Tak/Nie).

Wszystkie zmienne są w formie tekstowej, lub binarnej, oprócz trzech zmiennych numerycznych: SeniorCitizen, tenure, MonthlyCharges oraz jednej zmiennej numerycznej typu float: TotalCharges. Na początku dokonamy analizy tych trzech zmiennych numerycznych, wykorzystując podstawowe statystyki.

	tenure	MonthlyCharges	TotalCharges
X	Min. : 1.00	Min. : 18.25	Min. : 18.8
X.1	1st Qu.: 9.00	1st Qu.: 35.59	1st Qu.: 401.4
X.2	Median :29.00	Median : 70.35	Median :1397.5
X.3	Mean :32.42	Mean : 64.80	Mean :2283.3
X.4	3rd Qu.:55.00	3rd Qu.: 89.86	3rd Qu.:3794.7
X.5	Max. :72.00	Max. :118.75	Max. :8684.8

Badając mediany i średnie poszczególnych zmiennych z tabeli **??** możemy wyciągnąć kilka wniosków. Na przykład średnia wartość miesięcznej opłaty to 64.76 dolara, a mediana to 70.35 dolara. Można z tego wnioskować, że rozkład tej zmiennej jest skośny w lewo, co sugeruje, że większość klientów płaci więcej niż średnia wartość.Średni czas trwania umowy wynosi 32.37 miesiąca, a mediana to 29 miesięcy. Można zauważyć, że większość klientów trzyma się firmy przez mniej niż 3 lata. Średnia wartość MonthlyCharge dla klientów, którzy odeszli (churn=Yes), wynosi 74.44 dolarów, podczas gdy dla klientów, którzy pozostali (churn=No), wynosi 61.27 dolarów. Można z tego wnioskować, że klienci, którzy płacą więcej za usługi, są bardziej skłonni do zrezygnowania z nich. Są to oczywiście tylko przykładowe wnioski, które możemy wyciągnąć z danych na podstawie prostych statystyk. W dalszych częściach pracy będziemy analizowali dane z pomocą modeli o różnej złożoności.

Spójrzmy teraz na pozostałe zmienne. Na podstawie rozkładu zmiennych w poszczególnych kategoriach możemy wyciągnąć kilka wniosków (udział ten można zobaczyć na histogramach w kolejnym podrozdziale). Między innymi: -Większość klientów to osoby indywidualne (71,5%).

-Większość klientów korzysta z usługi telefonii cyfrowej (90,3%).

-Większość klientów korzysta z faktury elektronicznej (70,4%).

-Większość klientów nie korzysta z usługi ochrony urządzeń (90,1%).

-Około połowa klientów korzysta z usługi internetu szerokopasmowego (46,8%).

Z powyższych danych można wywnioskować, że firma powinna skupić się na promowaniu usługi internetu szerokopasmowego oraz usługi ochrony urządzeń, aby zwiększyć liczbę klientów korzystających z tych usług. Dodatkowo, firma powinna zastanowić się nad przyczynami, dla których tak mało klientów korzysta z faktury elektronicznej i ewentualnie wdrożyć działania promocyjne, zachęcające do korzystania z tej formy rozliczenia.

Wykresy

Następnie przejdźmy do 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.

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*.

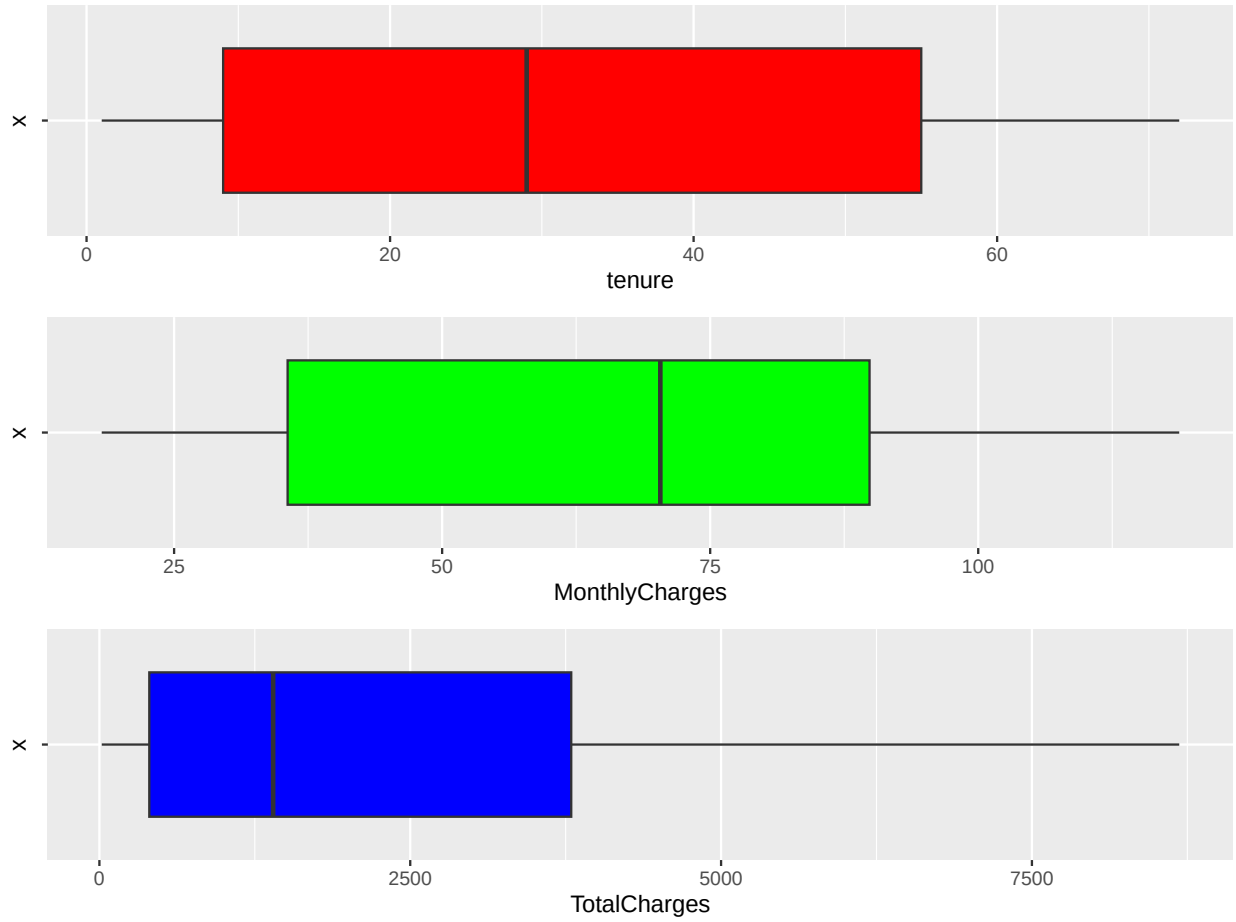


Figure 1: Boxploty zmiennych ciągłych

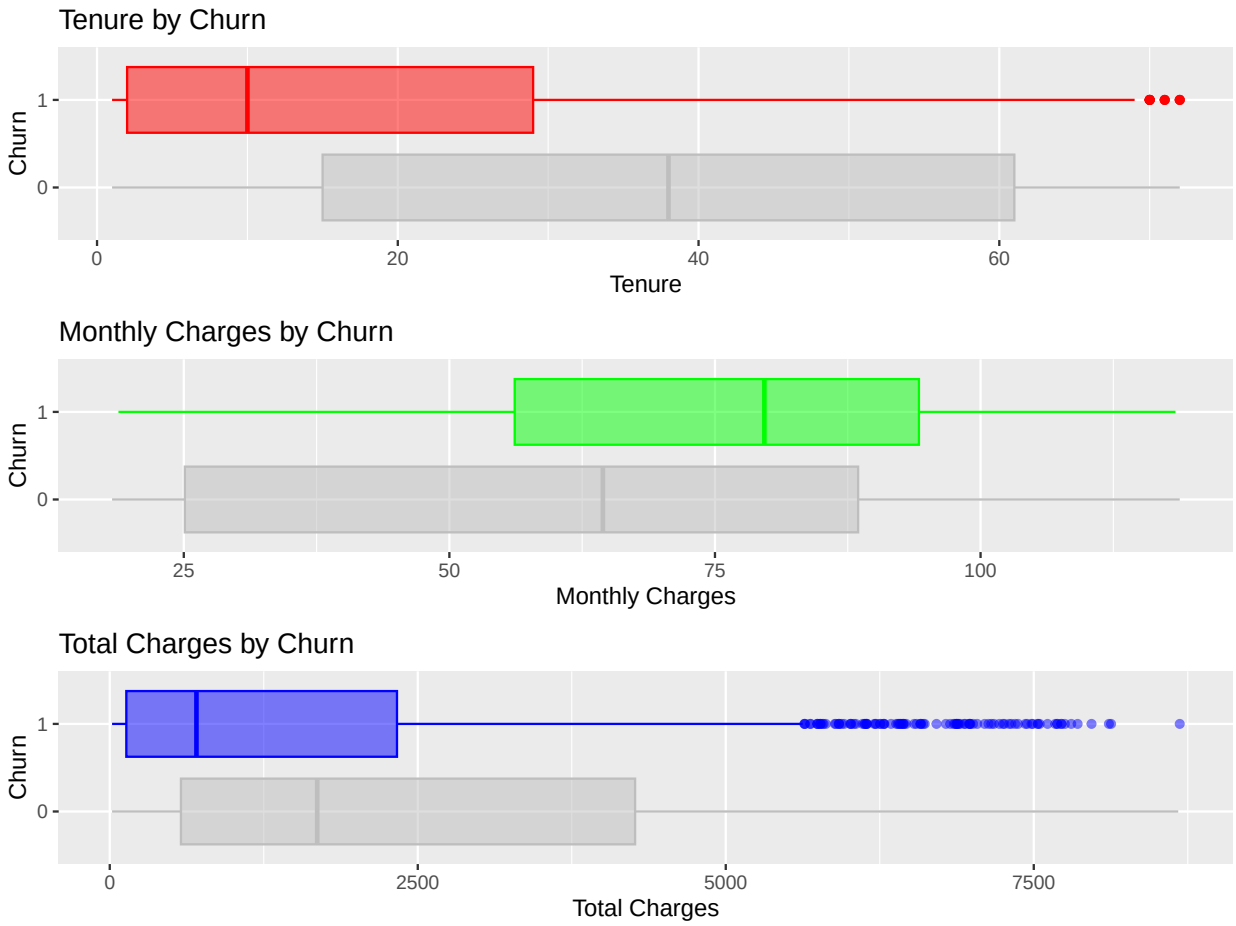


Figure 2: Boxploty zmiennych ciągłych z podziałem ze względu na Churn

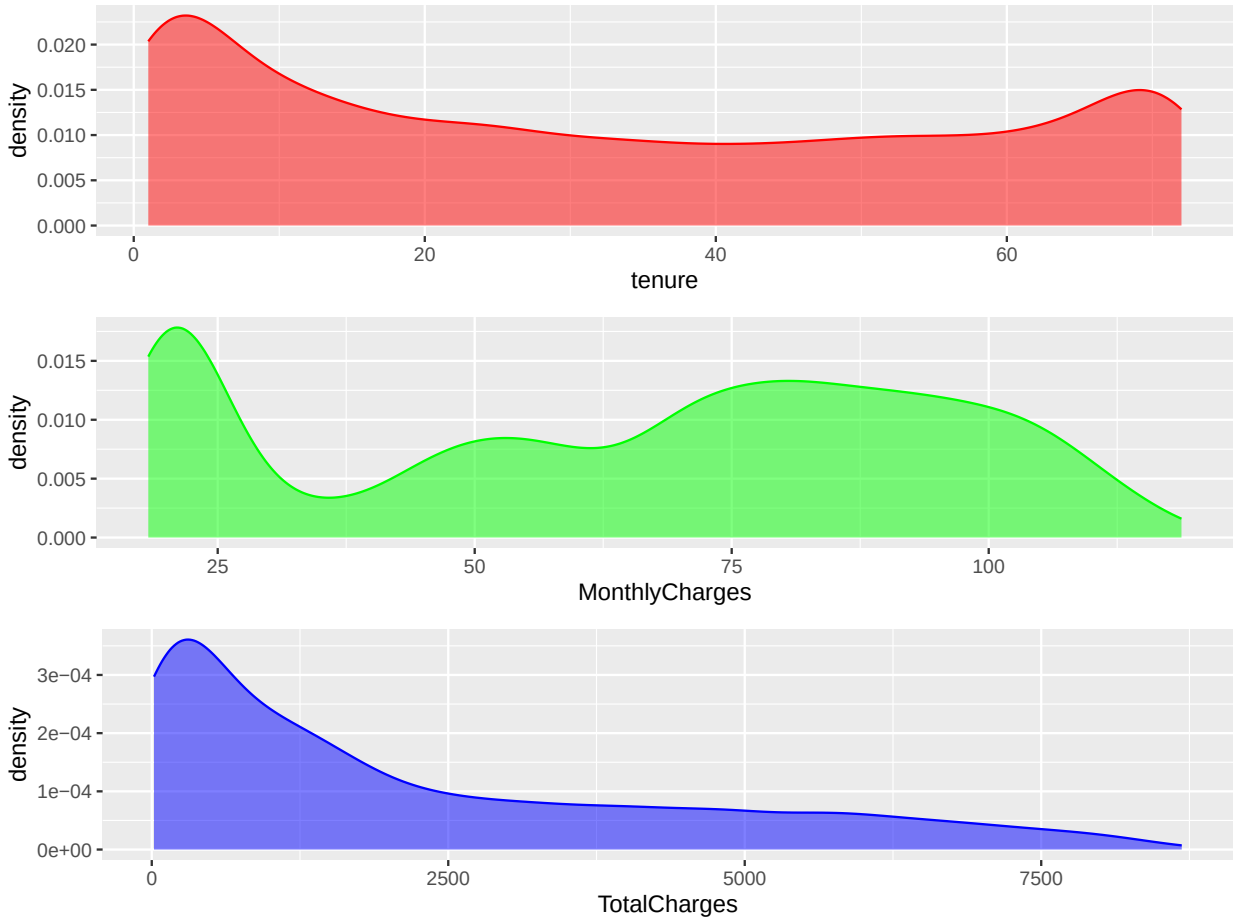


Figure 3: Estymator jądrowy gęstości

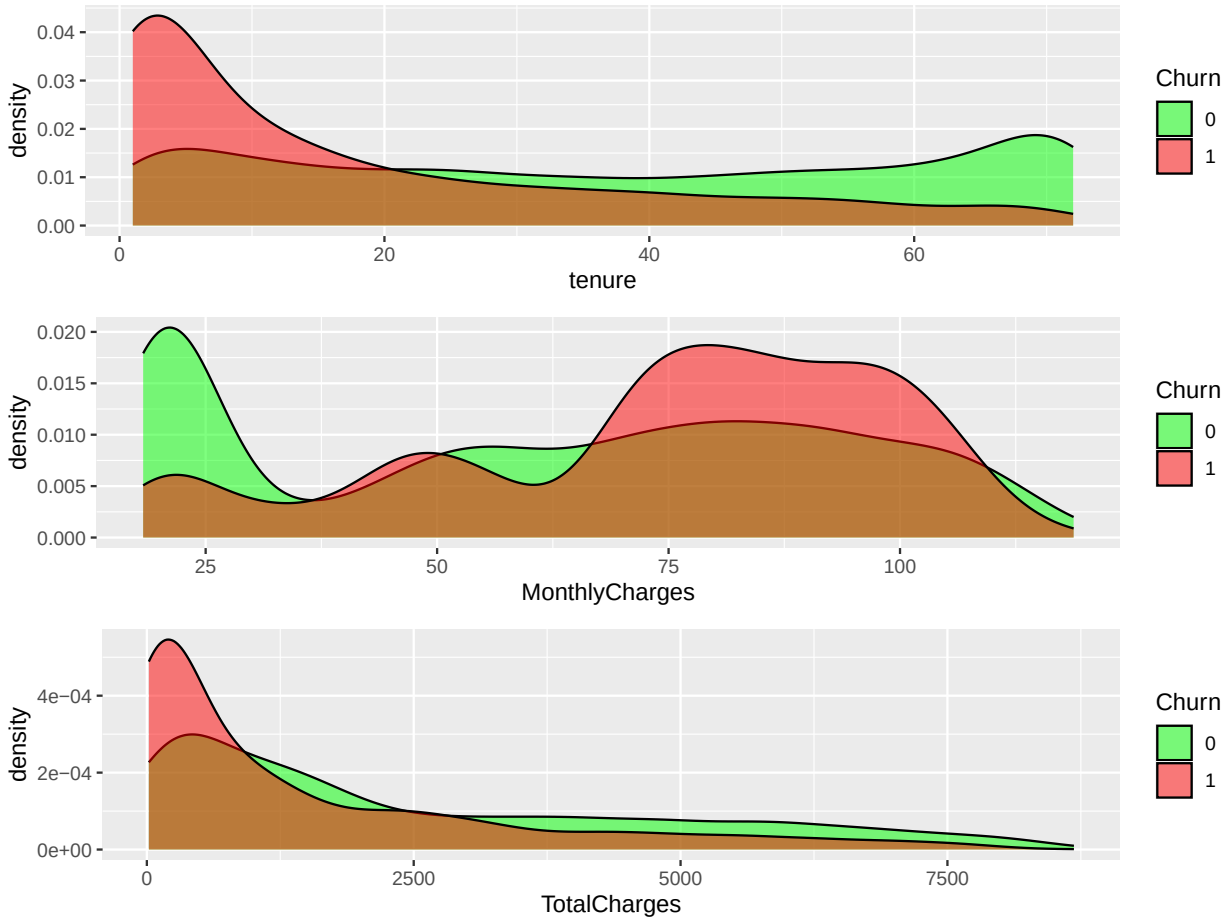


Figure 4: Estymator jądrowy gęstości z podziałem ze względu na Churn

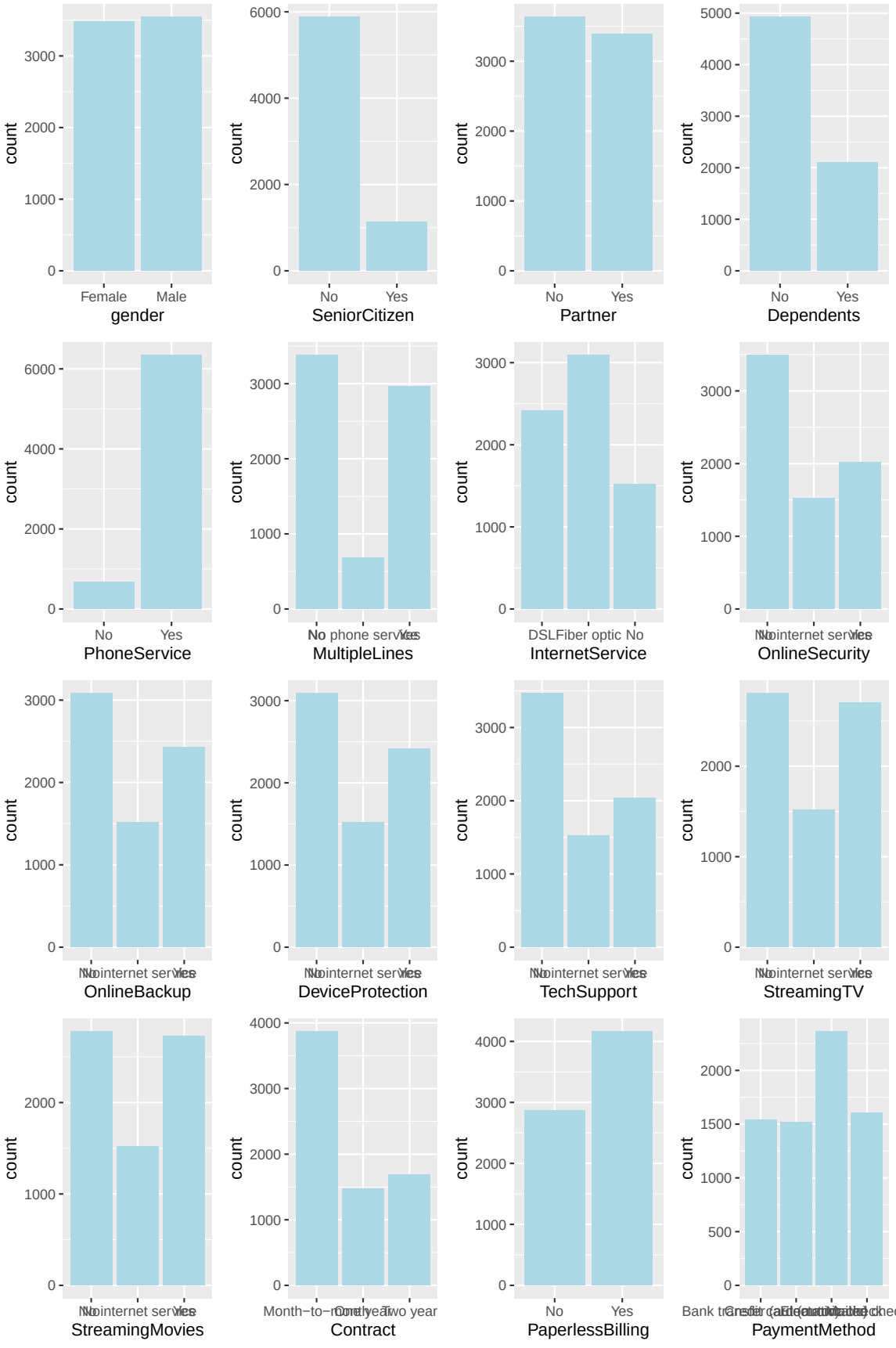


Figure 5: Wykres ilości obserwacji z podziałem na kategorie zmiennych

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 którejś ze 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/krótko 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

Najprawdopodbniej 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 należej do jednej z dwóch grup. Są albo nowymi klientami, albo klientami z dużym stażem. Ta druga grupa jest na wykresie pudełkowym interpertowana 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

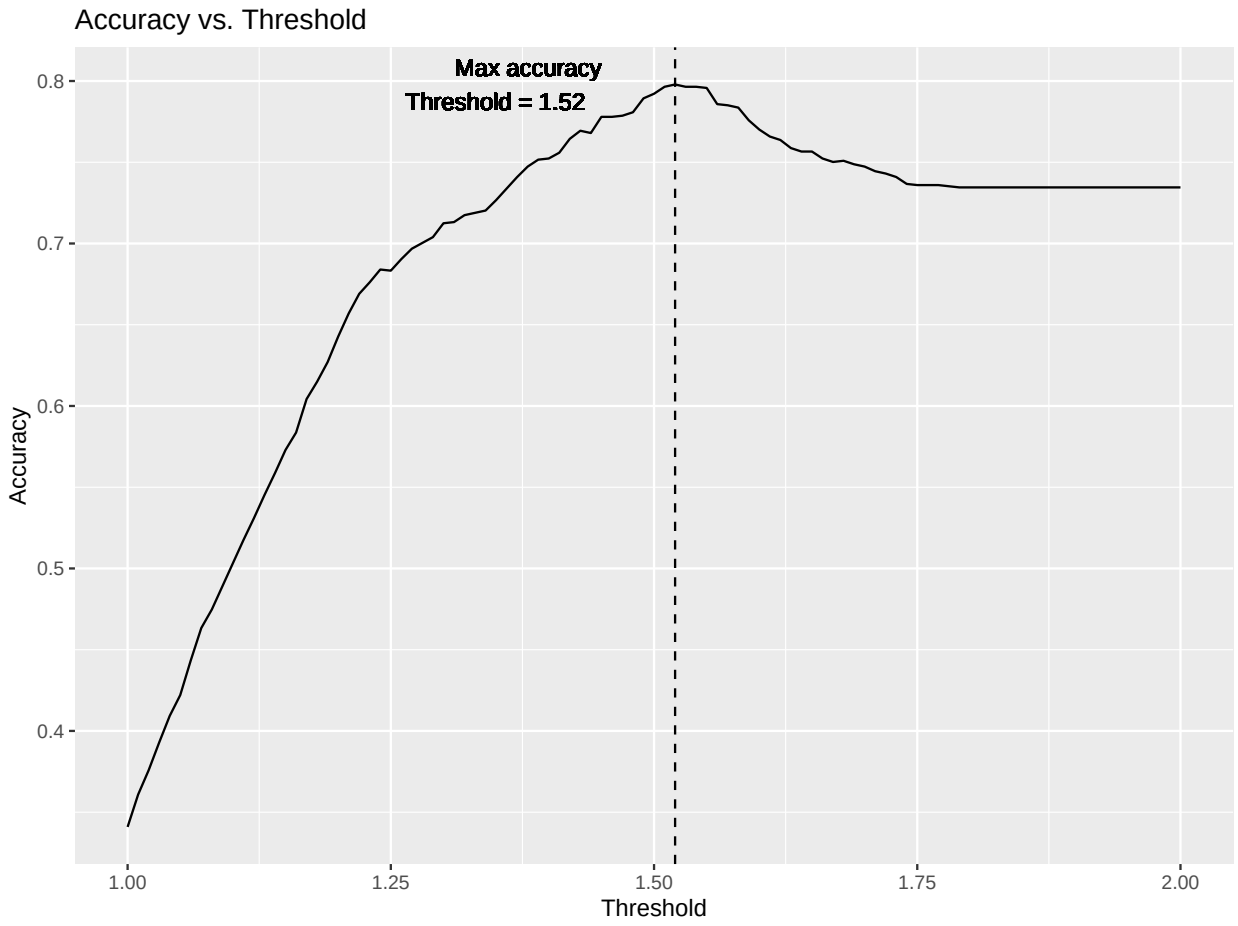


Figure 6: Skuteczność predykcji dla poszczególnych punktów odcięcia

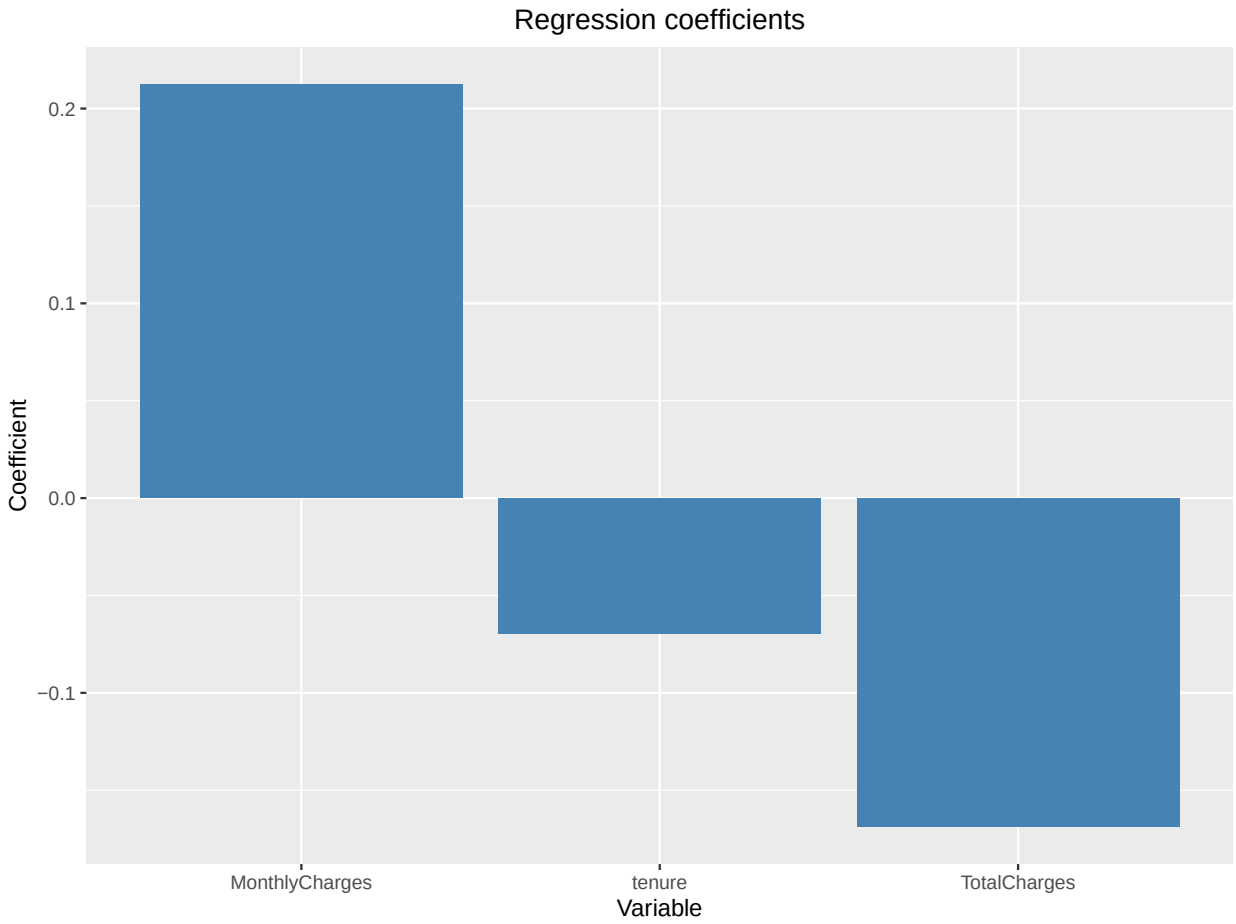


Figure 7: wartości współczynników w modelu regresji logistycznej

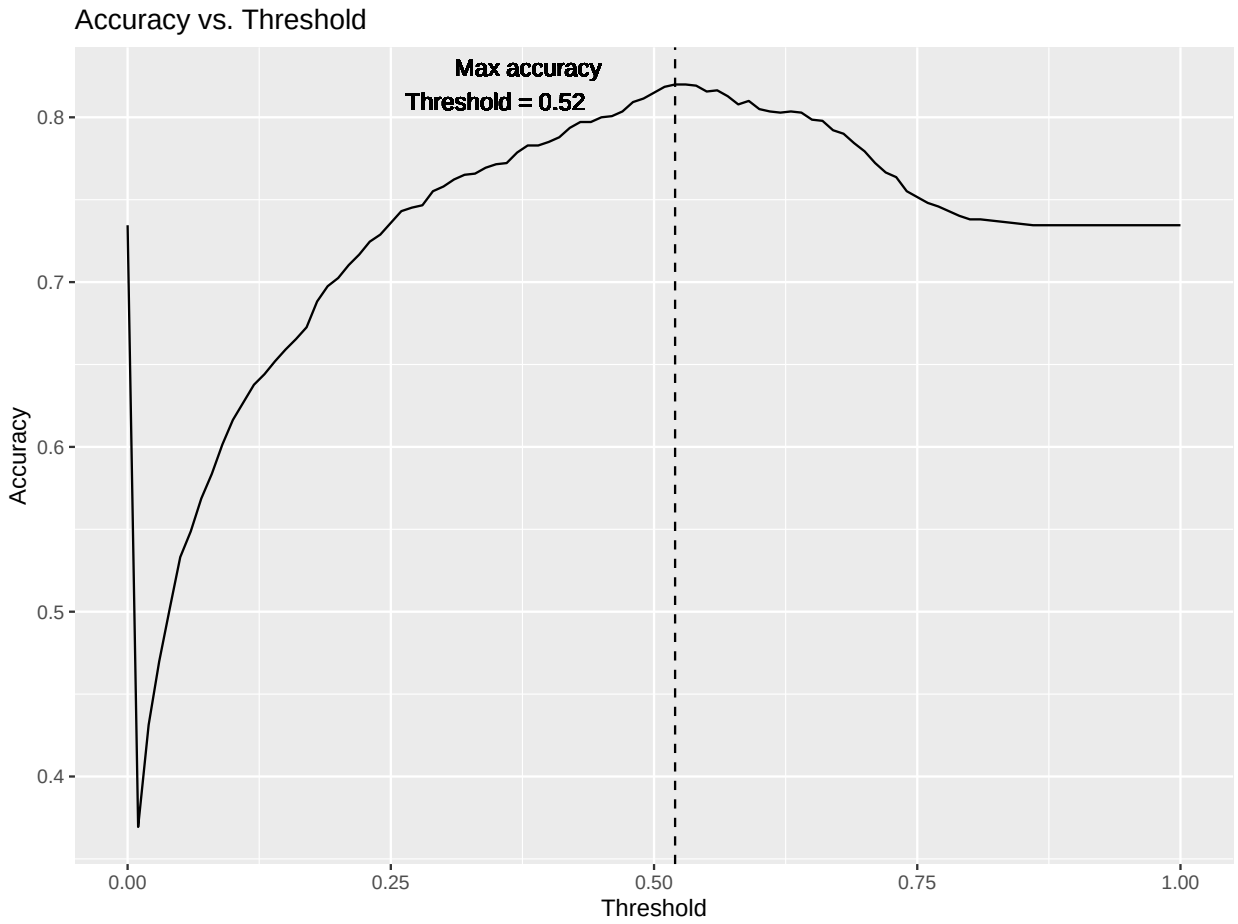


Figure 8: Skuteczność predykcji dla poszczególnych punktów odcięcia

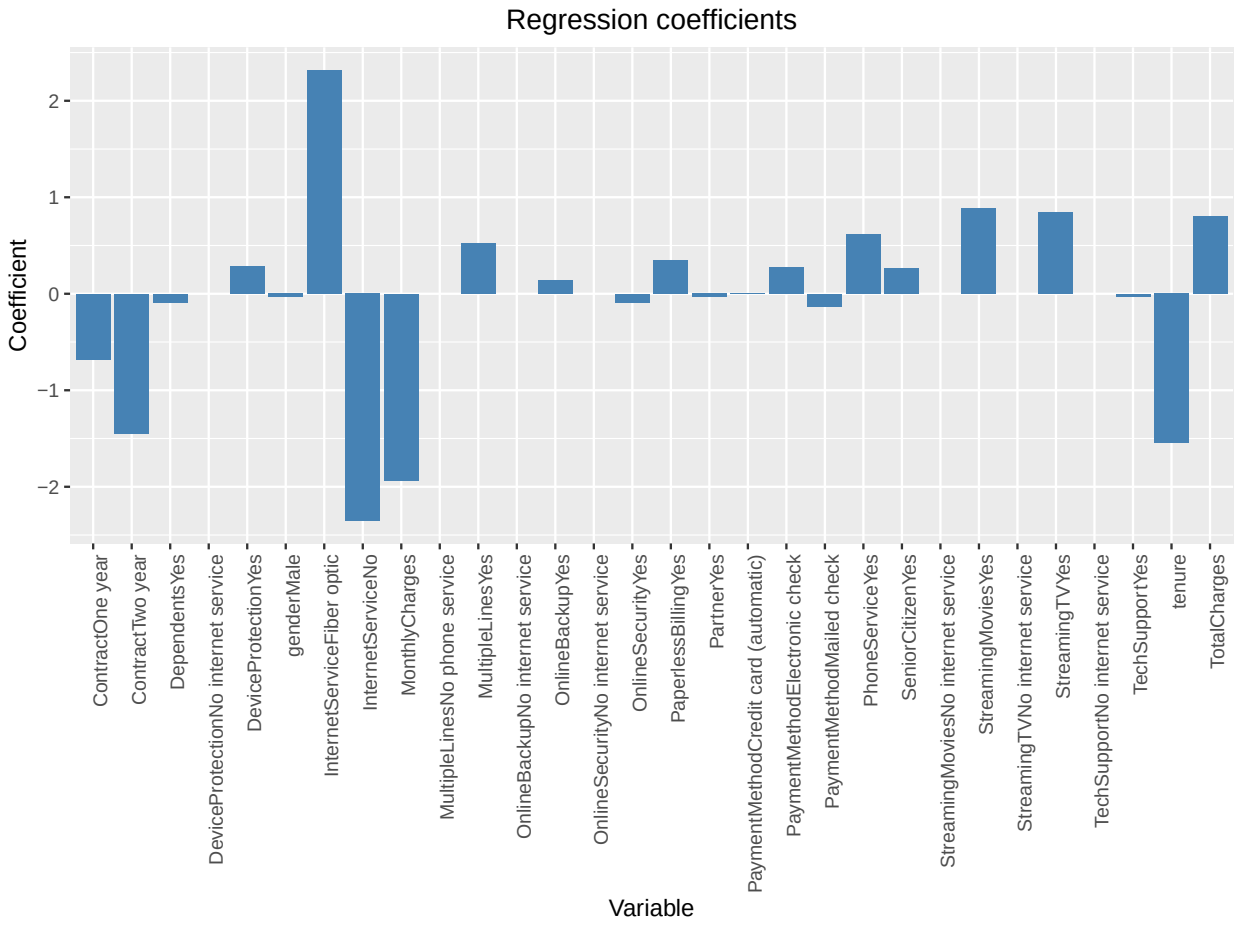


Figure 9: wartości współczynników w modelu regresji logistycznej

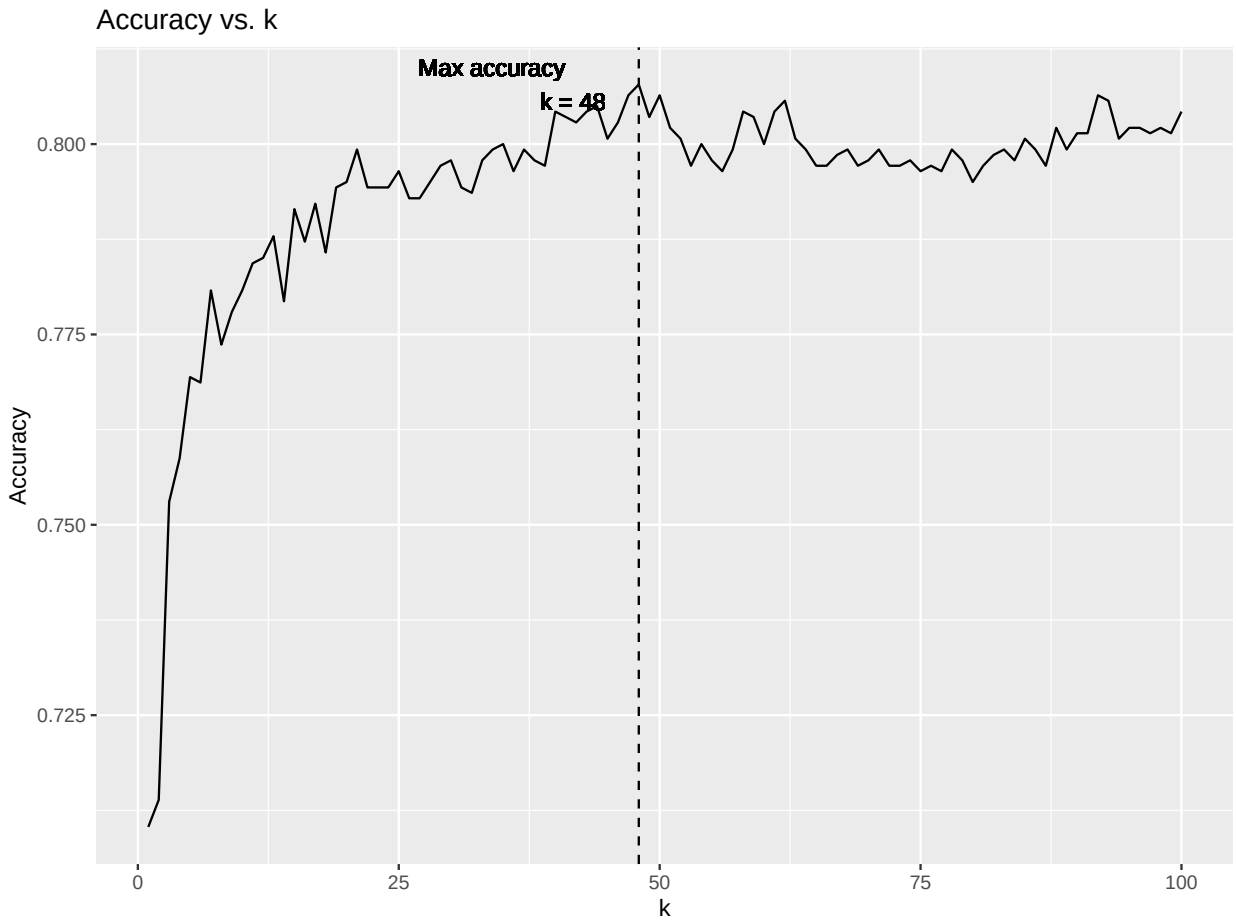


Figure 10: Skuteczność predykcji dla poszczególnych wartości k

	0	1
0	778	254
1	90	283

Table 3: Confusion matrix

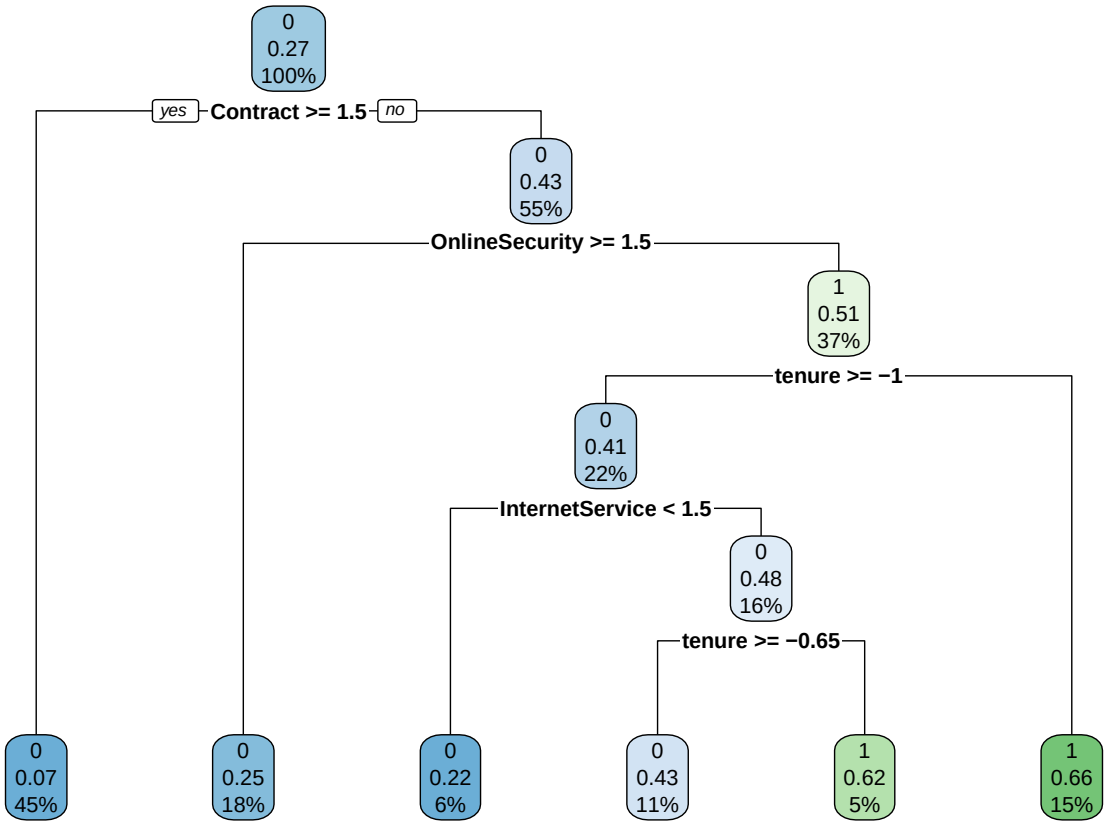
	0	1
0	917	115
1	158	215

Table 4: Confusion matrix for k = 48

Algorytm Naiwnego Bayesa

Algorytm k sąsiadów

Drzewo decyzyjne

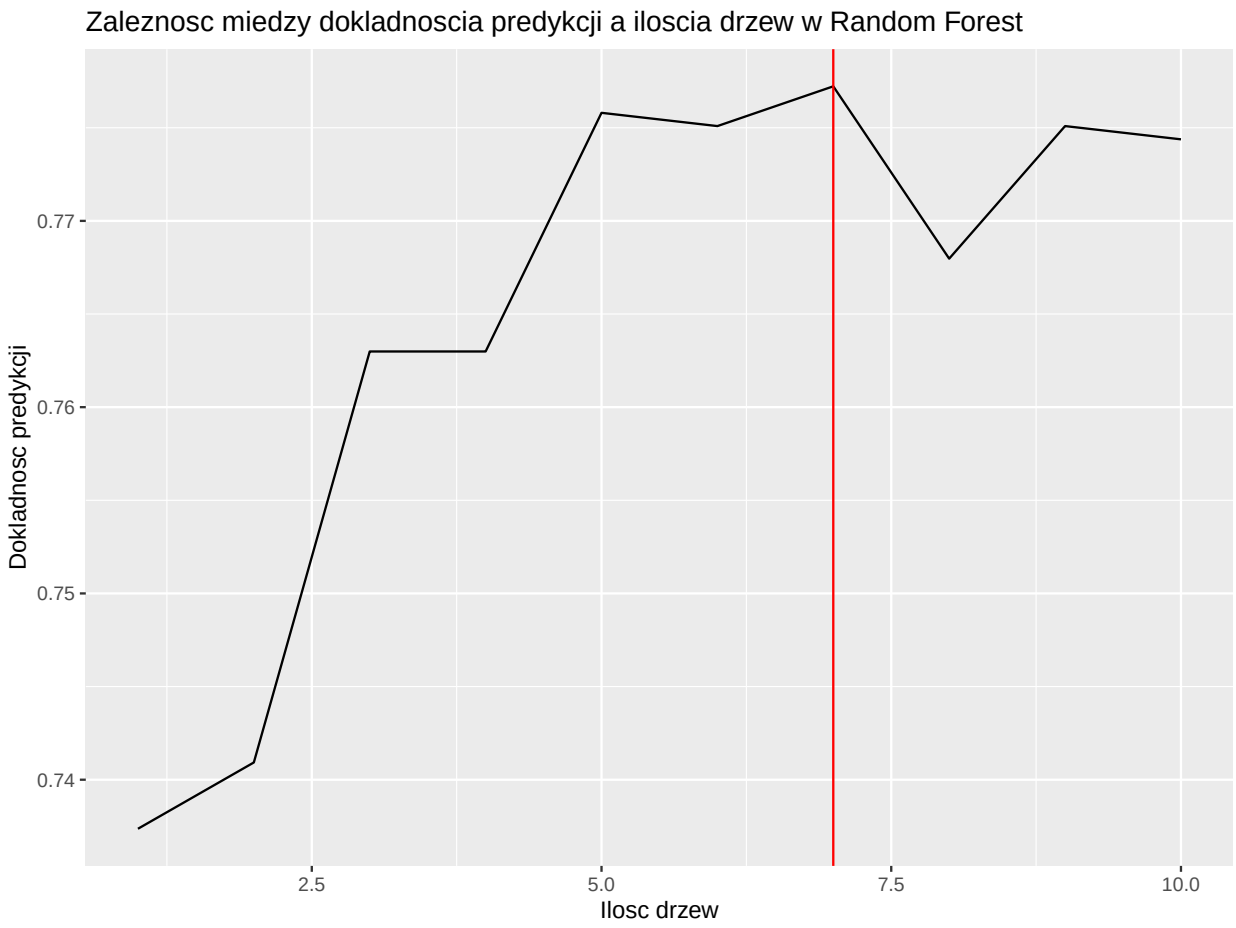


	0	1
0	934	98
1	201	172

Table 5: Confusion matrix

Random forest

	0	1
0	904	196
1	128	177



Boosting

[1] 0.8035587 [1] 0.6 Reference Prediction 0 1 0 922 166 1 110 207

Bagging

Accuracy: 0.7893238

	0	1
0	943	207
1	89	166