Analiza odchodzenia klientów

Bartosz Chądzyński 255680 & Michał Turek 246993

2023-05-07

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

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

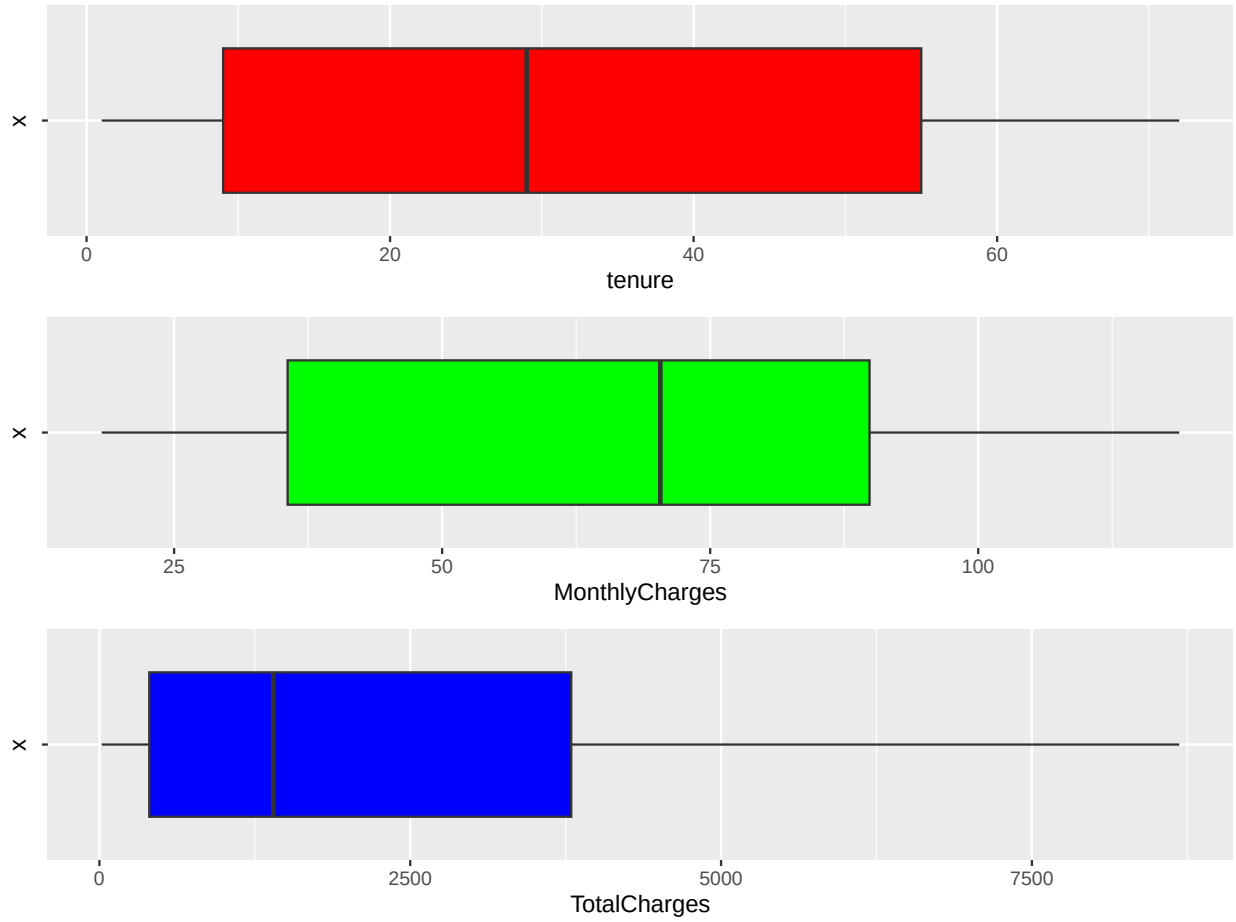


Figure 1: Boxploty zmiennych ciągłych

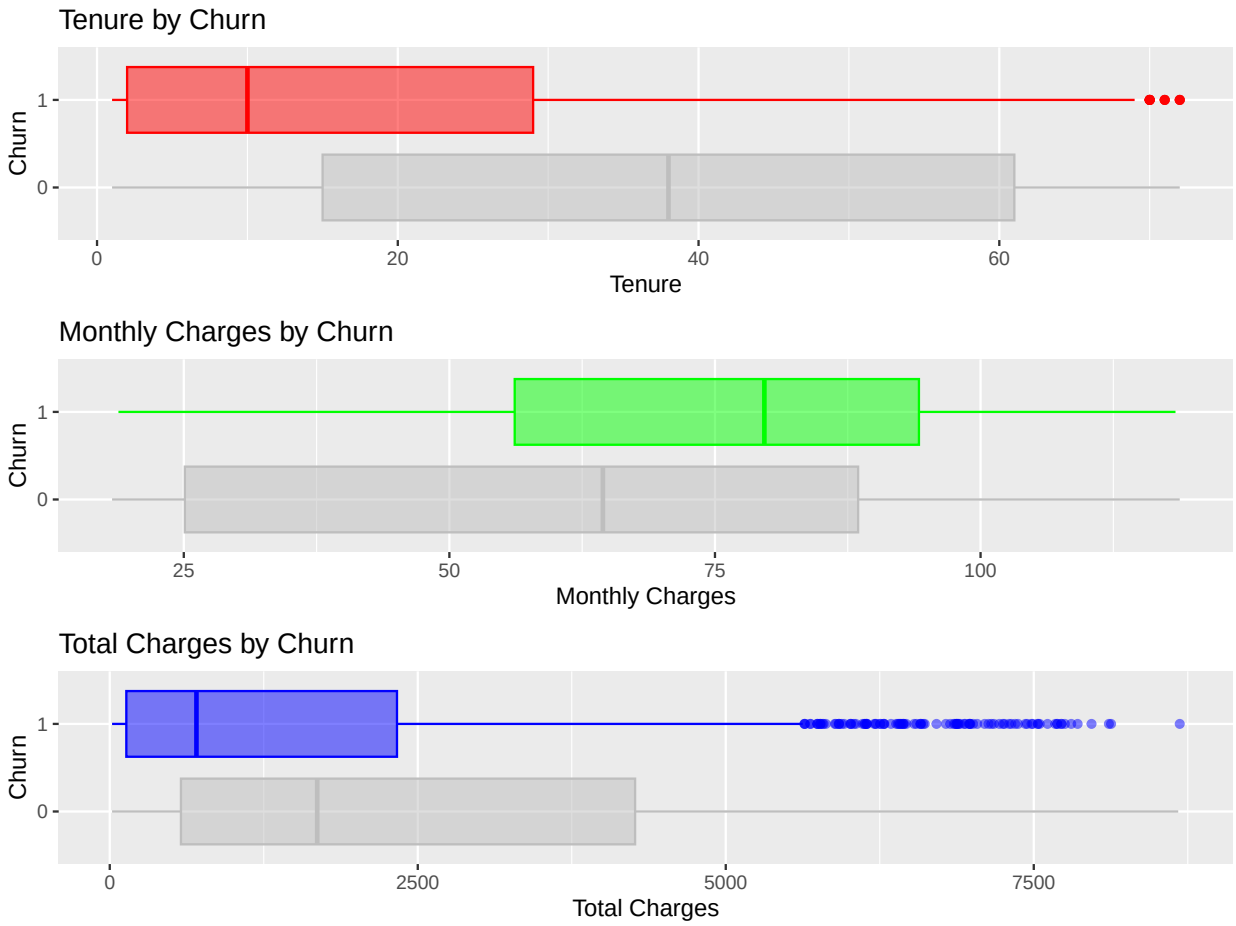


Figure 2: Boxploty zmiennych ciągłych z podziałem ze względu na Churn

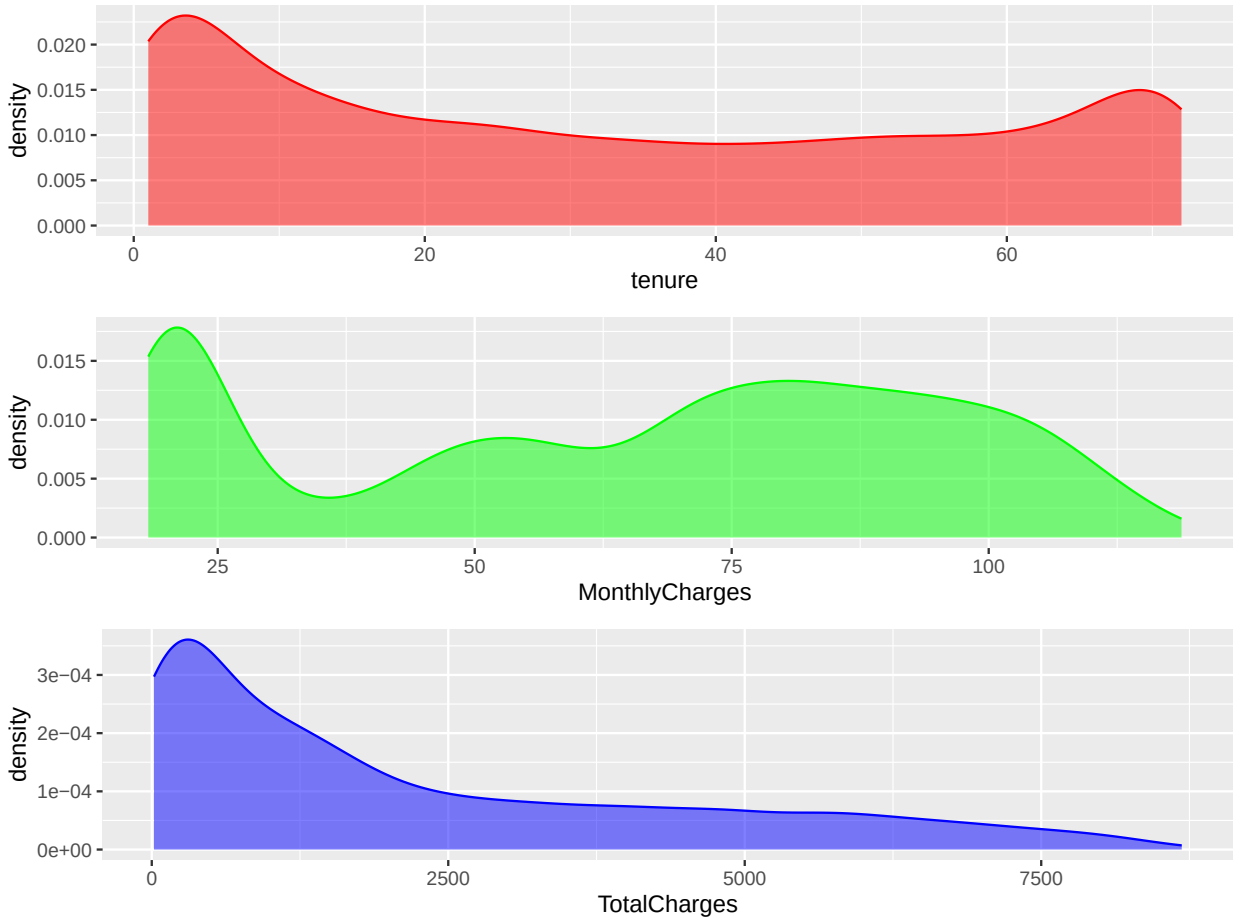


Figure 3: Estymator jądrowy gęstości

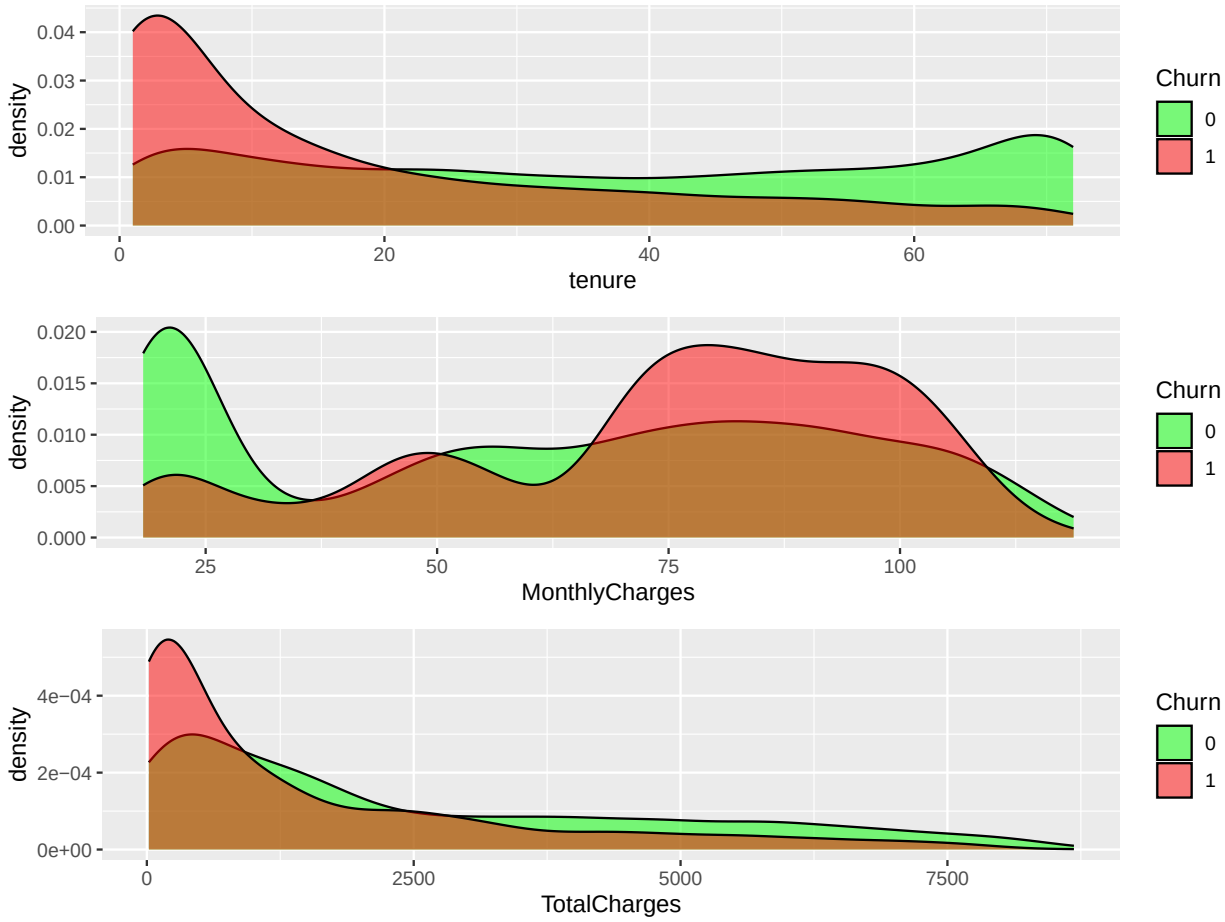


Figure 4: Estymator jądrowy gęstości z podziałem ze względu na Churn

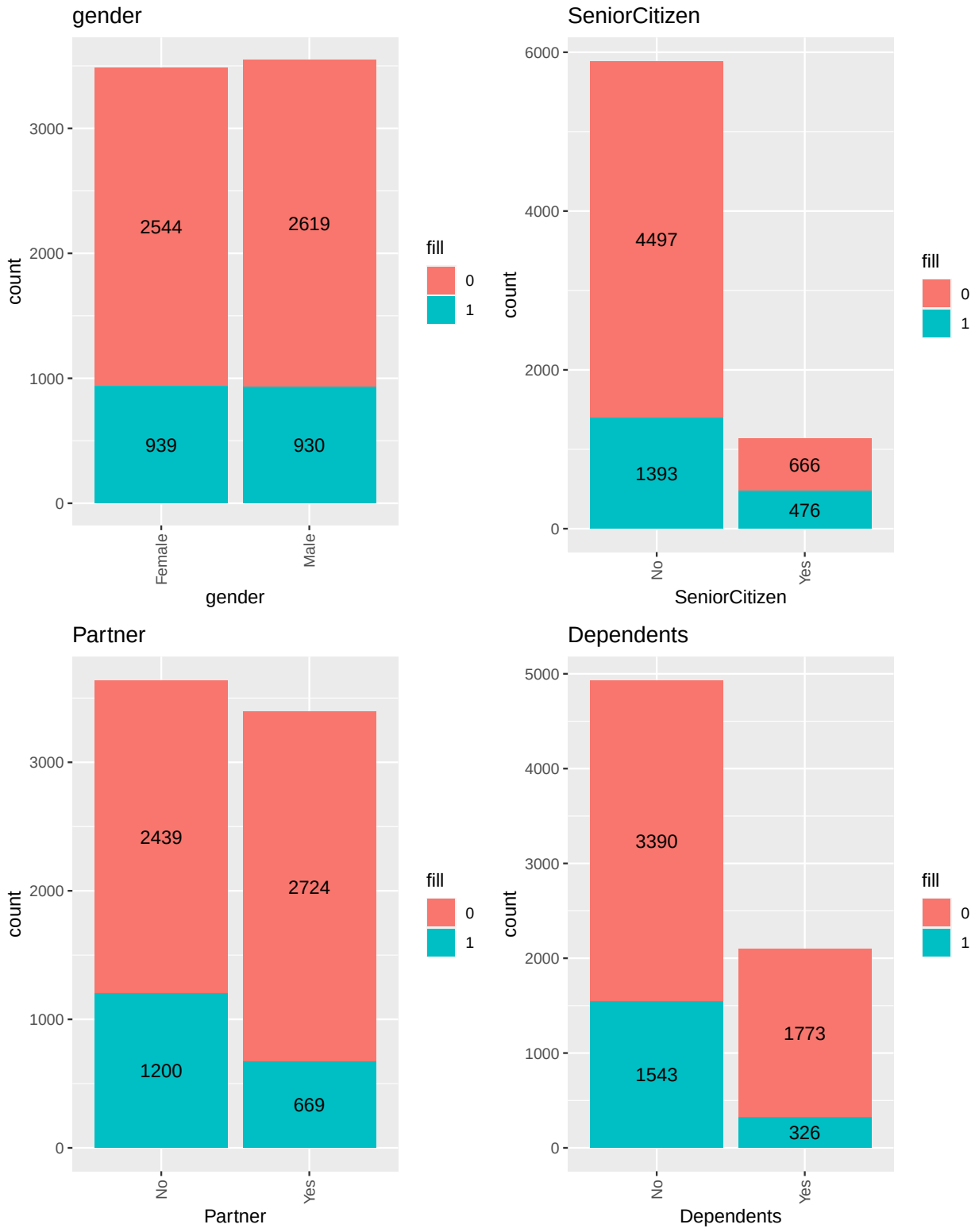


Figure 5: Wykres ilości obserwacji z podziałem na kategorie zmiennych

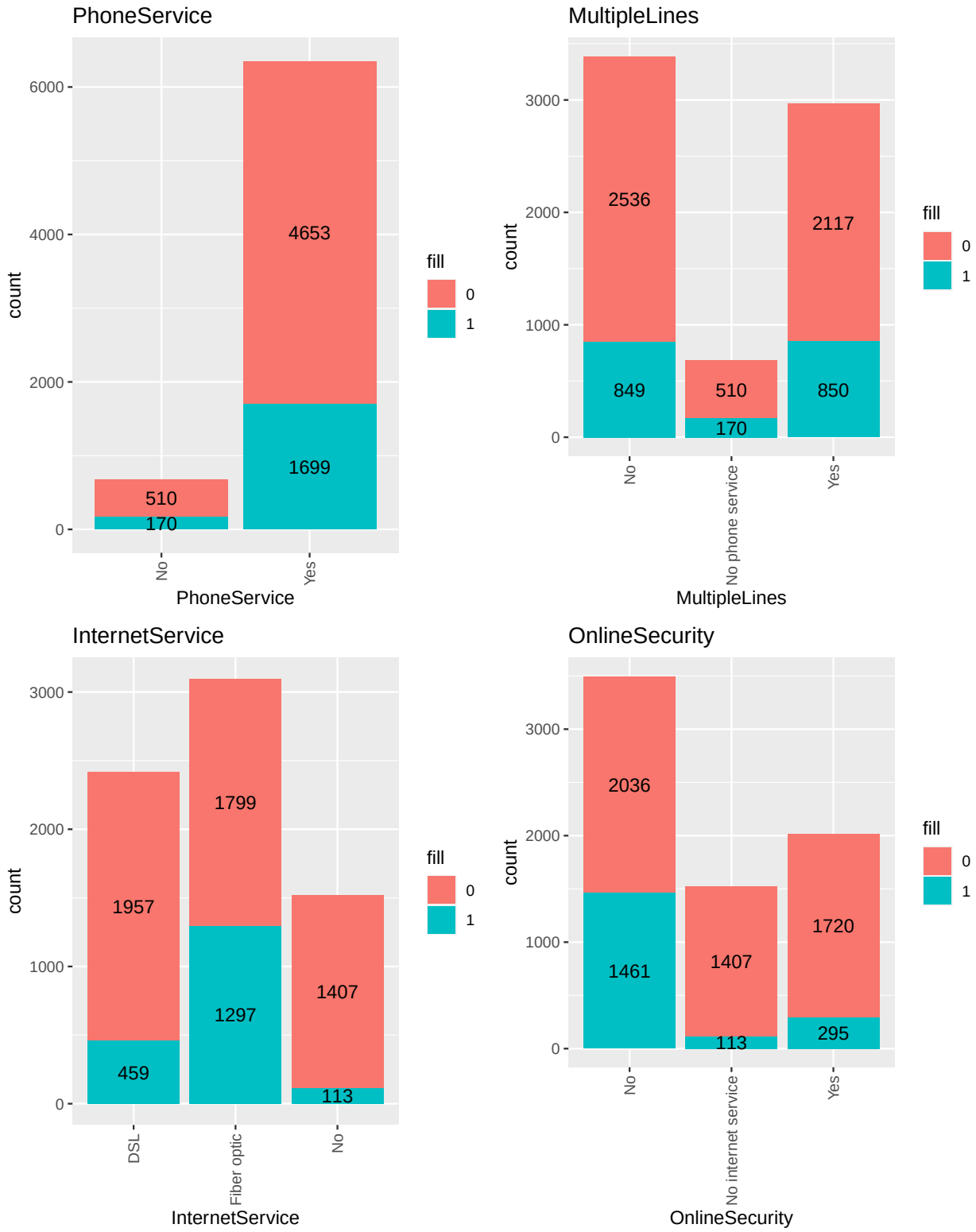


Figure 6: Wykres ilości obserwacji z podziałem na kategorie zmiennych

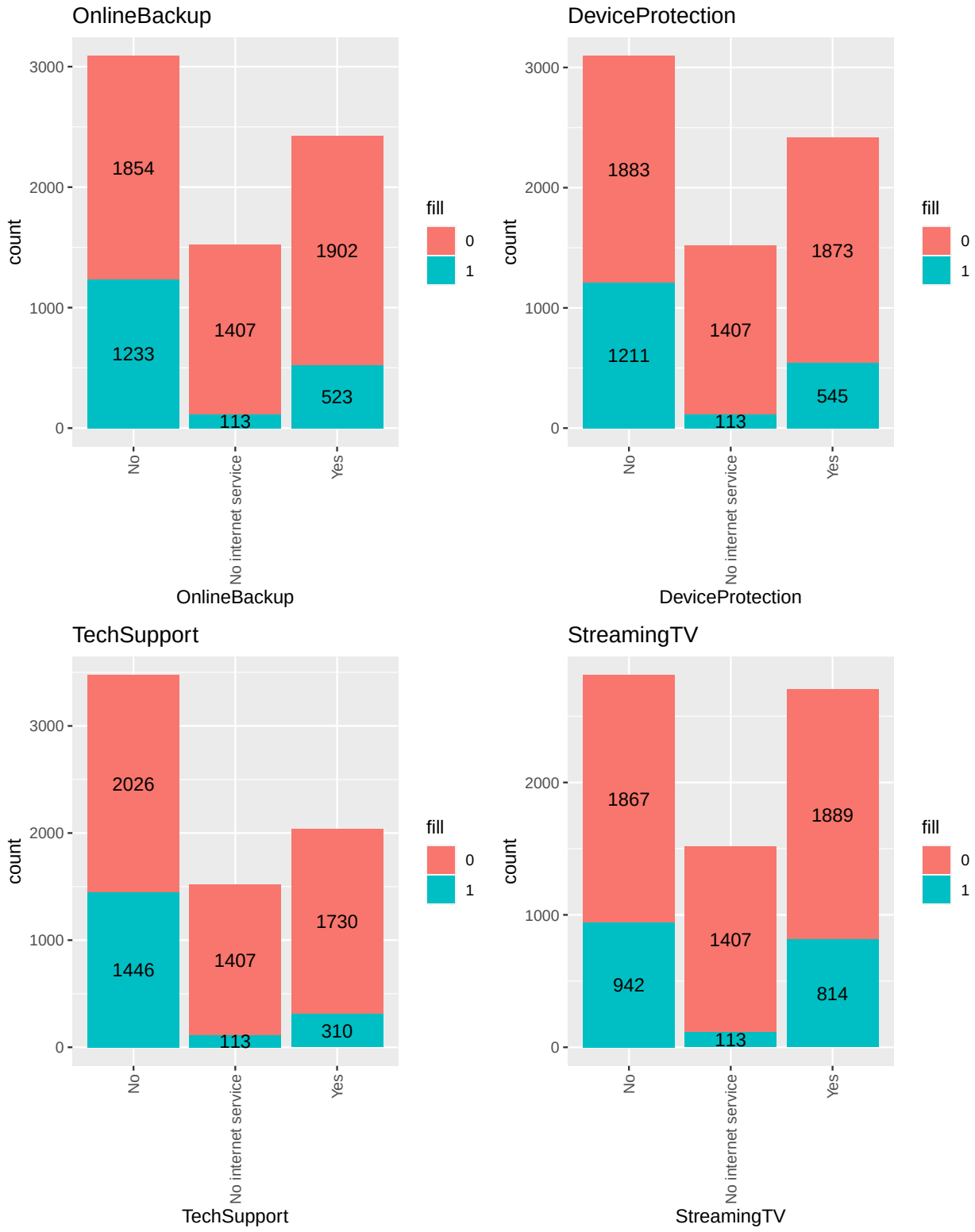


Figure 7: Wykres ilości obserwacji z podziałem na kategorie zmiennych

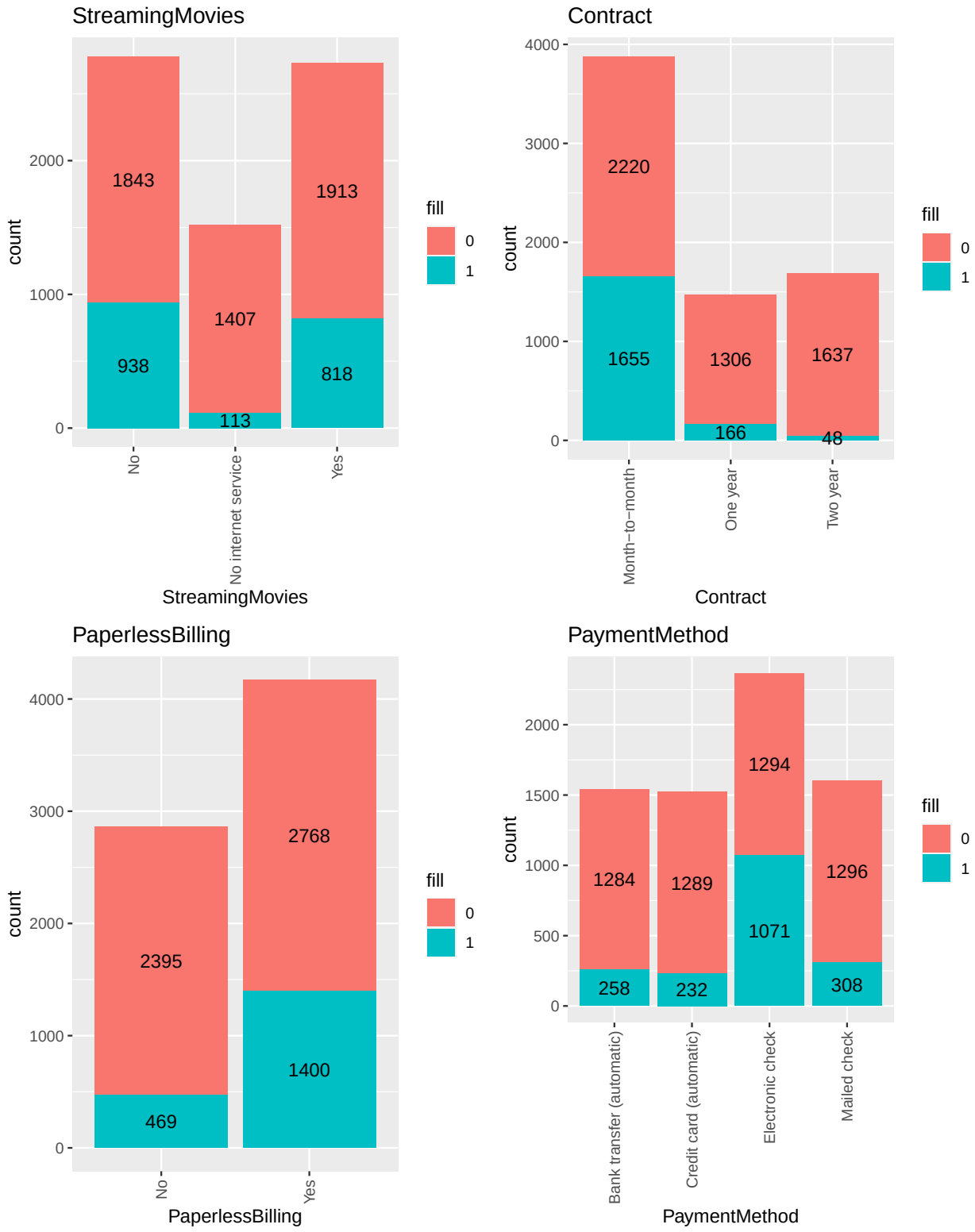


Figure 8: Wykres ilości obserwacji z podziałem na kategorie zmiennych

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.

	Estymowane 0	Estymowane 1
Rzeczwiste 0	968	220
Rzeczwiste 1	64	153

Table 1: Macierz pomyłek dla regresji liniowej, z punktem odcięcia = 1.52

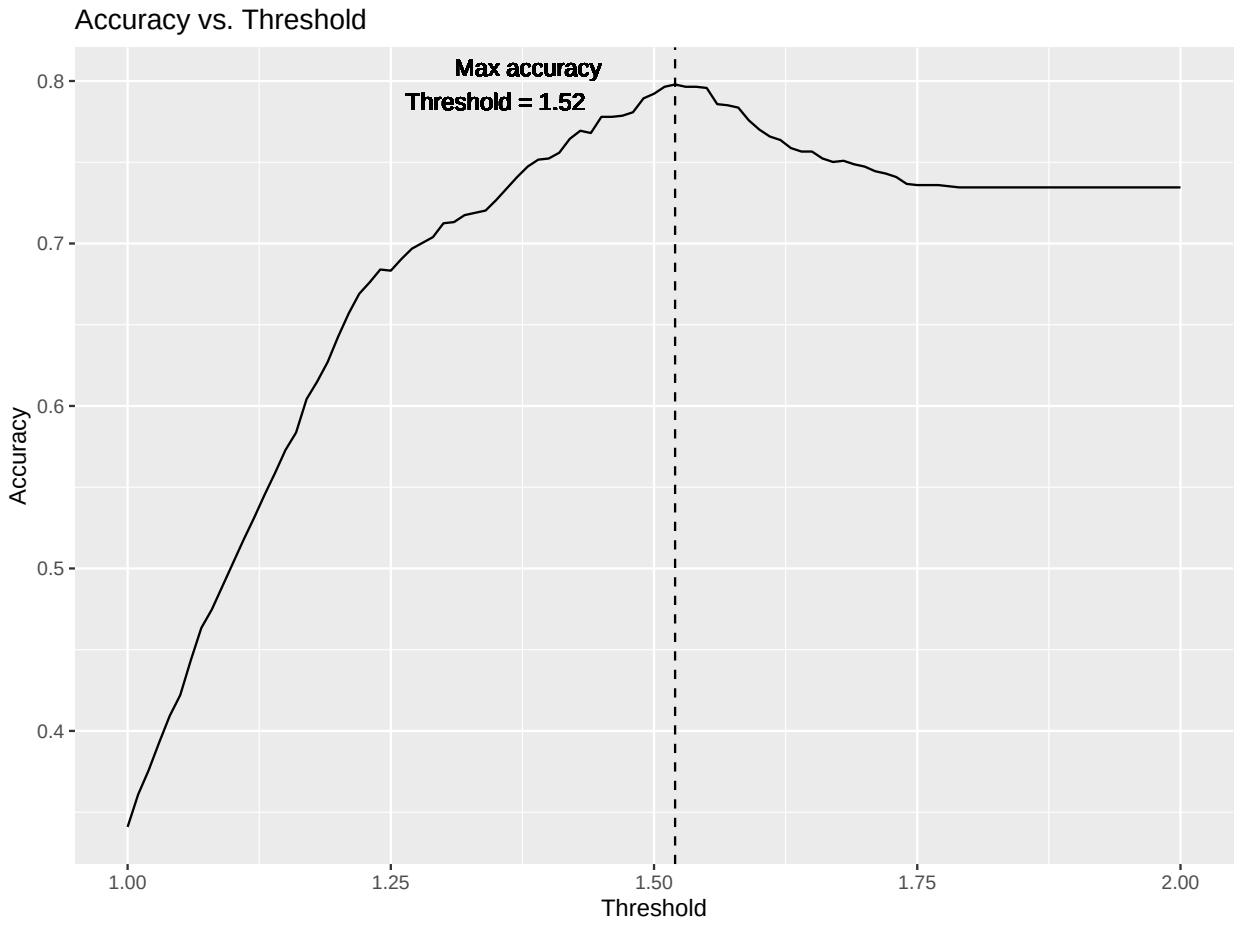


Figure 9: Skuteczność predykcji dla poszczególnych punktów odcięcia

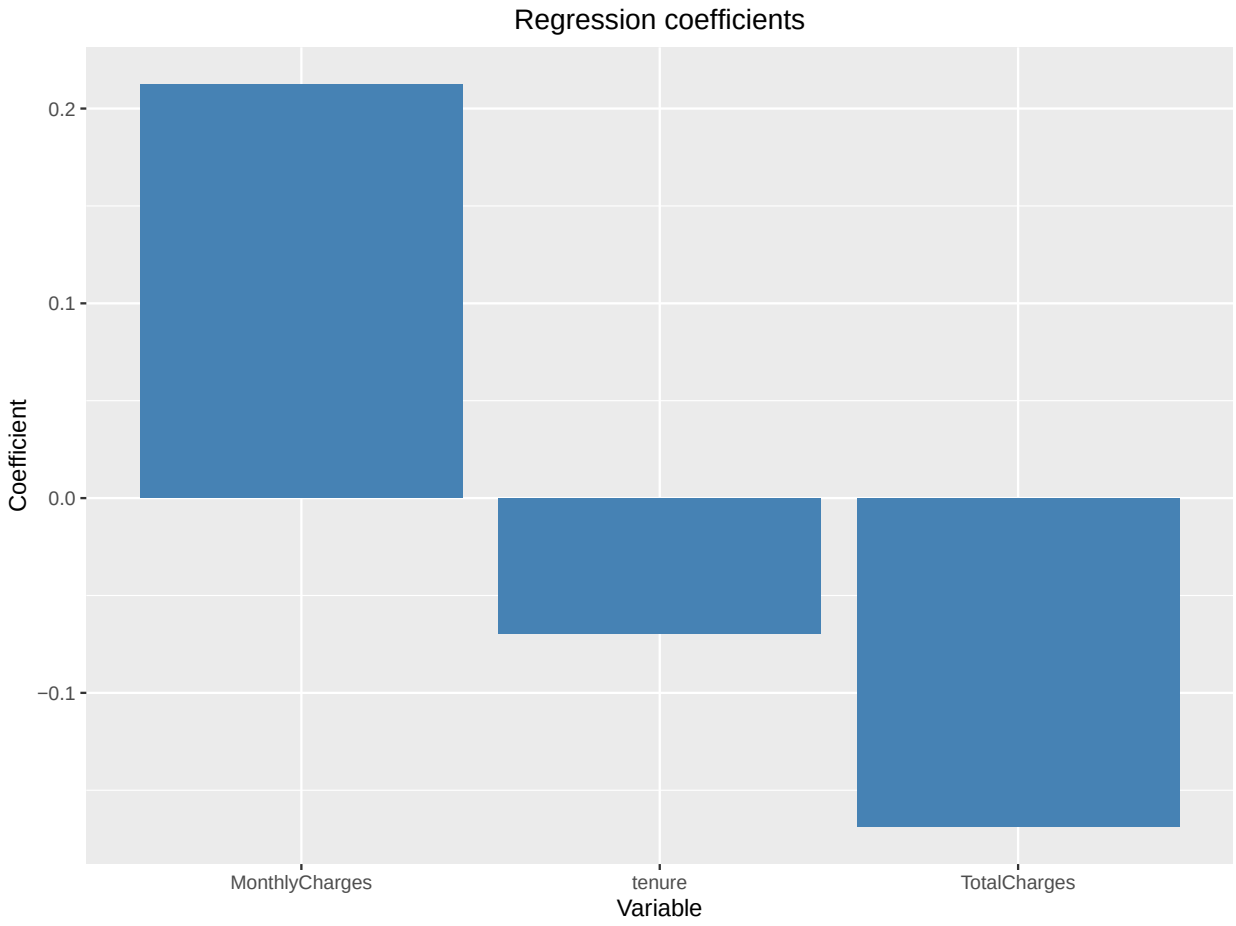


Figure 10: wartości współczynników w modelu regresji logistycznej

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*.

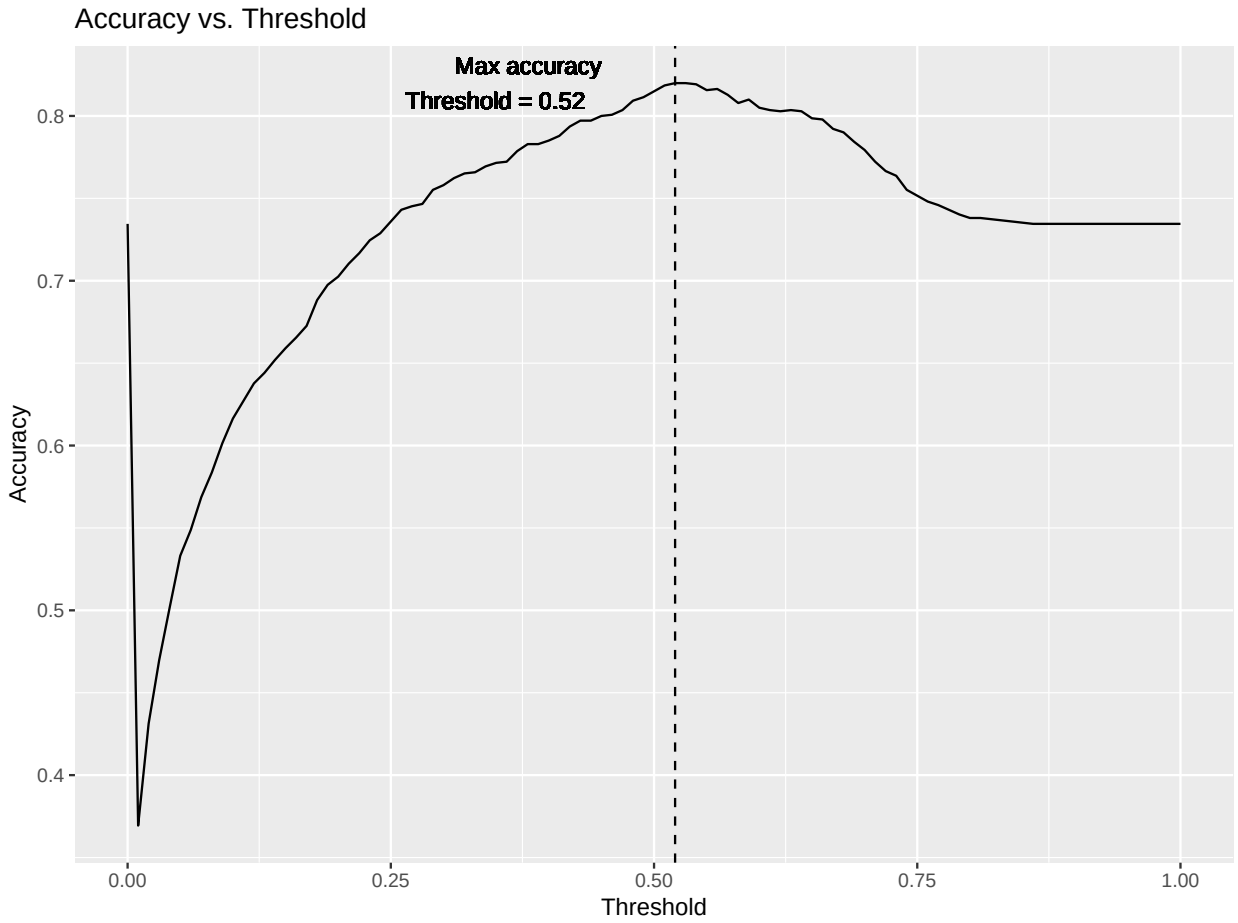


Figure 11: Skuteczność predykcji dla poszczególnych punktów odcięcia

	Estymowane 0	Estymowane 1
Rzeczwiste 0	942	163
Rzeczwiste 1	90	210

Table 2: Macierz pomyłek dla regresji logistycznej, z punktem odcięcia = 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.

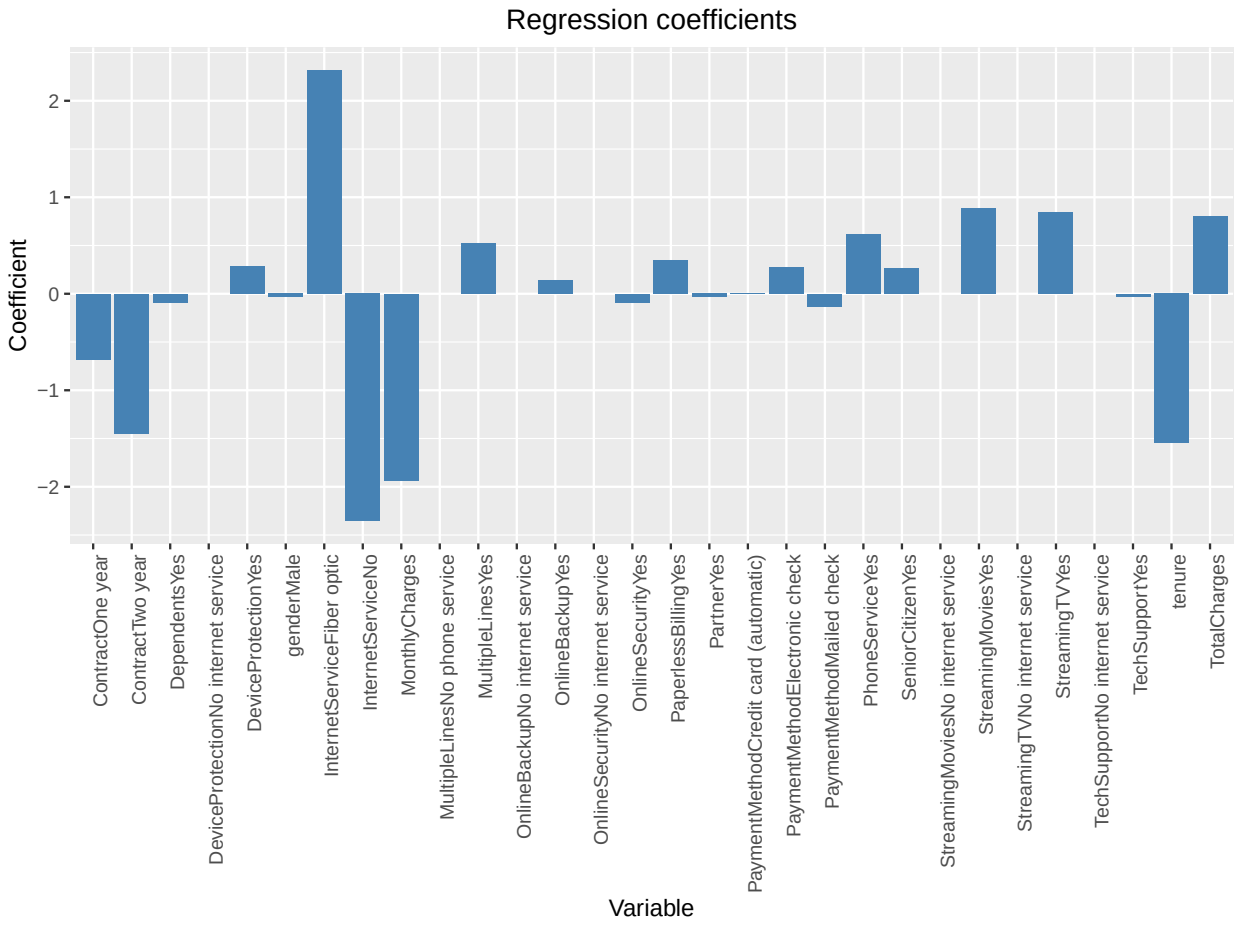


Figure 12: wartości współczynników w modelu regresji logistycznej

	Estymowane 0	Estymowane 1
Rzeczwiste 0	778	254
Rzeczwiste 1	90	283

Table 3: Macierz pomyłek dla algorytmu Naiwnego Bayesa

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.

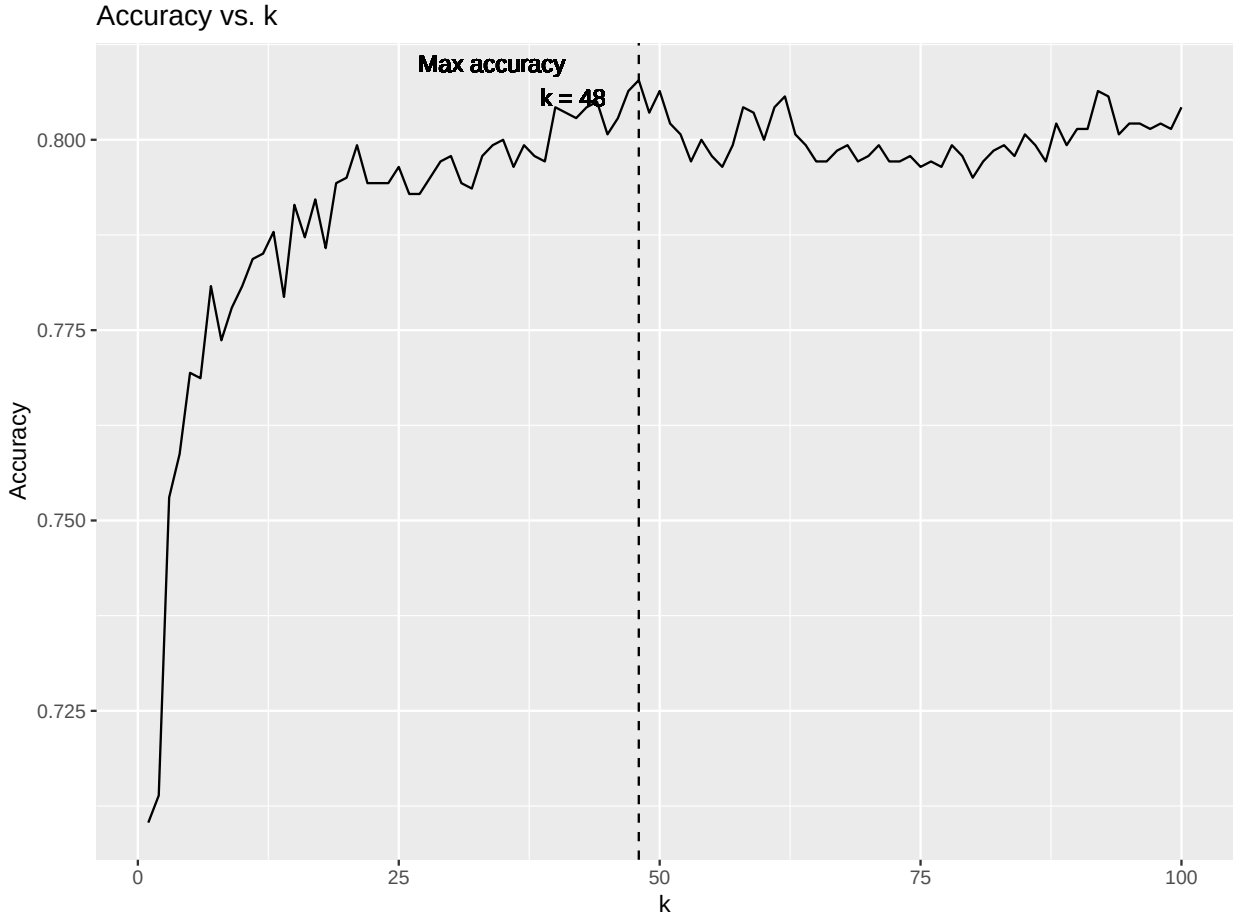


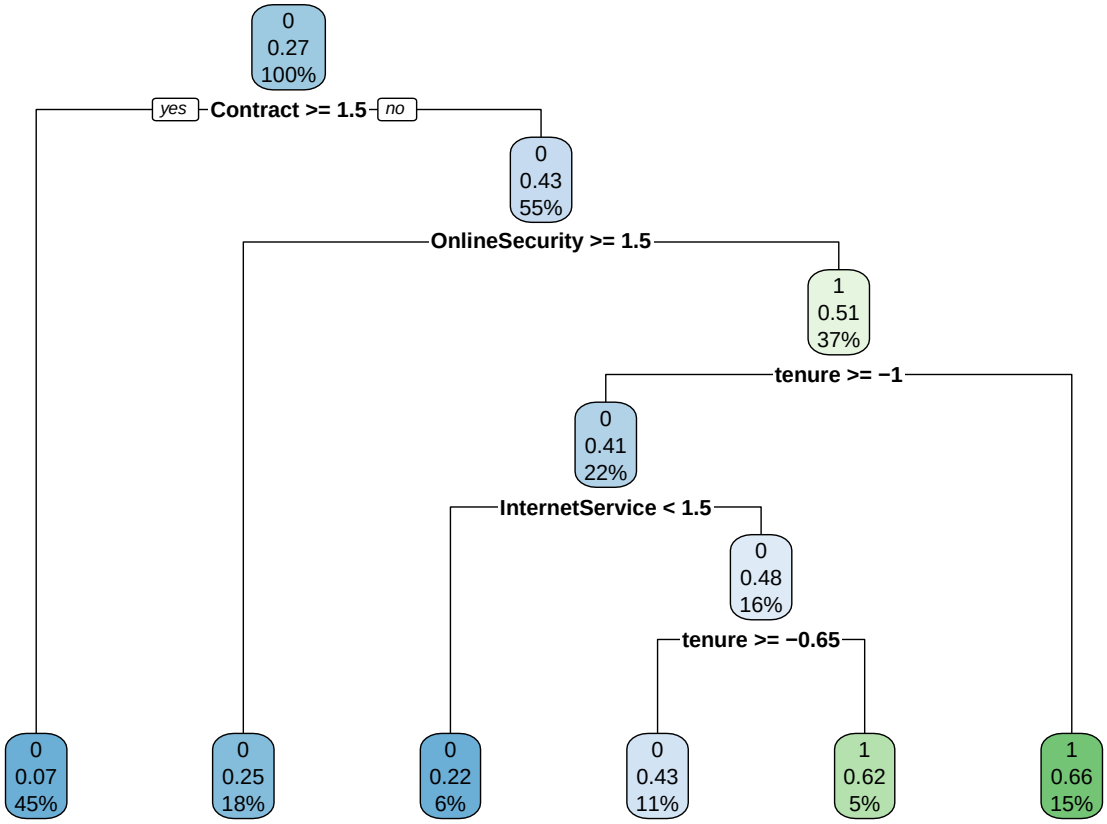
Figure 13: Skuteczność predykcji dla poszczególnych wartości k

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.

	Estymowane 0	Estymowane 1
Rzeczwiste 0	917	115
Rzeczwiste 1	158	215

Table 4: Macierz pomyłek dla algorytmu kNN, k = 48



	Estymowane 0	Estymowane 1
Rzeczwiste 0	934	98
Rzeczwiste 1	201	172

Table 5: Macierz pomyłek dla drzewa decyzyjnego

SVM

Algorytm SVM zastosowaliśmy z 4 rodzajami jąder: liniowym, wielomianowym (stopień 3), radialnym i sigmoidalnym. W tabelach 6 - 9 widzimy wyniki. Jak widać największą skuteczność otrzymujemy stosując jądro radialne i jest ona na poziomie 0.806. Równie dobrze algorytm działa w przypadku jądra liniowego, gdzie skuteczność wynosi 0.804. W pozostałych przypadkach wyniki są dużo gorsze.

Następnie zastosujemy 3 algorytmy wzmacniające. Będą to lasy losowe (Random forest), bagging i boosting. Dla tego pierwszego sprawdzimy również, jak na dokładność działania algorytmu wpłynie ilość drzew.

	Estymowane 0	Estymowane 1
Rzeczwiste 0	923	167
Rzeczwiste 1	109	206

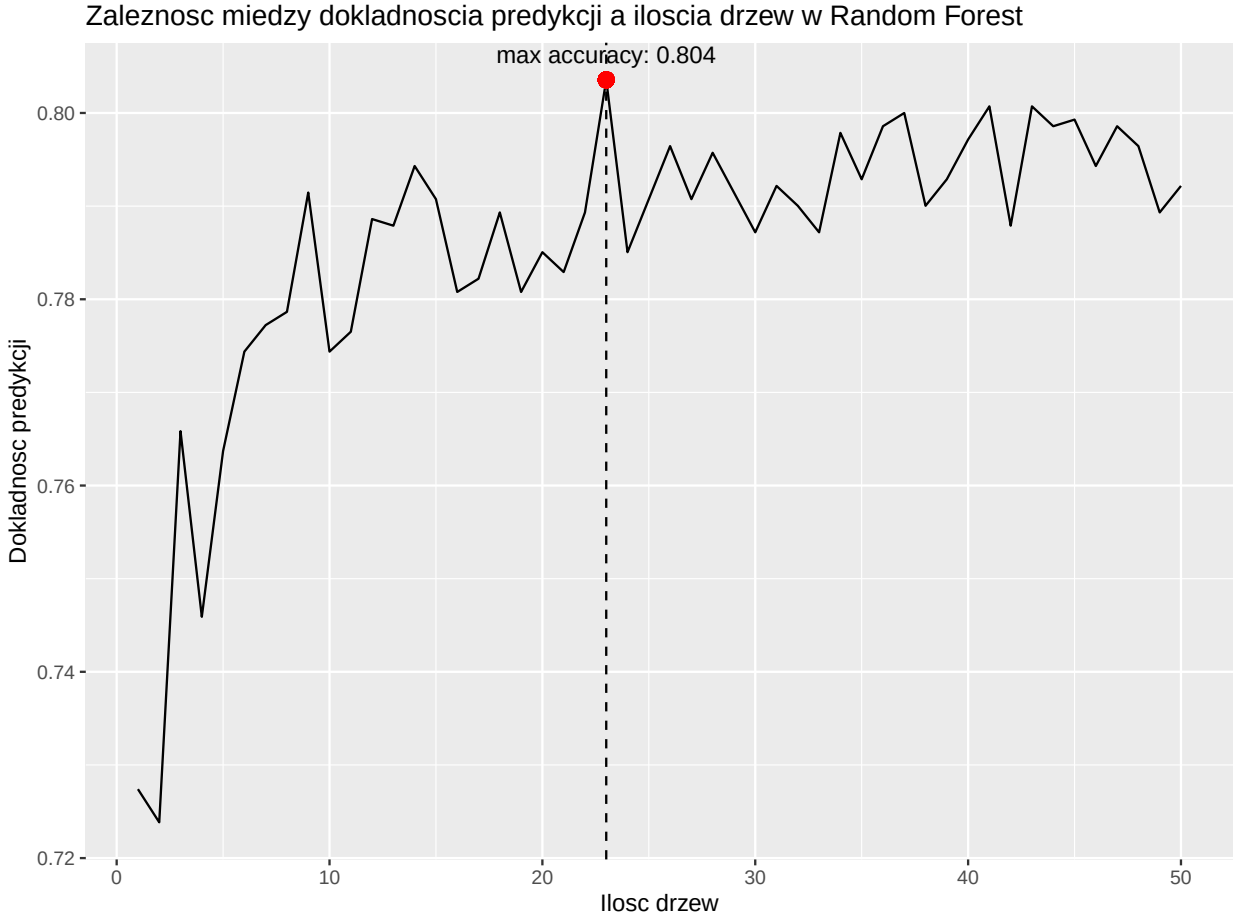
Table 6: Macierz pomyłek dla SVM (jądro liniowe)

	Estymowane 0	Estymowane 1
Rzeczwiste 0	931	196
Rzeczwiste 1	101	177

Table 7: Macierz pomyłek dla SVM (jądro wielomianowe)

Random Forest

Zacznijmy od algorytmu Random Forest. Skorzystamy z funkcji randomForest z biblioteki o tej samej nazwie. Na wykresie widzimy, że najlepsza dokładność predykcji wychodzi dla 23 drzew i wynosi ona nieco ponad 0.8. Przedstawiamy wyniki tylko dla $n_trees \leq 50$, bo dla większej ilości drzew różnica w dokładności jest już znikoma, a znacznie wzrasta złożoność obliczeniowa. Na podstawie tabeli pomyłek widzimy, że nasz model ma tendencję do lepszego klasyfikowania obserwacji negatywnych (0) niż pozytywnych (1).



	Estymowane 0	Estymowane 1
Rzeczwiste 0	945	188
Rzeczwiste 1	87	185

Table 8: Macierz pomyłek dla SVM (jądro radialne)

	Estymowane 0	Estymowane 1
Rzeczwiste 0	862	192
Rzeczwiste 1	170	181

Table 9: Macierz pomyłek dla SVM (jądro sigmoidalne)

Boosting

Teraz przejdziemy do metody Boosting. Skorzystamy z bibliotek xgboost i pROC. Dostaliśmy dokładność na poziomie około 0.8, a z macierzy pomyłek możemy odczytać, że ten model ponownie lepiej poradził sobie z klasyfikacją obserwacji negatywnych, niż pozytywnych.

Accuracy: 0.802847

	Estymowane 0	Estymowane 1
Rzeczwiste 0	911	179
Rzeczwiste 1	121	194

Table 10: Macierz pomyłek dla lasu losowego

	Estymowane 0	Estymowane 1
Rzeczwiste 0	924	169
Rzeczwiste 1	108	204

Table 11: Macierz pomyłek dla boostingu

Bagging

Ostatnim algorytmem, który zastosujemy, będzie algorytm Bagging. W celu kompilacji modelu skorzystamy z biblioteki ipred.Dokładność tego modelu wyszła około 0.79, czyli nieznacznie mniej, niż w przypadku dwóch poprzednich modeli. Mamy ponownie sytuację, w której model lepiej klasyfikuje klientów, którzy zostali w firmie, niż tych, którzy odeszli.

Accuracy: 0.7893238

	Estymowane 0	Estymowane 1
Rzeczwiste 0	943	207
Rzeczwiste 1	89	166

Table 12: Macierz pomyłek dla baggingu