Projekt zaliczeniowy

Fragment projektu opracowanego na przedmiot Data Mining

Spis treści

Dokument zawiera fragmenty raportu. Pozostawiono rozdziały, których część badawcza została przygotowana samodzielnie przeze mnie (pominięto rozdział o sieciach neuronowych).

I. Wstęp

Według raportu We Are Social i Hootsuit z 2018 roku z telefonów komórkowych korzysta już 5,135 miliarda ludzi na całym świecie. Również liczba dostępnych sieci komórkowych stale rośnie, stąd firmy telekomunikacyjne muszą nie tylko starać się przyciągnąć nowych klientów, ale także, co wcale nie jest mniej ważne w strategii firmy, utrzymać tych obecnych. Pozyskiwanie nowych klientów jest kosztownym procesem, dlatego firmy poświęcają dużo uwagi analizie odejść swoich klientów, by następnie móc wykorzystywać uzyskane wnioski do redukcji liczby traconych klientów.

Celem badania przedstawionego w poniższej pracy jest zrozumienie, którzy klienci odchodzą do konkurencji i z jakiego powodu. Do znalezienia pewnych reguł, które mogą być pomocne, aby zrozumieć problem, w kolejnych rozdziałach zostaną wykorzystane takie metody jak regresja logistyczna, drzewa decyzyjne i sieci neuronowe. Na końcu pracy zbudowane modele zostaną porównane w celu wybrania tego, który najlepiej opisuje dany problem.

II. Opis zbioru danych

Zbiór danych pochodzi z IBM Watson Analytics i dotyczy branży telekomunikacyjnej. Zmienną celu jest zmienna *churn*, a w zbiorze znajduje się 19 zmiennych objaśniających, z czego 11 jest nominalnych, 3 przedziałowe, a 5 binarnych. Zmienne objaśniające dotyczą przede wszystkim ilości i rodzaju posiadanych produktów w firmie telekomunikacyjnej oraz typu i długości zawartej umowy.

Tabela 1. Opis zmiennych w zbiorze danych oraz podstawowe statystyki

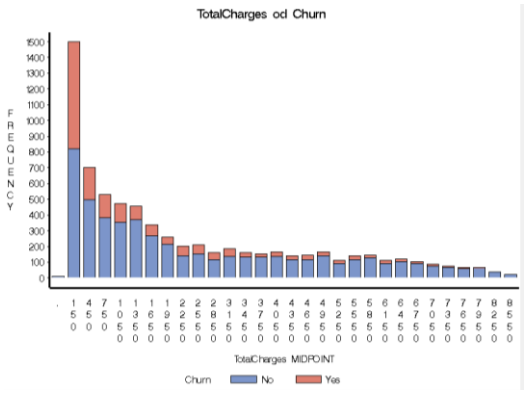
Nazwa zmiennej	Opis biznesowy zmiennej	Opis wartości zmiennej i podstawowe statystyki	Typ zmiennej
Churn	Czy klient odszedł?	tak/nie 26,5% z badanych klientów odeszło (churn = yes)	binarna, zmienna celu
Gender	Płeć	kobieta/mężczyzna 49,5% klientów to kobiety	binarna
SeniorCitizen	Czy klient jest emerytem?	tak/nie 16,2% klientów jest emerytami	binarna
Partner	Czy klient ma partnera?	tak/nie 48,3% klientów ma partnera	binarna
Dependents	Czy klient ma osoby na utrzymaniu?	tak/nie 30,0% klientów ma kogoś na utrzymaniu	binarna

Tenure	Liczba miesięcy, od kiedy osoba jest klientem firmy	Zmienna z przedziału 0-72, średnia wartość 32, przedział o największej liczebności: 0-7,2 (22,9% obserwacji)	przedziałowa
PhoneService	Czy klient posiada usługę telefoniczną?	tak/nie 90,3% klientów posiada usługę	binarna
MultipleLines	Czy klient posiada kilka numerów?	tak (42,2%)/nie (48,1%)/nie ma telefonu (9,7%)	nominalna
InternetService	Typ dostępu do internetu	Cyfrowa linia abonencka (34,3%) /światłowód (44%)/żaden (21,7%)	nominalna
OnlineSecurity	Czy klient korzysta z ochrony w internecie?	tak (28,7%)/nie (49,7%)/nie ma internetu (21,7%)	nominalna
OnlineBackup	Czy klient tworzy kopie zapasowe online?	tak (34,5%)/nie (43,8%)/nie ma internetu (21,7%)	nominalna
DeviceProtection	Czy klient korzysta z ochrony urządzenia?	tak (34,3%)/nie (44%)/nie ma internetu (21,7%)	nominalna
TechSupport	Czy klient korzysta ze wsparcia technicznego online?	tak (29%)/nie (49,3%)/nie ma internetu (21,7%)	nominalna
StreamingTV	Czy klient korzysta ze streamingu telewizji?	tak (38,4%)/nie (39,9%)/nie ma internetu (21,7%)	nominalna
StreamingMovies	Czy klient korzysta ze streamingu filmów?	tak (38,8%)/nie (39,5%)/nie ma internetu (21,7%)	nominalna
Contract	Długość trwania umowy	Odnawialna z miesiąca na miesiąc (55%)/ roczna (21%)/dwuletnia (24%)	nominalna
PaperlessBilling	Czy klient korzysta z faktur elektronicznych?	tak/nie 59,2% klientów korzysta	binarna
PaymentMethod	Sposób płatności	płatność elektroniczna (33,6%)/płatność na poczcie (22,8%)/stały przelew (22%)/automatyczna zapłata kartą kredytową (21,6%)	nominalna
MonthlyCharges	Miesięczna wartość rachunków	Wartość z przedziału 18,25-118,75 ze średnią 64. Największa częstotliwość w przedziale 18,25-18,30 (22% obserwacji)	przedziałowa
TotalCharges	Suma rachunków	Wartość z przedziału 18,8-8684,8 ze średnią 2283. Największa częstotliwość w przedziale 18,8-885,4 (38,5% obserwacji).	przedziałowa

źródło: opracowanie własne

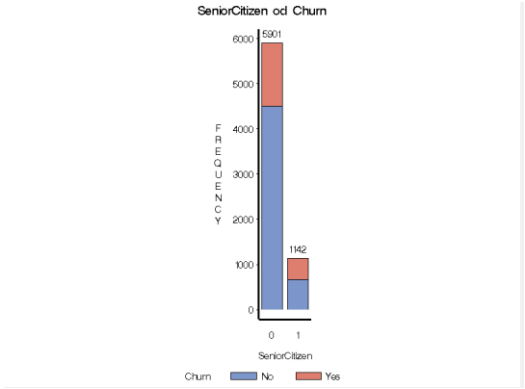
W analizie wzięło udział 7043 obserwacji. Braki danych w większości zmiennych nie występują, jedynie dla zmiennej *TotalCharges* zanotowano 11 brakujących wartości, które zostały usunięte z próby ze względu na niewielki udział i brak zauważonej korelacji z wartościami pozostałych zmiennych. Wstępna eksploracja danych pozwala zauważyć kilka pierwszych zależności:

Wykres 1. Histogram zmiennej TotalCharges w zależności od Churn



źródło: opracowanie własne z użyciem SAS Enterprise Miner

Wykres 2. Histogram zmiennej SeniorCitizen w zależności od Churn



źródło: opracowanie własne z użyciem SAS Enterprise Miner

W grupie z najniższymi totalnymi opłatami stosunek osób, które odeszły i nie jest niemal równy. W kolejnych grupach liczba klientów odchodzących jest relatywnie dużo niższa. Nie musi to być skorelowane z wysokością opłaty, a jedynie z wielkością grupy (mniej liczebne grupy są mało reprezentatywne i łatwiej o wartości skrajne), niemniej warto zwrócić na to uwagę. Podczas analizy zmiennej *SeniorCitizen* zauważono, że emeryci charakteryzują się zdecydowanie wyższym stosunkiem klientów odchodzących niż młodsze osoby, co kłóci się z intuicją (bardziej prawdopodobna wydaje się wyższa lojalność klientów tego typu i ich niechęć do zmian).

III. Regresja logistyczna

Ze względu na jakościową zmienną celu pierwszym z analizowanych modeli jest regresja logistyczna. Po podziale zbioru na część uczącą, walidującą i testową w proporcjach odpowiednio 60%, 30% i 10% zastosowano węzeł *Wybór zmiennych*.

Ze względu na niską wartość R-kwadrat, w modelu nie uwzględniono 9 zmiennych:

- Dependents

- DeviceProtection

- MonthlyCharges

- MultipleLines

- OnlineBackup

- Partner

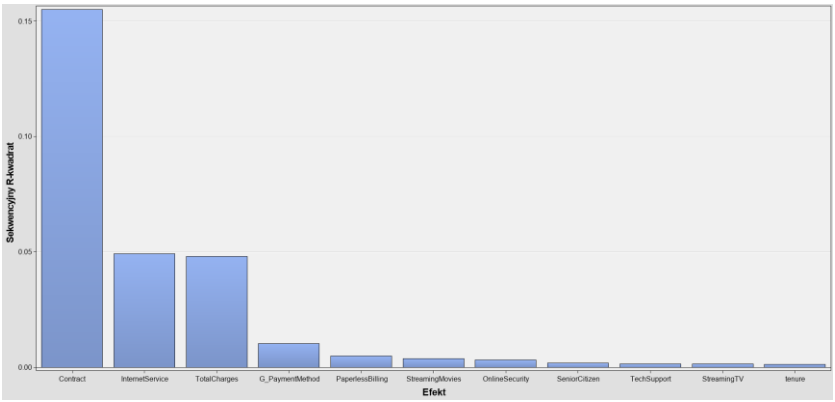
- *PaymentMethod* (ta zmienna została zastąpiona zgrupowaną)

- PhoneService

- gender.

Oryginalna zmienna *PaymentMethod* została zgrupowana dla osiągnięcia lepszych wyników - wartości różne od płatności elektronicznej zostały oznaczone jako 1, a płatność elektroniczna jako 0. Nowa zmienna, *G_PaymentMethod*, znalazła się w modelu.

Wykres 3. Efekty zmiennych.



źródło: opracowanie własne z użyciem SAS Enterprise Miner

Najlepiej zmienną celu wyjaśnia typ umowy, następnie typ dostępu do internetu oraz suma wpłat. Pozostałe ze zmiennych umieszczonych w modelu mają zdecydowanie mniejszy udział w wyjaśnianiu zmienności.

Następnie, korzystając z wyboru zmiennych do modelu metodą regresji krokowej, stworzono trzy modele regresji logistycznej:

- model ze zmodyfikowanymi zmiennymi i stałą (model 1)

- model ze zmodyfikowanymi zmiennymi bez stałej (model 2)

- model bez modyfikacji zmiennych (model 3).

Wyniki jakości dopasowania zostały przedstawione w tabeli poniżej.

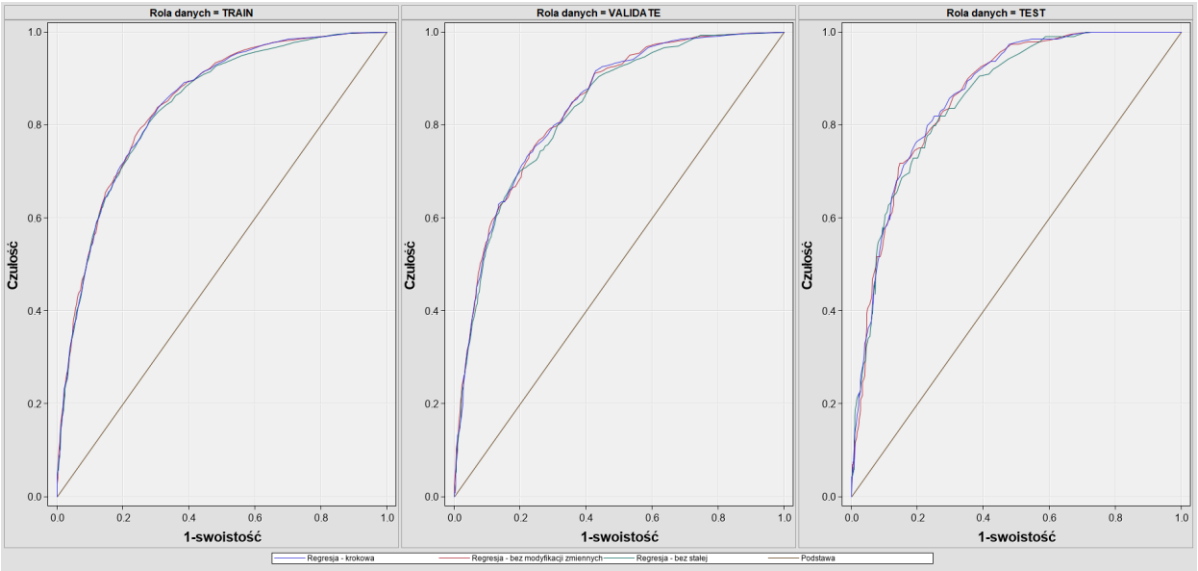
Tabela 2. Oceny jakości dopasowania modeli regresji logistycznej.

Opis modelu	AIC	Odsetek błędnych klasyfikacji			ROC		
	Uczenie	Uczenie	Walidacja	Testowanie	Uczenie	Walidacja	Testowanie
Model 1	3565,377	0,198	0,194	0,194	0,843	0,838	0,866
Model 2	3605,968	0,196	0,198	0,184	0,838	0,830	0,858
Model 3	3558,282	0,199	0,192	0,188	0,845	0,839	0,864

źródło: opracowanie własne z użyciem SAS Enterprise Miner

Na zbiorze uczącym najlepsze dopasowanie mierzone kryterium informacyjnym Akaike miał model 3. bez zmodyfikowanych zmiennych - 3558,28. Drugi w kolejności wynik osiągnął model 1. Na zbiorze testowym najlepsze wyniki mierzone krzywą ROC (który został wybrany jako sposób oceny jakości klasyfikatora) prezentuje model 1. - ze zmodyfikowanymi zmiennymi i stałą. W tabeli zamieszczono dla porównania również odsetki błędnych klasyfikacji poszczególnych model, jednak za ostateczne kryterium wyboru uznano krzywą ROC. Wykres krzywych ROC zaprezentowano na wykresie poniżej.

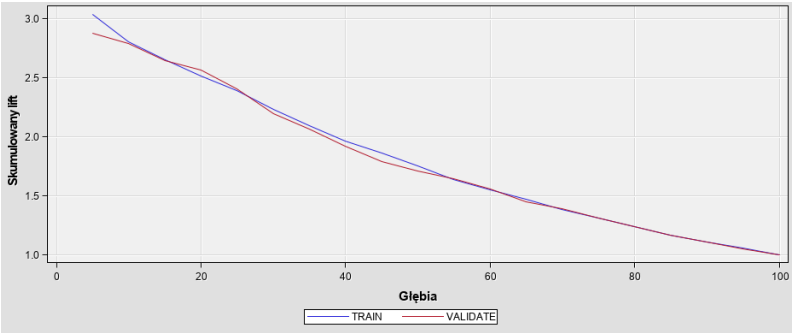
Wykres 4. Krzywa ROC dla trzech modeli regresji logistycznej.



źródło: opracowanie własne z użyciem SAS Enterprise Miner

Model regresji logistycznej ze zmodyfikowanymi zmiennymi i stałą (model 1.) zachowuje się prawidłowo na zbiorze walidacyjnym. Krzywe reprezentujące skumulowany LIFT na zbiorze uczącym i walidacyjnym nie różnią się znacznie.

Wykres 5. Skumulowany LIFT dla modelu 1.



źródło: opracowanie własne z użyciem SAS Enterprise Miner

Biorąc pod uwagę analizę efektów typu III, wszystkie zmienne w ostatecznym modelu są istotne statystycznie na poziomie 0,05.

Tabela 3. Analiza efektów typu III w modelu 1.

Efekt	DF	Chi-kwadrat Walda	Pr. > chi-kw.
Contract	2	29,4561	<0,0001
G_PaymentMethod	1	22,2004	<0,0001
InternetService	2	69,7329	<0,0001
OnlineSecurity	1	18,4210	<0,0001

PaperlessBilling	1	21,8720	<0,0001
SeniorCitizen	1	6,7087	0,0096
StreamingMovies	1	5,1283	0,0235
TechSupport	1	12,1502	0,0005
TotalCharges	1	5,5024	0,0190
tenure	1	43,5461	<0,0001

źródło: opracowanie własne z użyciem SAS Enterprise Miner

Oszacowania parametrów wybranego modelu prezentują się jak poniżej. Zmienne charakteryzujące się pozytywnym wpływem na prawdopodobieństwo odejścia klienta zostały pogrubione (exp(oszacowanie) > 1).

Tabela 4. Oszacowania parametrów zmiennych w modelu 1.

Parametr	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.	Exp(oszaco-wanie)
Intercept	-0,8467	0,1322	41,05	<0,0001	0,429
Contract::Month-to-month	0,5354	0,0991	29,19	<0,0001	1,708
Contract::One year	0,0883	0,0993	0,79	0,3738	1,092
G_PaymentMethod	1	0,209	0,0443	22,2	1,232
InternetService::DSL	0,0908	0,0729	1,55	0,2126	1,095
InternetService::Fiber optic	0,7427	0,09	68,14	<0,0001	2,102
OnlineSecurity::No	0,2314	0,0539	18,42	<0,0001	1,26
OnlineSecurity::No internet service	0				
PaperlessBilling::No	-0,2242	0,0479	21,87	<0,0001	0,799
SeniorCitizen	1	-0,1375	0,0531	6,71	0,872
StreamingMovies::No	-0,1123	0,0496	5,13	0,0235	0,894
StreamingMovies::No internet service	0				
TechSupport::No	0,1894	0,0543	12,15	0,0005	1,209
TechSupport::No internet service	0				
TotalCharges	0,000194	0,000083	5,5	0,019	1
tenure	-0,0496	0,00751	43,55	<0,0001	0,952

źródło: opracowanie własne z użyciem SAS Enterprise Miner

Klienci zawierający umowy krótkoterminowe, przedłużane co miesiąc, charakteryzują się wyższym prawdopodobieństwem odejścia niż ci z kontraktami dwuletnimi (o 70,8%) przy pozostałych zmiennych niezmienionych. Podobnie kontrakty roczne - o 9,2% w porównaniu z dwuletnimi. Płatność metodami innymi niż przelew elektroniczny (w tym płatność jako zlecenie stałe) to również jedna z charakterystyk klientów odchodzących. Konsumenci z najszybszymi łączami (światłowód) odchodzą częściej niż ci korzystający z cyfrowej linii abonenckiej oraz częściej niż ci nieposiadający internetu (odpowiednio 2,1 razy częściej oraz 1,1 razy częściej w porównaniu do osób bez łącza ceteris paribus). Brak ochrony w Internecie

oraz niekorzystanie ze wsparcia technicznego to odpowiednio 26% i 20,9% większe szanse na odejście.

Rzadziej z kolei odchodzą konsumenci wybierający elektroniczne faktury (20,1% niższe prawdopodobieństwo w porównaniu z papierowymi fakturami), emeryci (12,8% rzadziej niż osoby przed emeryturą), a także osoby niekorzystające ze streamingu filmów (10,6% rzadziej niż korzystający). Co więcej, z każdym kolejnym miesiącem, w którym klient nie odszedł, prawdopodobieństwo jego odejścia spada o 4,8%.

IV. Drzewa decyzyjne

Budowę modeli drzew decyzyjnych, podobnie jak w poprzednim rozdziale, rozpoczęto od podziału zbioru na część uczącą, walidacyjną i testową w proporcjach odpowiednio 60%, 30% i 10%. W pierwszym kroku trzy drzewa różniące się kryterium wyboru poddrzewa:

- w modelu 1. miarą oceny była decyzja,

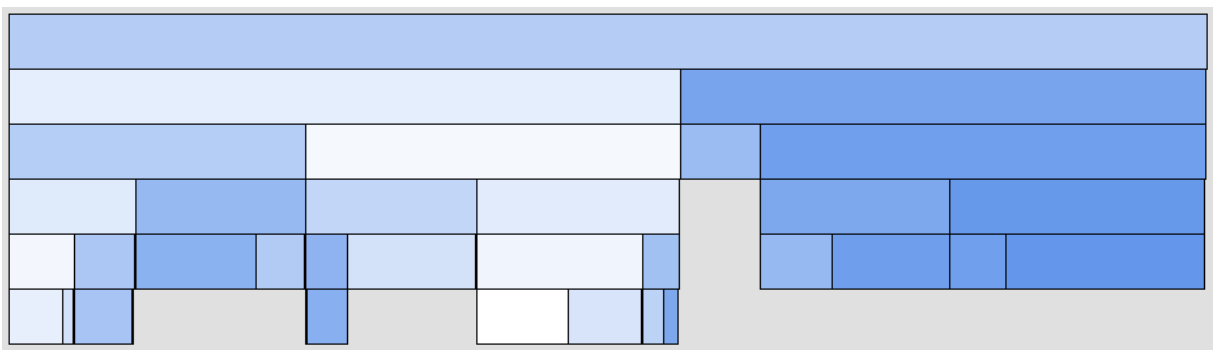
- w modelu 2. błędne klasyfikacje,

- w modelu 3. lift.

Stworzone modele nie różniły się znacznie pod względem wartości kryteriów dopasowania (odsetka błędnych klasyfikacji na zbiorze walidacyjnym). Modele 1. i 2. okazały się identyczne, a model 3. osiągnął nieco lepsze wyniki niż pozostałe dwa w przypadku indeksu ROC (0,839 na zbiorze uczącym i 0,849 na zbiorze testowym w porównaniu do 0,815 i 0,834 dla modeli 1. i 2.).

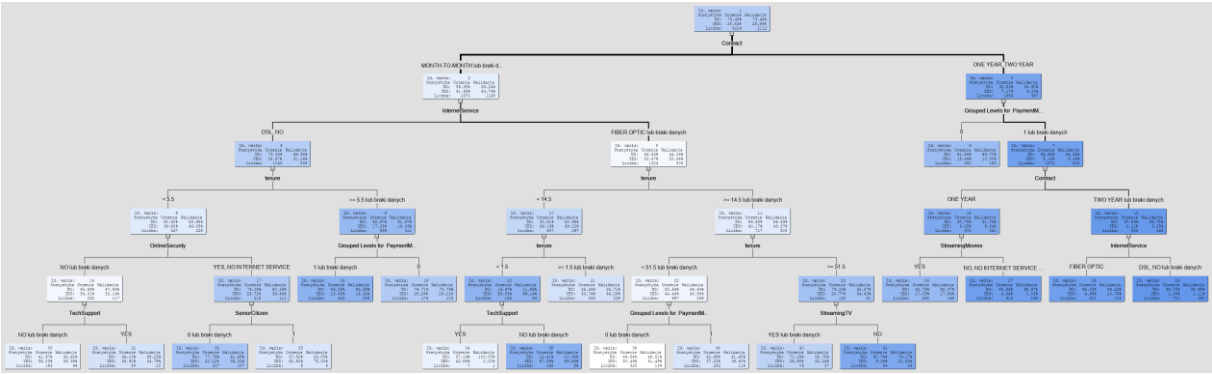
Model 3. charakteryzuje się jednak dużą liczbą klas i liśćmi niewielkich rozmiarów, co przedstawione zostało na rysunkach poniżej.

Rysunek 1. Wykres kafelkowy dla modelu 3.



źródło: opracowanie własne z użyciem SAS Enterprise Miner

Rysunek 2. Wizualizacja drzewa decyzyjnego w modelu 3.



źródło: opracowanie własne z użyciem SAS Enterprise Miner

Niektóre z klas mają liczebność mniejszą niż 10 obserwacji, co może świadczyć o niepotrzebnym skomplikowaniu modelu. W związku z tym stworzony zostaje model 4. z nałożonym ograniczeniem na minimalną wielkość liścia, który powinien przyciąć model 3. do bardziej optymalnej formy. Czwarty model charakteryzuje się niemal takim samym odsetkiem błędnych klasyfikacji na zbiorze testowym co poprzednie trzy modele (różnica jest na poziomie 0,002), ale lepszym wskaźnikiem ROC na zbiorze testowym (0,851 w porównaniu do 0,849 w modelu bez obcięcia).

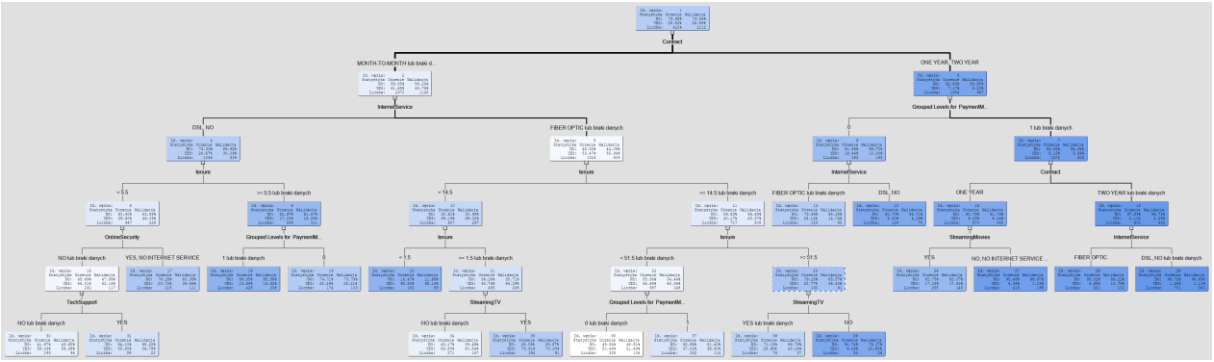
Tabela 5. Statystyki dopasowania modeli drzew decyzyjnych

Model	Odsetek błędnych klasyfikacji			Indeks ROC		
	Uczenie	Walidacja	Testowanie	Uczenie	Walidacja	Testowanie
Drzewo decyzyjne (lift)	0,201	0,203	0,180	0,839	0,838	0,849
Drzewo decyzyjne (decyzja)	0,201	0,203	0,180	0,815	0,823	0,834
Drzewo decyzyjne (bł. klas)	0,201	0,203	0,180	0,815	0,823	0,834
Drzewo decyzyjne (lift z ogr)	0,202	0,205	0,180	0,842	0,836	0,851

źródło: opracowanie własne z użyciem SAS Enterprise Miner

W żadnym z modeli nie występuje znaczna różnica między statystykami dopasowania na zbiorze walidacyjnym i uczącym, więc przeuczenie nie występuje.

Rysunek 3. Wizualizacja drzewa decyzyjnego w modelu 4.



źródło: opracowanie własne z użyciem SAS Enterprise Miner

Nałożone ograniczenie nie wpłynęło na głębokość drzewa, zmniejszyła się jedynie liczba liści (z 17 do 16). Ostateczny model bierze pod uwagę zmienną *Contract, InternetService, PaymentMethods, Tenure, OnlineSecurity, StreamingMovies* i *StreamingTV* - o trzy zmienne mniej niż model regresji logistycznej. Przykładową interpretacją jest, że osoba z kontraktem krótkoterminowym korzystająca ze światłowodu, która jest klientem firmy telekomunikacyjnej krócej niż 1,5 miesiąca odejdzie z prawdopodobieństwem 88%. Podobny klient, który pozostał w firmie przez okres od 14,5 do 51,5 miesiąca oraz korzysta z płatności innej niż jednorazowy przelew elektroniczny (płaci na poczcie lub jako zlecenie stałe) ma już tylko 38,6% szans na odejście.

V. Porównanie modeli

By ostatecznie wybrać model, który najlepiej oddaje zjawisko odejścia z firmy, zebrano wartości statystyk dopasowania dla najlepszych modeli z każdego rozdziału.

Tabela 8. Porównanie statystyk dopasowania.

Opis modelu	Odsetek błędnych klasyfikacji			ROC		
	Uczenie	Walidacja	Testowanie	Uczenie	Walidacja	Testowanie
Regresja logistyczna	0,198	0,194	0,194	0,843	0,838	0,866
Drzewo decyzyjne	0,202	0,205	0,180	0,842	0,836	0,851
Sieć neuronowa	0,194	0,190	0,197	0,851	0,844	0,866

źródło: opracowanie własne z użyciem SAS Enterprise Miner

Wszystkie stworzone modele charakteryzowały się bardzo podobnymi statystykami dopasowania i żaden model nie jest zauważalnie lepszy od innych. Jednak

wykorzystując statystykę dla krzywej ROC dla zbioru testowego, która we wcześniejszych rozdziałach była używana do wybrania najlepszych modeli, można zauważyć, że największą wartością równą 0,866 charakteryzuje się zarówno model regresji logistycznej, jak i model sieci neuronowej. Z tych dwóch modeli to regresja logistyczna odznacza się mniejszym odsetkiem błędnych klasyfikacji na zbiorze testowym.

Ostatecznie, możemy więc wnioskować, że to model regresji logistycznej ze zmodyfikowanymi zmiennymi i stałą, dokładniej opisany i interpretowany w III rozdziale, okazał się najlepszy.

VI. Podsumowanie

Powyższa analiza dotyczyła wykorzystania modeli analitycznych regresji logistycznej, drzew decyzyjnych i sieci neuronowych do wyjaśnienia zjawiska odchodzenia klientów firmy telekomunikacyjnej. Najlepsze wyniki dopasowania osiągnął model regresji logistycznej.

Najsilniejszym wpływem w modelu regresji charakteryzuje się rodzaj łącza internetowego (ponad dwukrotnie wyższe prawdopodobieństwo odejścia dla klientów korzystających ze światłowodu względem osób nieposiadających żadnego łącza) oraz długość zawieranego kontraktu (umowy odnawiane co miesiąc są zrywane z prawdopodobieństwem wyższym o 70% względem umów dwuletnich). Tak znaczny wpływ pierwszej z wymienionych zmiennych może wynikać z ograniczonej oferty firmy telekomunikacyjnej, która nie wykorzystuje w pełni możliwości, jakie daje światłowód (usługi telekomunikacyjne często kupowane są w pakietach, opłacanie jednego rachunku za telefon i Internet jest z pewnością czynnikiem zachęcającym do zamówienia wszystkich usług w jednej firmie). Skłania to klientów do poszukiwania innego dostawcy i przeniesienia również pozostałych usług. Drugi z czynników z najsilniejszym wpływem, czas trwania kontraktu, nie powinien być z biznesowego punktu widzenia rozpatrywany jako przyczyna odejścia. Intuicyjnym jest, że klienci mogący rozwiązać umowę z krótkim okresem wypowiedzenia odejdą do konkurencji wtedy, gdy będzie to najbardziej opłacalne (np. w momencie uruchomienia ciekawej promocji).

Z kolei w grupie czynników, które obniżają prawdopodobieństwo odejścia, znajduje się fakt otrzymywania elektronicznych faktur, bycia emerytem oraz niekorzystania ze streamingu filmów. Ponadto, im dłużej dana osoba pozostaje klientem, tym mniej prawdopodobne, że odejdzie (innymi słowy, prawdopodobieństwo odejścia jest najwyższe na początku współpracy). Tutaj również można doszukiwać się braków w ofercie analizowanej firmy, być

może konkurencja oferuje lepsze pakiety zawierające m.in. streaming filmów, co może mieć wpływ na decyzję o odejściu.

Spis tabel

Spis wykresów

Spis rysunków