Analiza odchodzenia klientów

Bartosz Chądzyński & Michał Turek

2023-05-07

Wstęp

Preprocessing

Wykresy

Zacznijmy od analizy wykresów. Na początek zmienne ciągłe. Na wykresach 1 i 3 oraz w tabeli poniżej widzimy, że zmienne te są w znacząco różnych skalach, więc prawdopodobnie potrzebna będzie normalizacja. Zmienna *tenure*, a więc czas jaki dana osoba była/jest klientem, waha się od 1 do 72 miesięcy. Przy czym jej rozkład jest dwumodalny. Teoretycznie powinno się wydawać, że rozkład tej zmiennej powinien mieć charakter podobny do rozkładów z rodziny Gamma (np. rozkładu wykładniczego). W końcu każdy klient po pewnym czasie odchodzi, a więc w miarę upływu czasu klientów ubywa. Być może jednak jakieś procesy rynkowe spowodowały, że mamy liczniejszą grupę klientów ze stażem ok. 70 miesięcy (np. 70 miesięcy temu dana firma proponowała bardzo korzystne umowy). Kolejną zmienną jest *MonthlyCharges*. Jej estymowany rozkład jest bardzo nieregularny. Ma kilka maksimów lokalnych. Występują one w okolicach okrągłych liczb, takich jak 50 czy 80. Zapewne są związane z jakimiś limitami, które posiadają klienci, gdyż najczęściej nie przekraczają tych klejnych dziesiątek, albo mówiąc inaczej cyfra 9 pojawia się tu nadzwyczaj często jako cyfra jedności. Na koniec zostaje jeszcze *TotalCharges*. Zmienna ta ma spodziwany rozkład, tzn. przypomina on swoim kształtem rozkład Gamma. Wartości tej zmiennej są dużo większe od pierwszych dwóch, dlatego przeprowadzimy normalizacje danych przed ich użyciem.

Na wykresie 2 i 4 widać, że każda ze zmiennych ma istotnie różny rozkład, gdy pogrupujemy ją ze względu na Churn. Najbardziej wyróżnia się *tenure*, gdzie widać że odchodzili głównie nowi klienci. Podobnie odchodzili głównie klienci, których łączne opłaty były stosunkowo niskie, ale wynika to z korelacji zmiennej *TotalCharges* ze zmienną *tenure* (opiszemy to później). Natomiast jeśli chodzi o *MonthlyCharges*, to klienci, którzy odeszli, przeważają wśród tych co płacili większe miesięczne rachunki, co nie dziwi.

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

Na wykresach 5, 6, 7, 8 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*. Natomiast w znacznej więszkości proporcje klientów, którzy zostali i odeszli są podobne w każdej kategorii. Wyróżniają się tu osoby, które miały miesięczne kontrakty. To głównie one rezygnowały z usług operatora. W przypadku umów długoterminowych takie sytuacje zdarzały się bardzo rzadko. Podobnie

wyróżnia się sposób płatności. Prawie połowa osób płacących za pomocą *electrionic check* odeszła. Oczywiście w innych metodach płatności te liczby nie były aż tak duże. Można też zauważyć, że wśród straconych klientów jest bardzo mało osób, które nie korzystał z usług internetowych. s

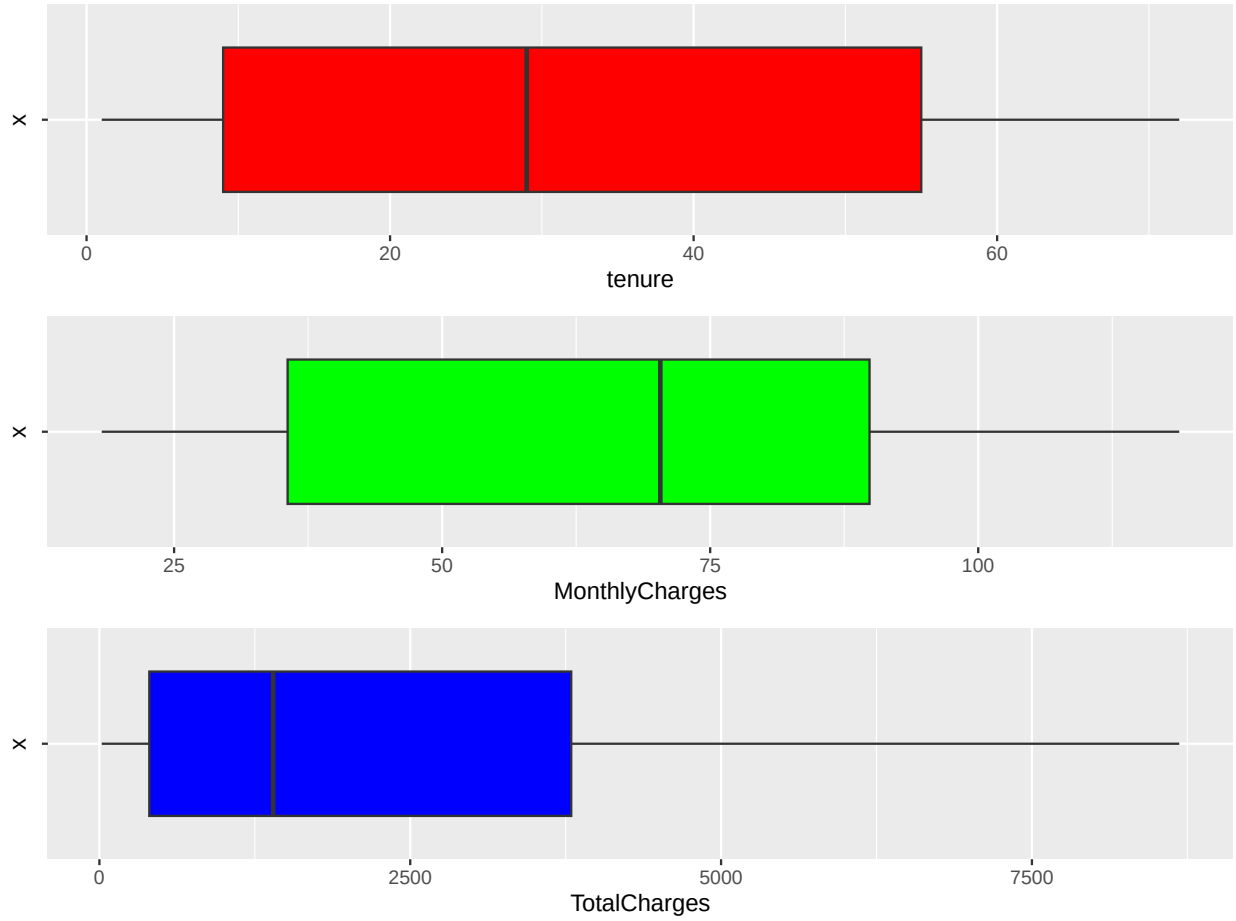


Figure 1: Boxploty zmiennych ciągłych

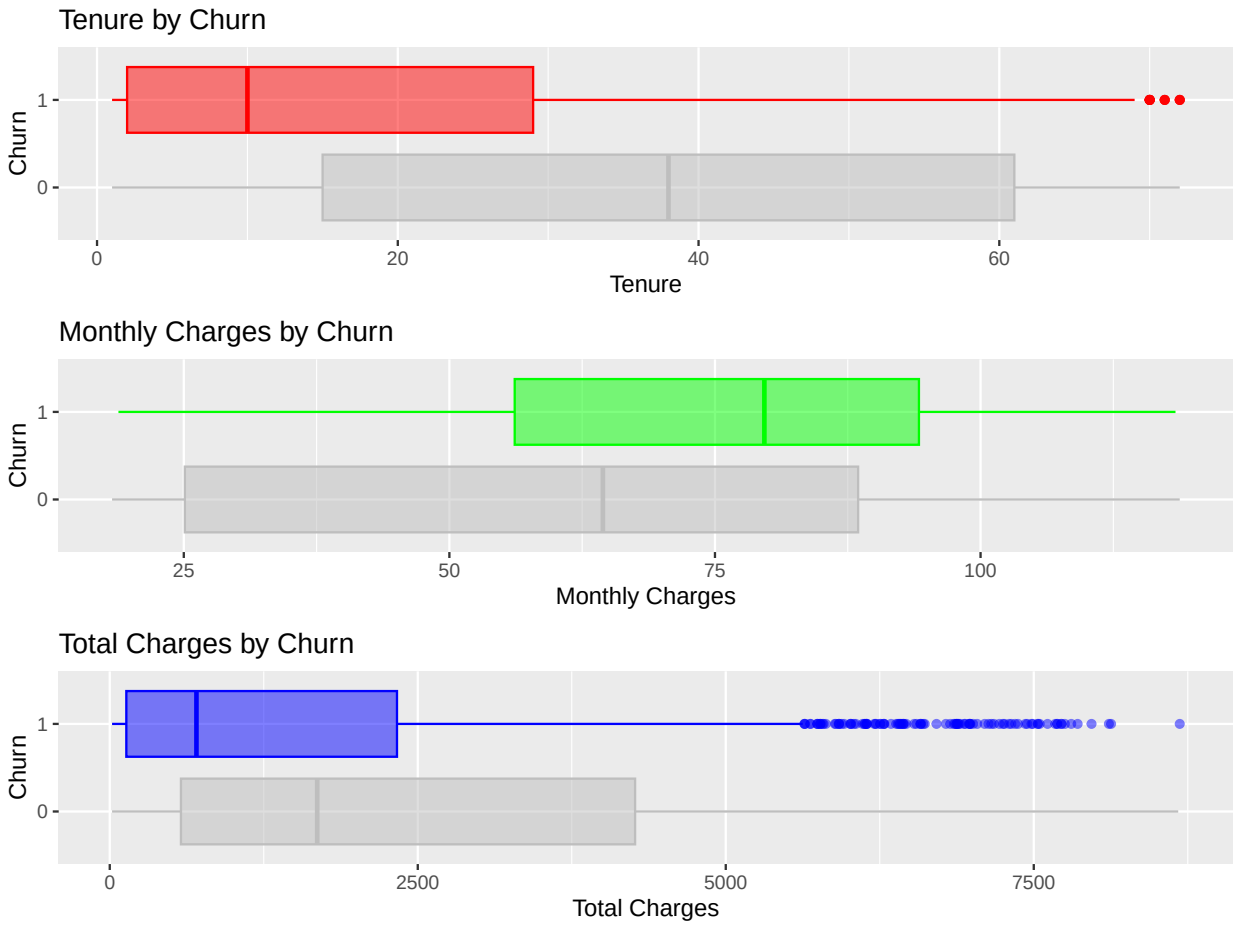


Figure 2: Boxploty zmiennych ciągłych z podziałem ze względu na Churn

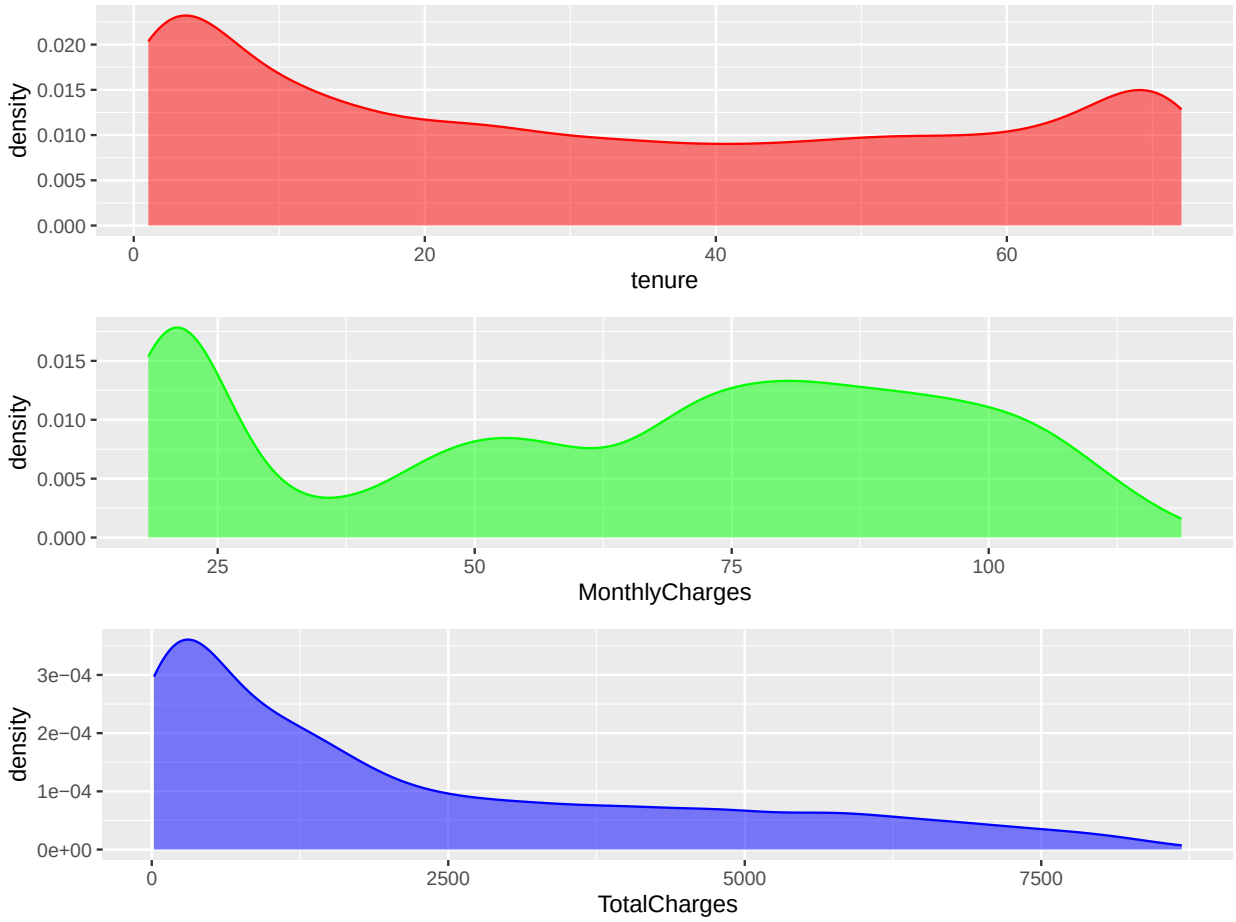


Figure 3: Estymator jądrowy gęstości

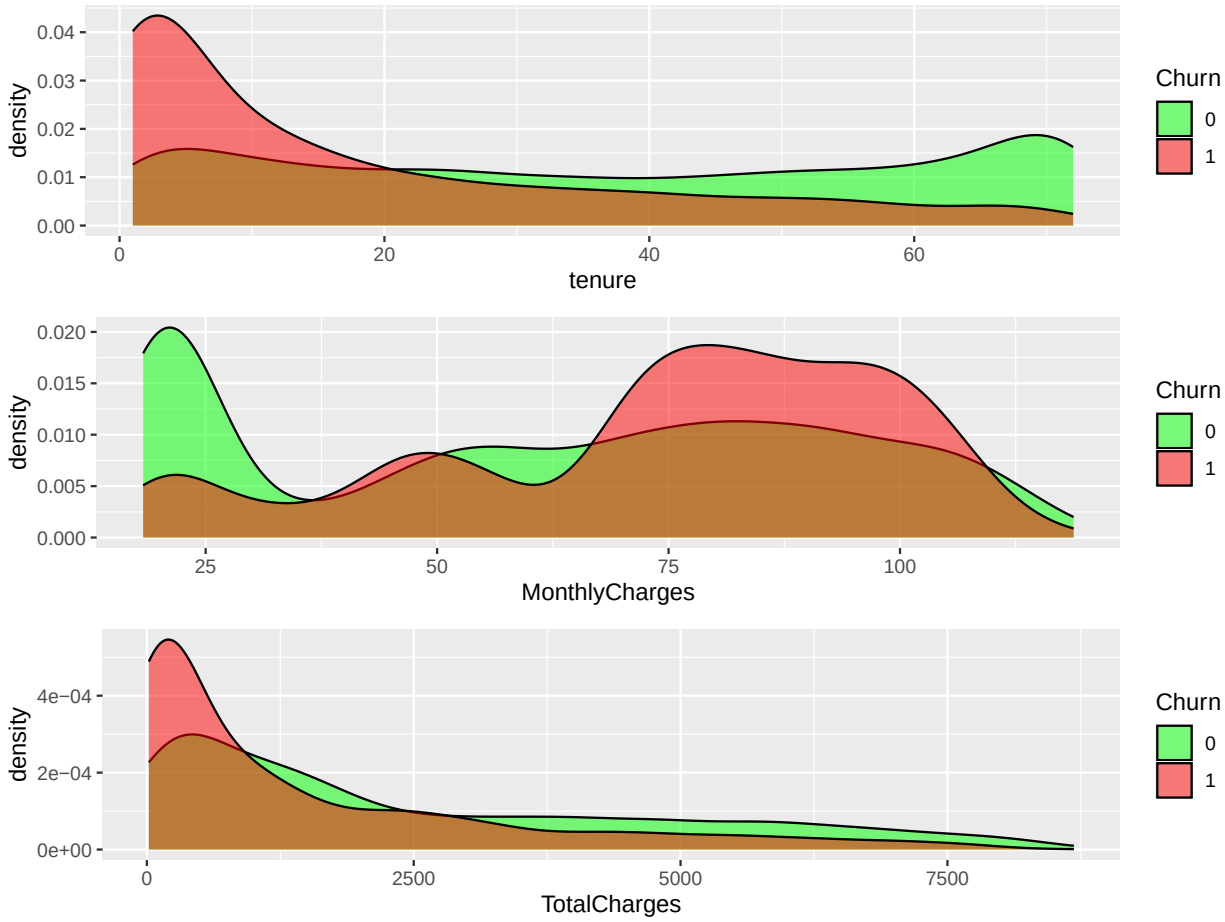


Figure 4: Estymator jądrowy gęstości z podziałem ze względu na Churn

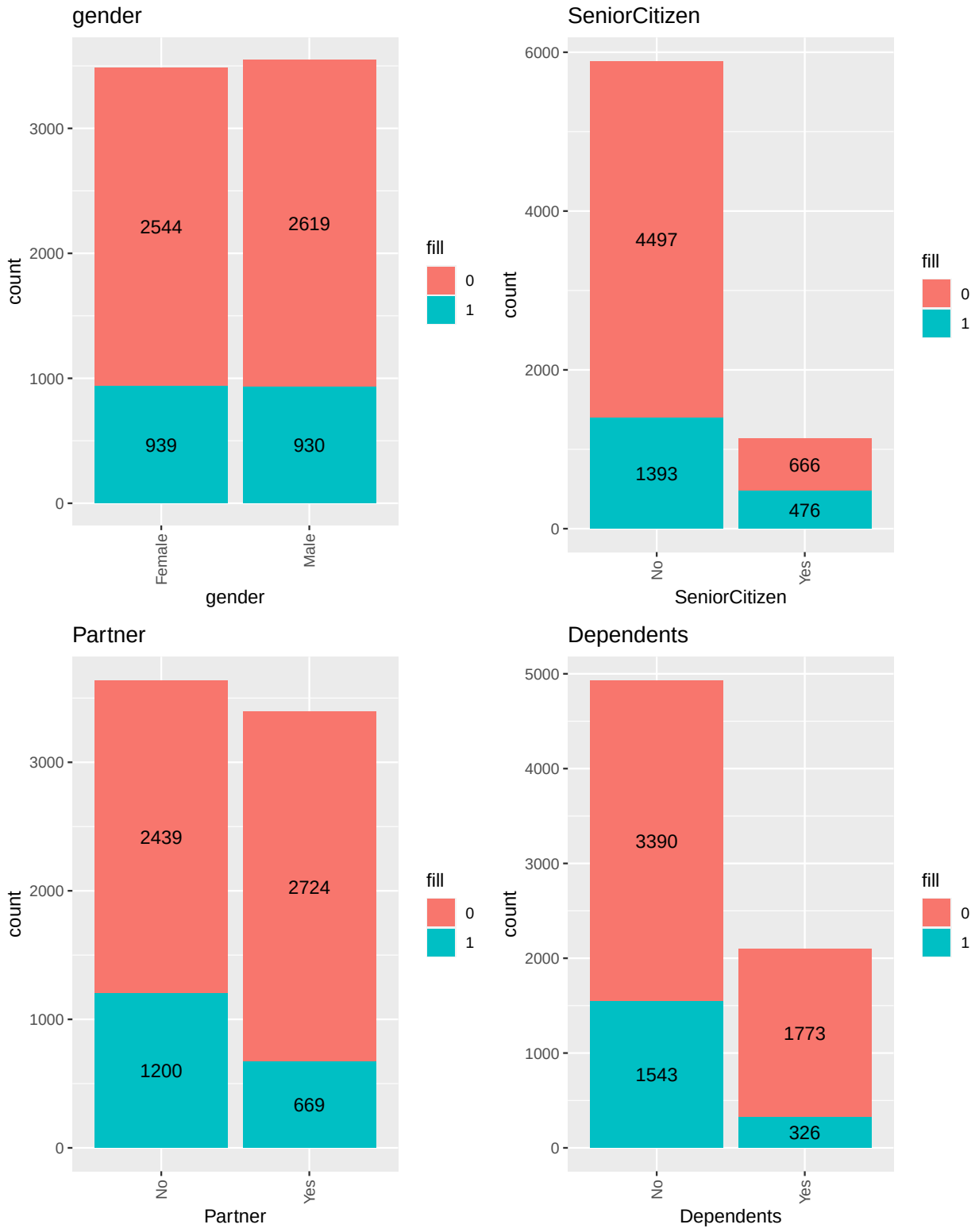


Figure 5: Wykres ilości obserwacji z podziałem na kategorie zmiennych

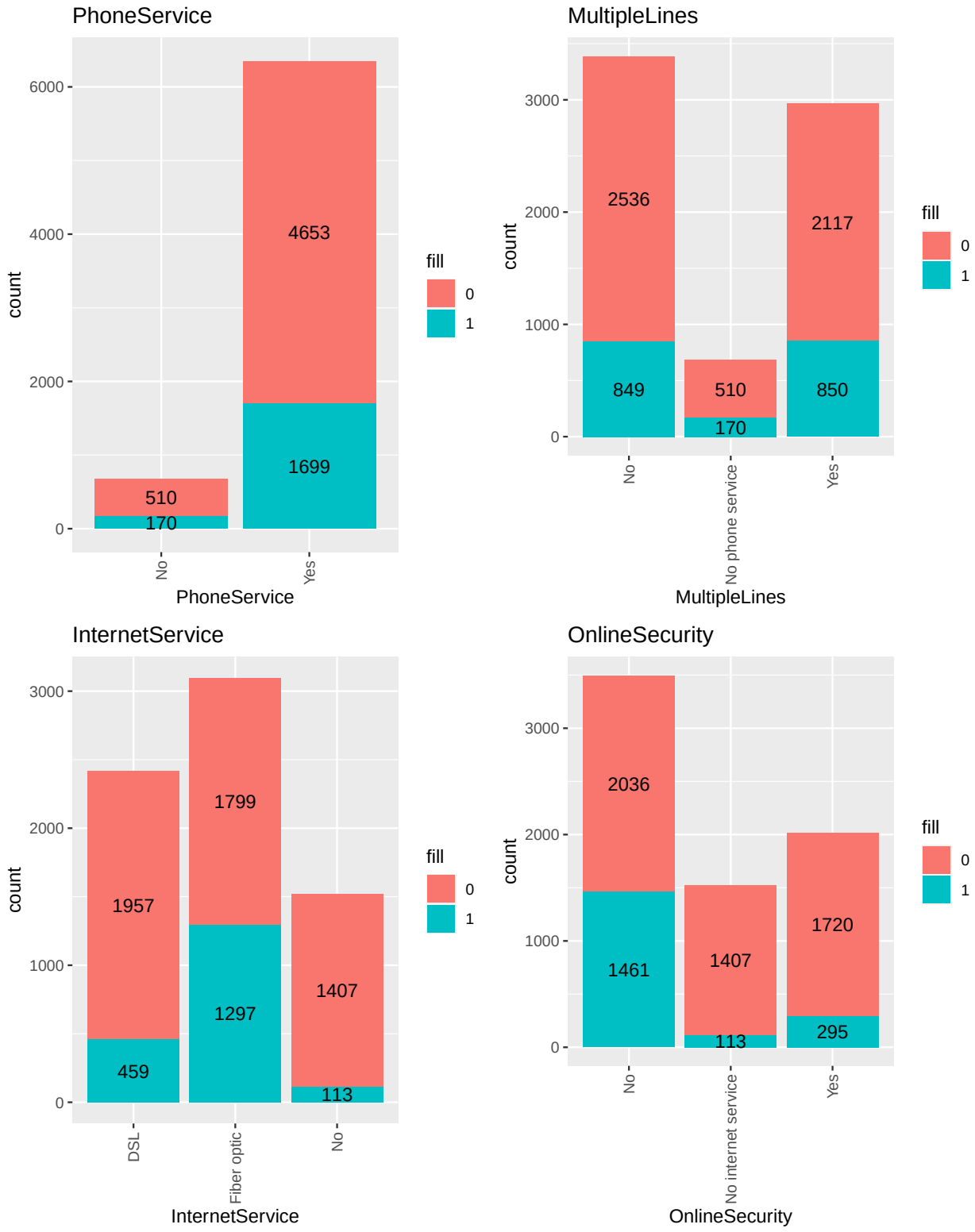


Figure 6: Wykres ilości obserwacji z podziałem na kategorie zmiennych

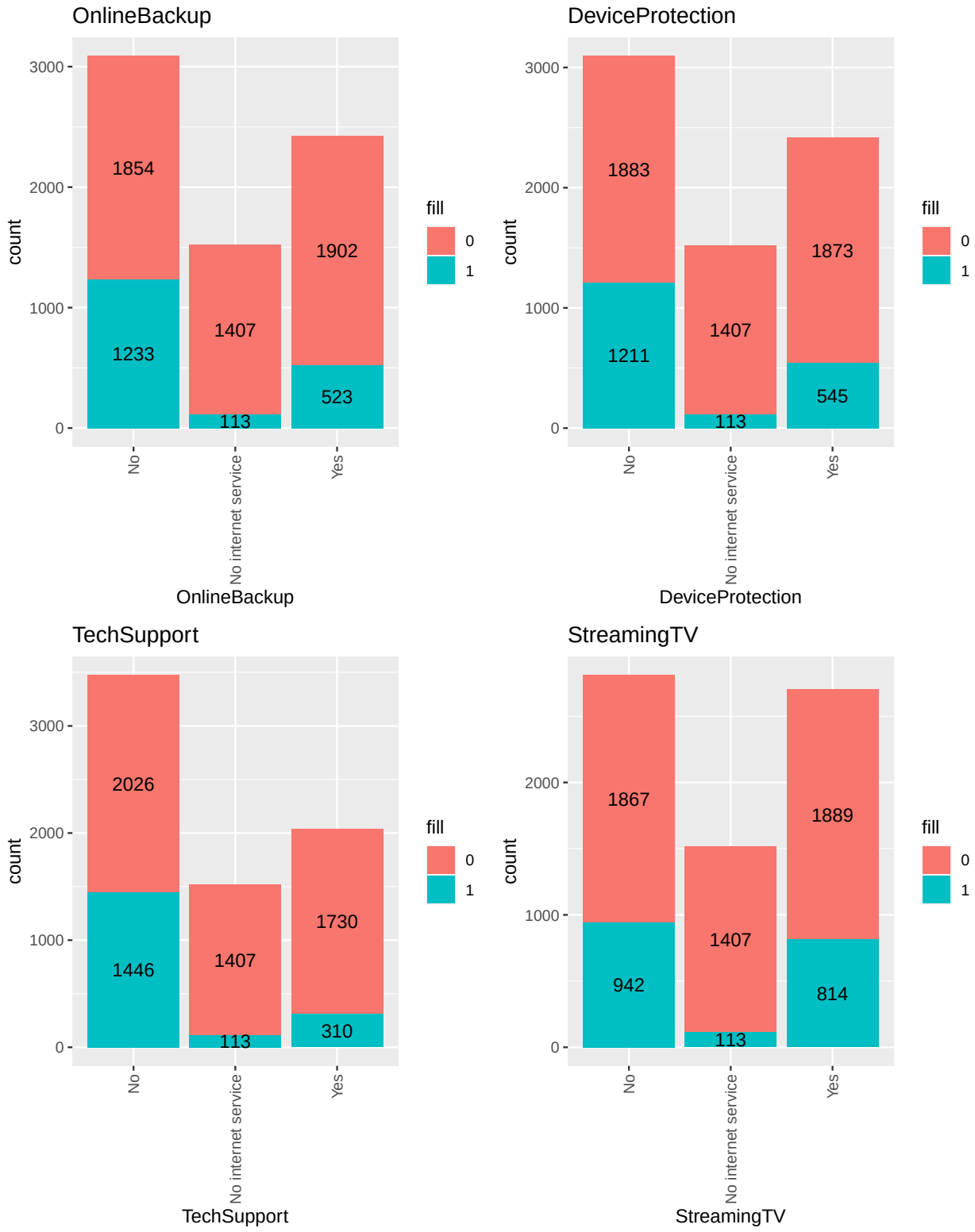


Figure 7: Wykres ilości obserwacji z podziałem na kategorie zmiennych

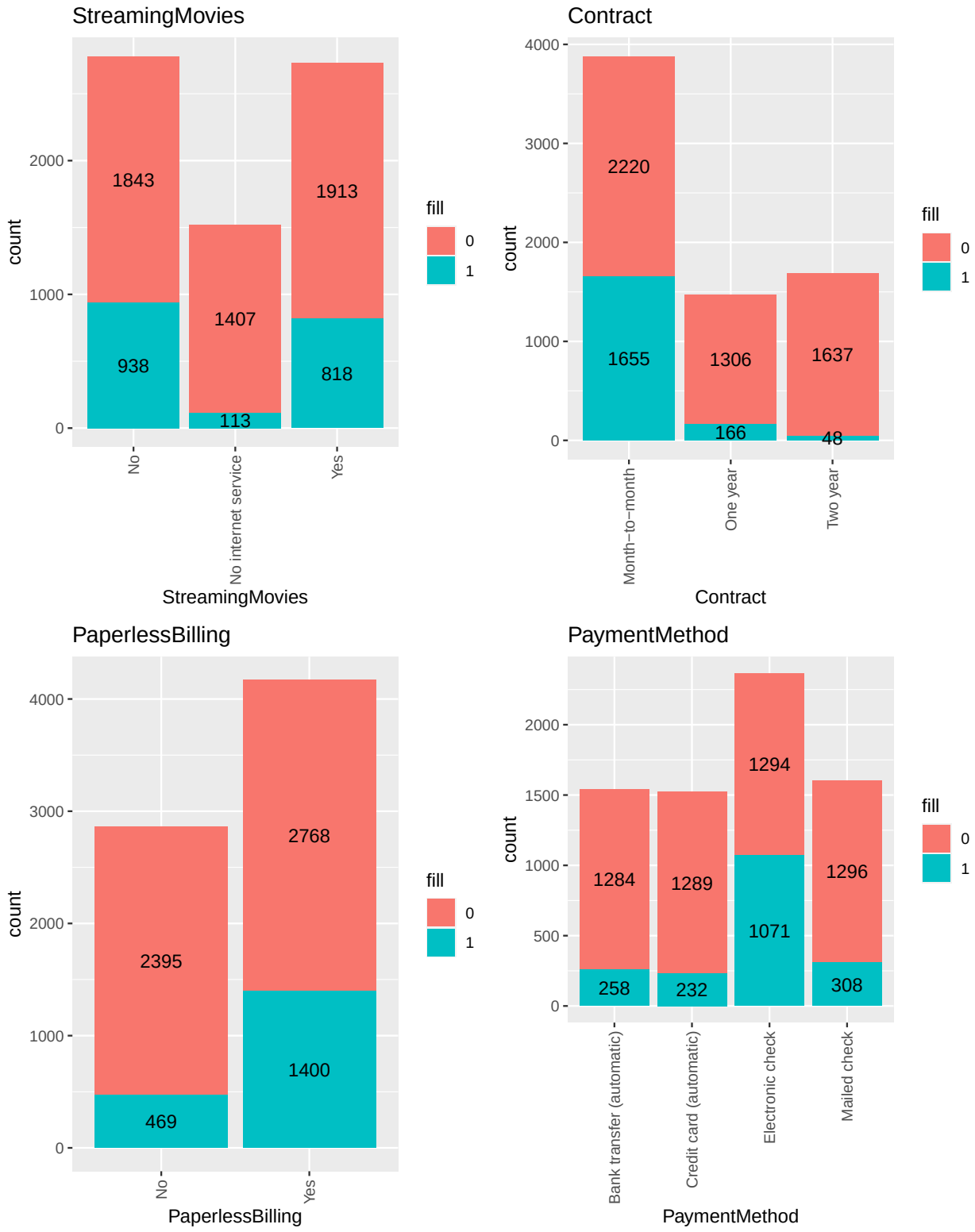


Figure 8: Wykres ilości obserwacji z podziałem na kategorie zmiennych

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 wydaje się odpowiednia ilościowo (nie za mała i nie za duża). W analizie dokonujemy losowego podziału na zbiór treningowy i testowy.

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

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. Zastosowaliśmy model regresji liniowej, który bierze pod uwagę 3 zmienne ciągłe, jako zmienne objaśniające i *Churn*, jako zmienną objaśnianą. Otrzymane w ten sposób wartości dzielimy na dwie grupy stosując punkt odcięcia na ustalonym poziomie. Na wykresie 9 widzimy skuteczność predykcji dla punktów odcięcia pomiędzy 1 i 2. Wybieramy ten z największą skutecznością i sprawdzamy jak wygląda macierz pomyłek dla niego (1). Jak widać osiągamy w ten sposób całkiem niezłą skuteczność na poziomie 0.7978648, co jest o ok. 0.05 więcej niż, gdybyśmy estymowali każdą obserwacje do liczniejszej klasy.

Zastanówmy się jeszcze jaki wpływ miały poszczególne zmienne w modelu. W tym celu przyjrzyjmy się współczynnikom w modelu (wykres 10). Nie ma tam stałej, ponieważ nie ma ona wpływu na model (i tak później ustalamy punkt odcięcia). Widać natomiast, że największy wpływ na to że ktoś jest sklasyfikowany z *Churn*=1, ma zmienna *MonthlyCharges*. Im większe miesięczne opłaty, tym większe prawdopodobieństwo że osoba zrezygnuje z umowy. Odwrotnie jest w przypadku łącznych opłat i stażu klienta, które to zmienne zwiększają prawdopodobieństwo aby obserwacja była zakwalifikowana *Churn*=0.

	0	1
0	968	220
1	64	153

Table 1: Confusion matrix at threshold = 1.52

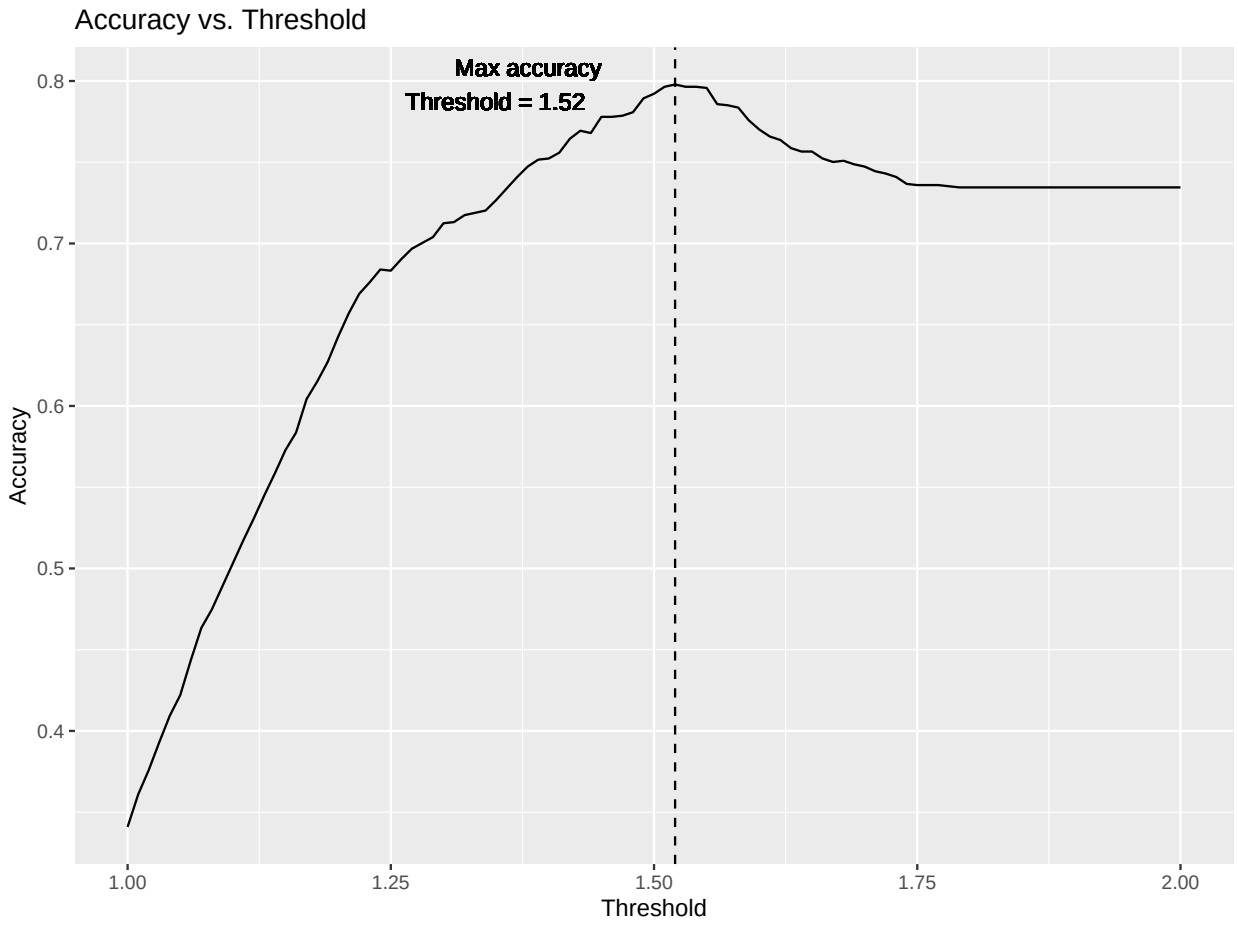


Figure 9: Skuteczność predykcji dla poszczególnych punktów odcięcia

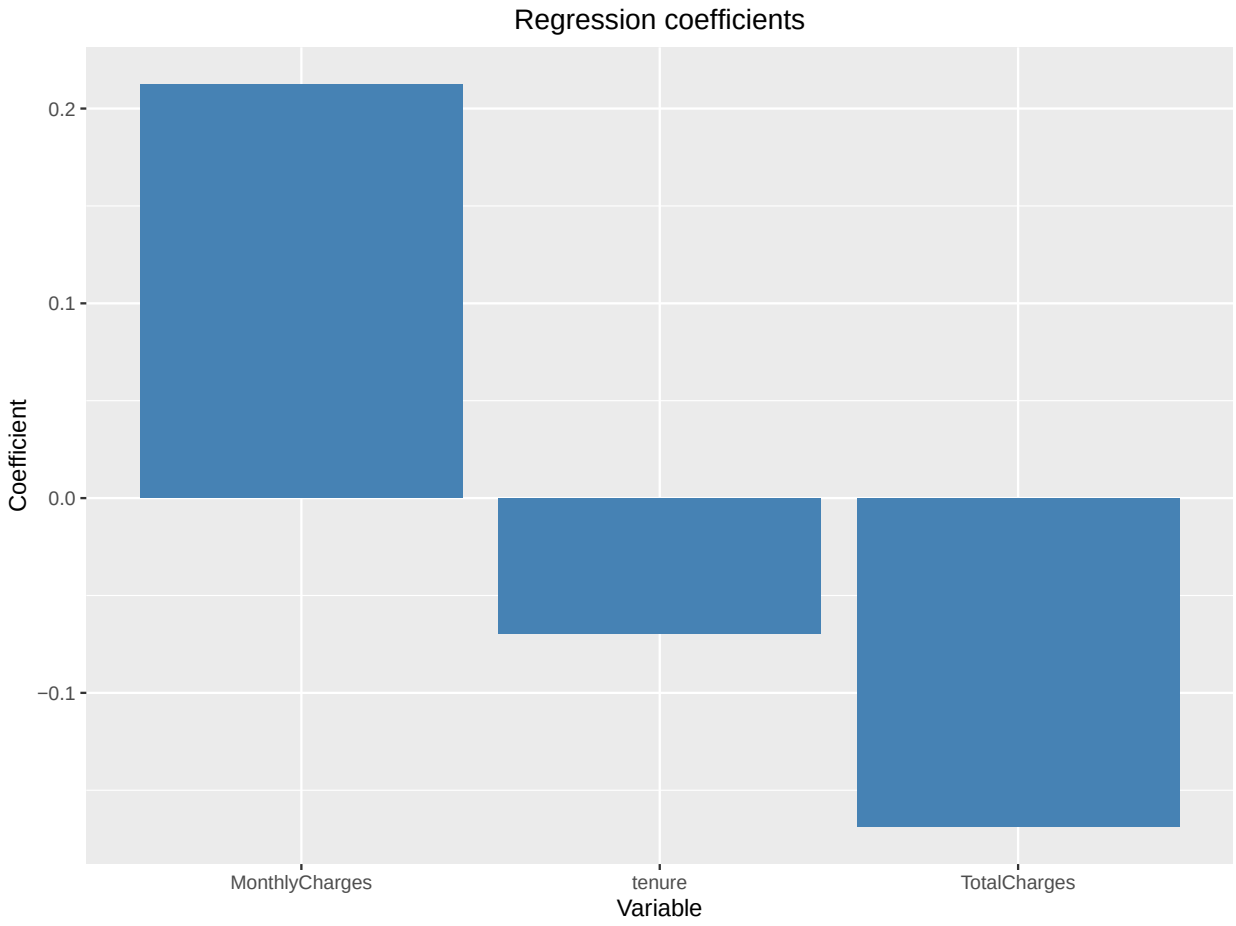


Figure 10: wartości współczynników w modelu regresji logistycznej

Regresja Logistyczna

Teraz model regresji logistycznej. Standardowo, wzięliśmy pod uwagę wszystkie zmienne i zbudowaliśmy model z domyślnymi parametrami. Punkt odcięcia wybraliśmy testując skuteczności przy różnych wartościach. Na wykresie 11 widzimy, że najskuteczniejszy był model z punktem odcięcia równym 0.52. Z tabeli 1 widzimy, że skuteczność wynosi niemal dokładnie 0.82. Można się jeszcze przyjrzeć współczynnikom modelu. Na ich podstawie widzimy, że największy wpływ oprócz *MonthlyCharges* i *tenure* mają zmienne *InternetServiceFiber* i *InternetService*.

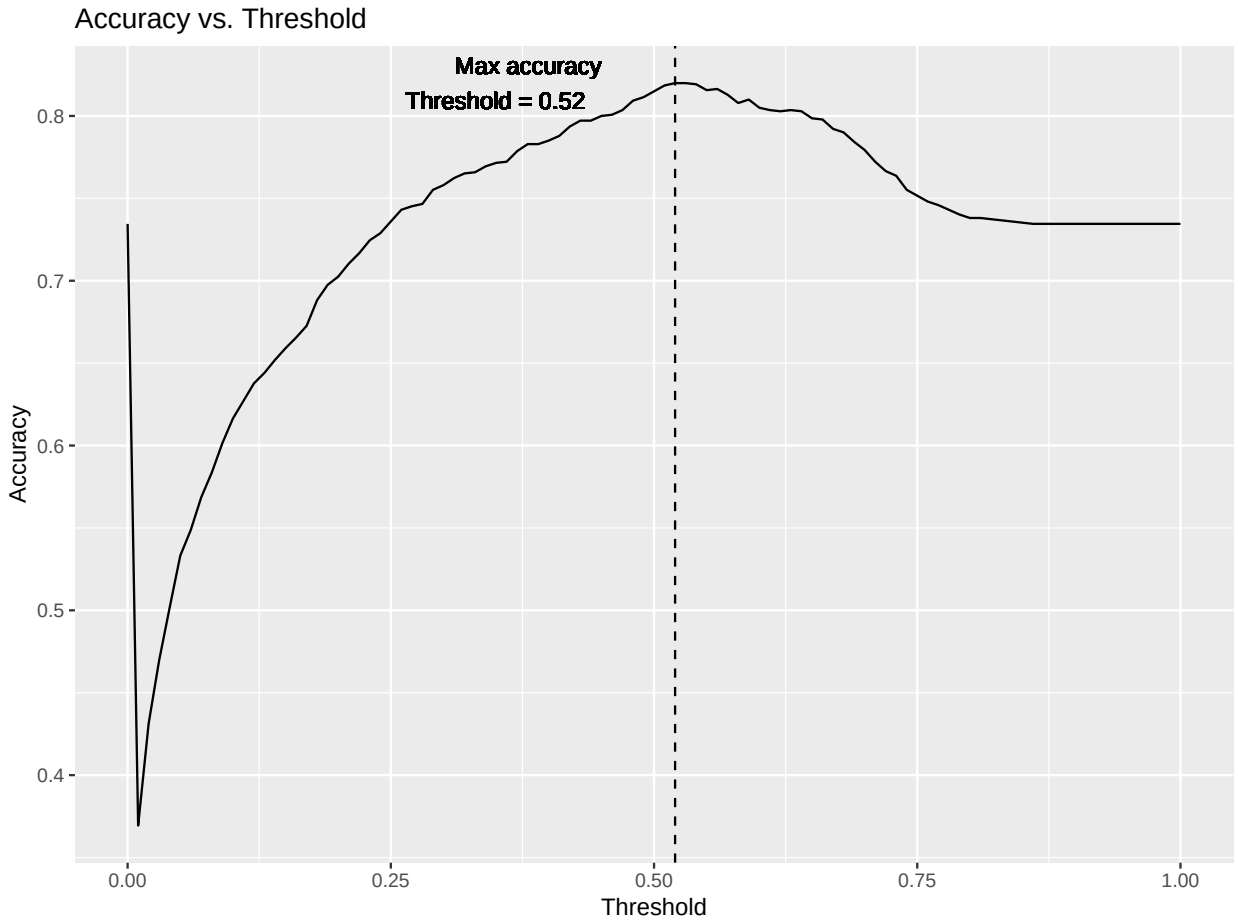


Figure 11: Skuteczność predykcji dla poszczególnych punktów odcięcia

	0	1
0	942	163
1	90	210

Table 2: Confusion matrix at threshold = 0.52

Algorytm Naiwnego Bayesa

Zastosowaliśmy również Naive Bayes Algorithm z domyślnymi parametrami (funkcja *NaiveBayes* z pakietu *klaR*). Nie dał on jednak zbyt dobrych rezultatów. W tabeli 3 widzimy, że skutecznosc tego modelu wyniosła zaledwie 0.76.

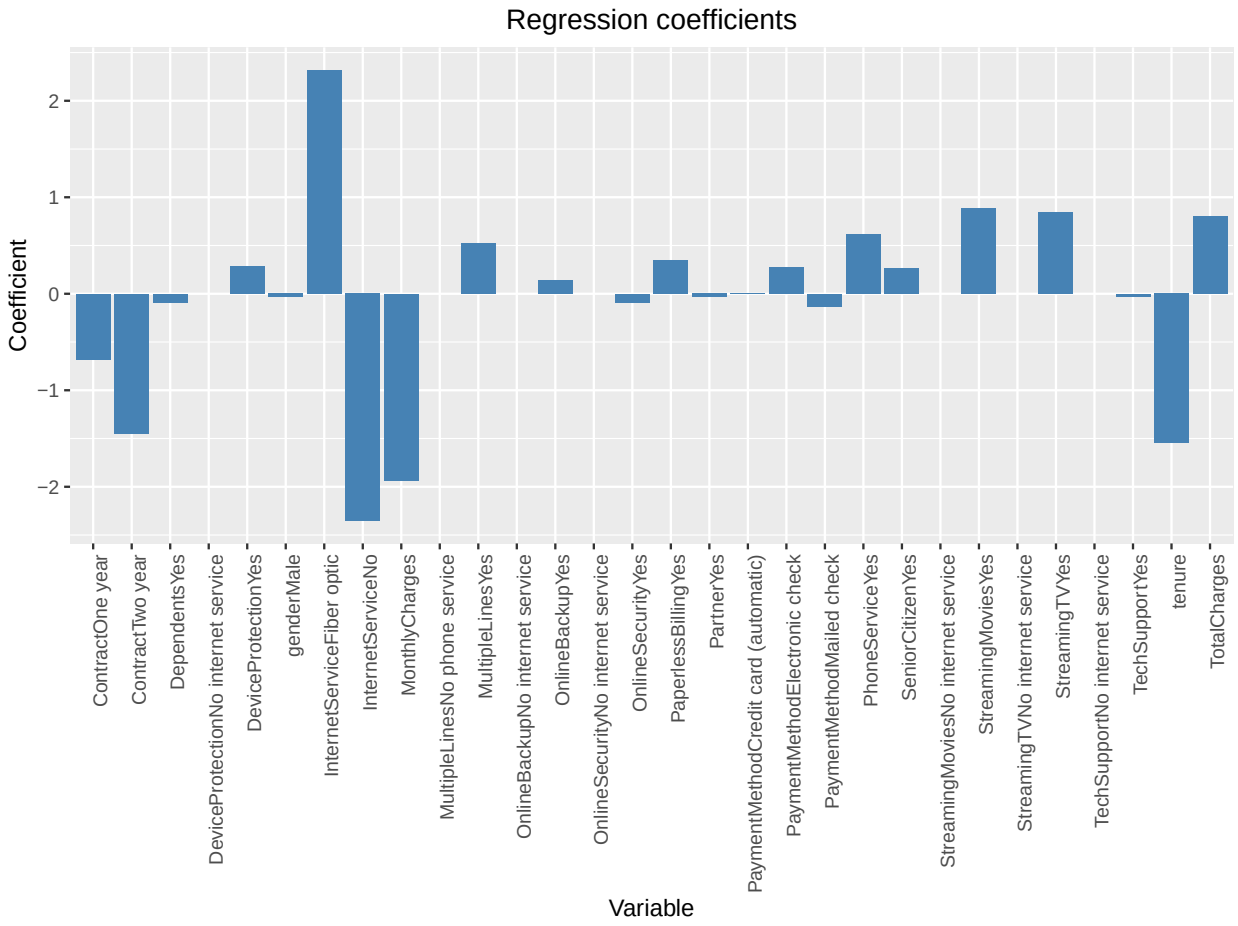


Figure 12: wartości współczynników w modelu regresji logistycznej

	0	1
0	778	254
1	90	283

Table 3: Confusion matrix

Algorytm k sąsiadów

W tym przypadku użyliśmy funkcji *knn* z pakietu *class*. Przeprowadziliśmy symulacje dla wszystkich wartości k od 1 do 100. Najlepszy model powstał dla k równego 48 (wykres 13). Jego skuteczność wyniosła 0.806. W tabeli 4 widzimy, że po raz pierwszy mamy sytuacje kiedy obiekty, które w rzeczywistości mają *Churn*=1 są cześciej błędnie klasyfikowane niż obiekty z *Churn*=0.

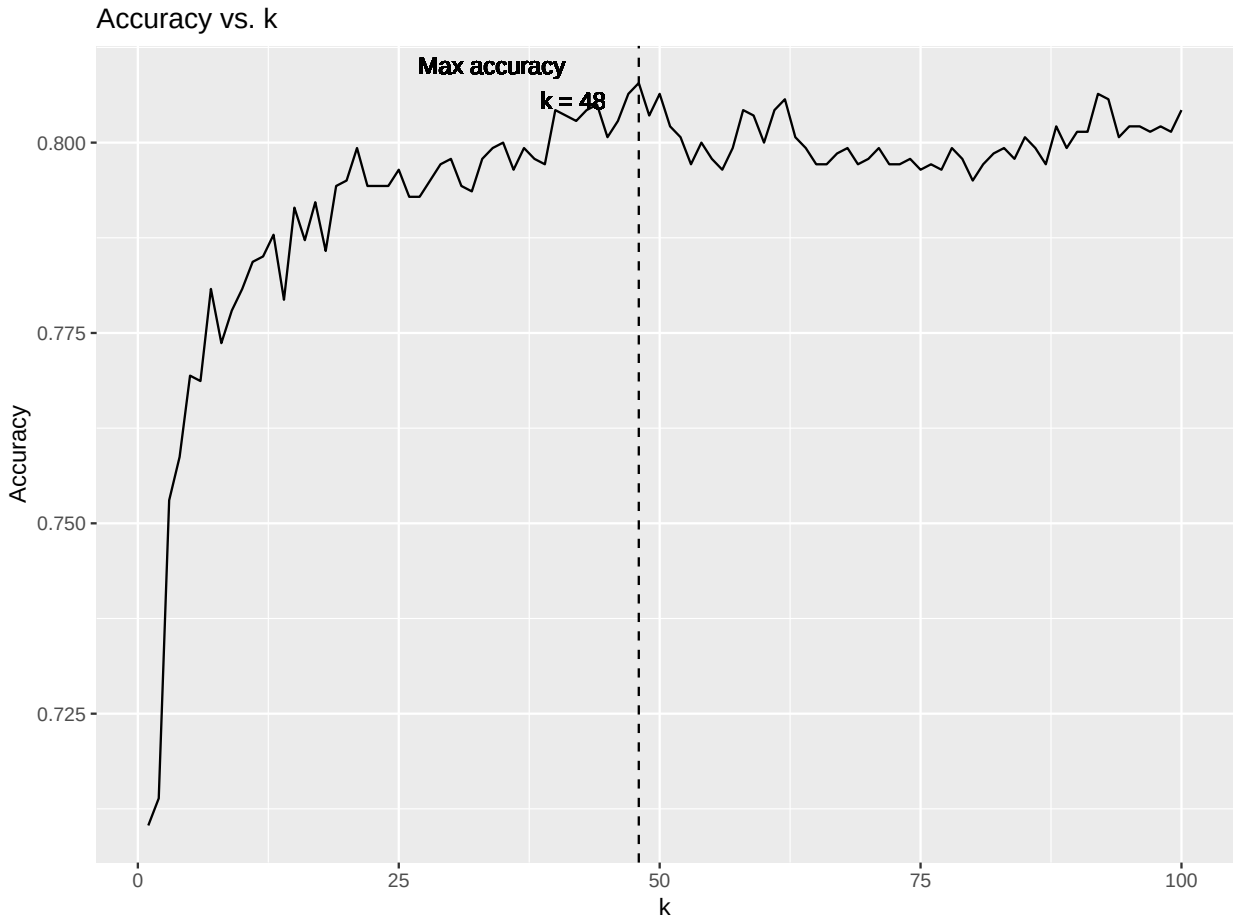


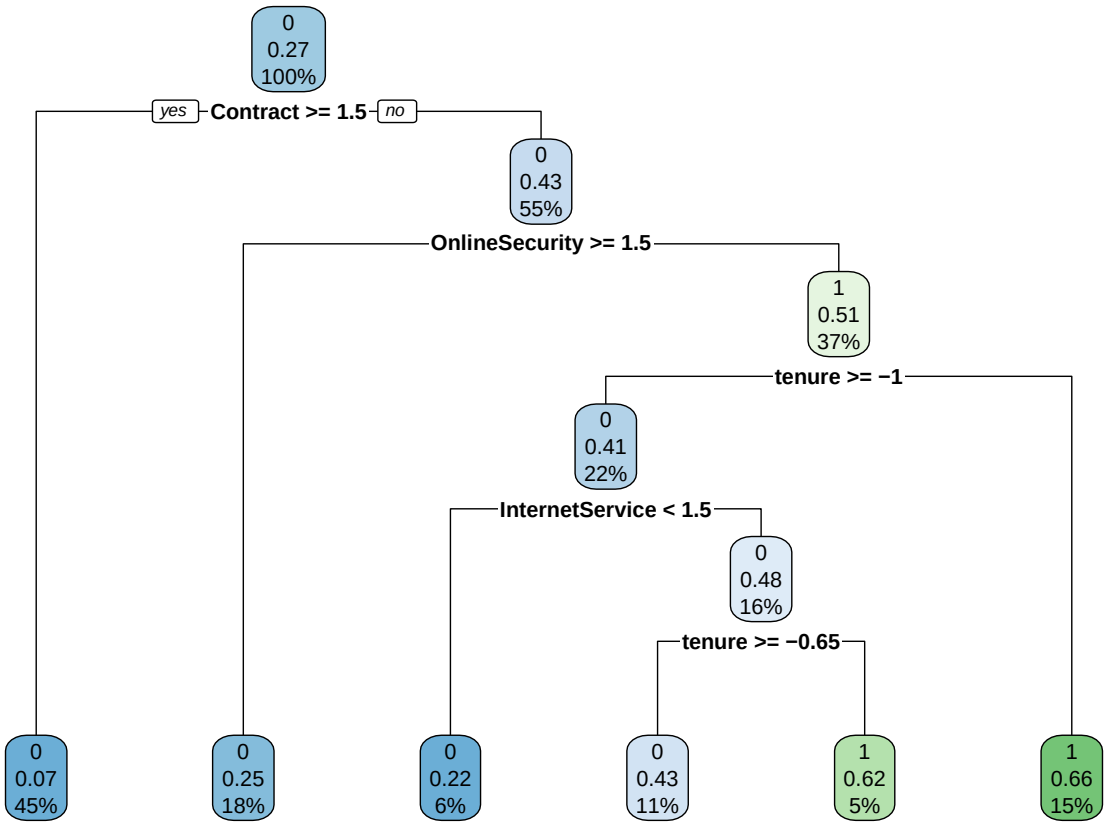
Figure 13: Skuteczność predykcji dla poszczególnych wartości k

Drzewo decyzyjne

W tym przypadku użyliśmy funkcji *rpart* z pakietu *rpart*. Na wykresie poniżej widzimy jakie zmienne warunkowały kolejne przejścia w drzewie. Natomiast macierz pomyłek 5 wskazuje, że skuteczność wyniosła 0.79. Oczywiście pjedyncze drzewo jest bardzo niestabilne, dlatego zastosujemy też metody ze wzmocnieniem.

	0	1
0	917	115
1	158	215

Table 4: Confusion matrix for k = 48



	0	1
0	934	98
1	201	172

Table 5: Confusion matrix