Odejście klientów z banku ABC Multinational Bank

Raport zaliczeniowy

Prezentacja i Wizualizacja Danych

Bogdan Yanovich 109072

Warszawa 25.12.2023

Spis treści

Opis zbioru danych

Dane użyte w raporcie pochodzą z ABC Multinational Bank. Dane te są publicznie dostępne na stronie internetowej Kaggle pod poniższym adresem:

https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset

Celem tego raportu jest wizualizacja rozkładu zmiennych ze zbioru danych, wizualizacja zależności między wybranymi zmiennymi objaśniającymi a zmienną churn, stworzenie interaktywnych wykresów za pomocą pakietu Shiny oraz kategoryzowanie klientów ze względu na ich podjęcie lub nie decyzji o rezygnacji z usług banku za pomocą regresji logistycznej i drzewa klasyfikacyjnego. Celem tej informacji jest pomoc bankowi w zmniejszeniu liczby osób, które postanowiły odejść z tego banku poprzez np. zwiększenie działań marketingowych nacelowanych na te osoby.

Zbiór danych obejmuje 10 000 obserwacji i 12 zmiennych, z których jedna pełni funkcję identyfikatora (customer_id), 10 to zmienne objaśniające, a zmienna zależna to 'Churn'. Zmienna 'Churn' przyjmuje wartości binarne: 0 dla klientów, którzy pozostają i kontynuują korzystanie z usług banku, oraz 1 dla klientów, którzy zdecydowali się zrezygnować z usług banku. Poniżej przedstawiona jest tabela zawierająca zmienne występujące w zbiorze danych wraz z ich opisem.

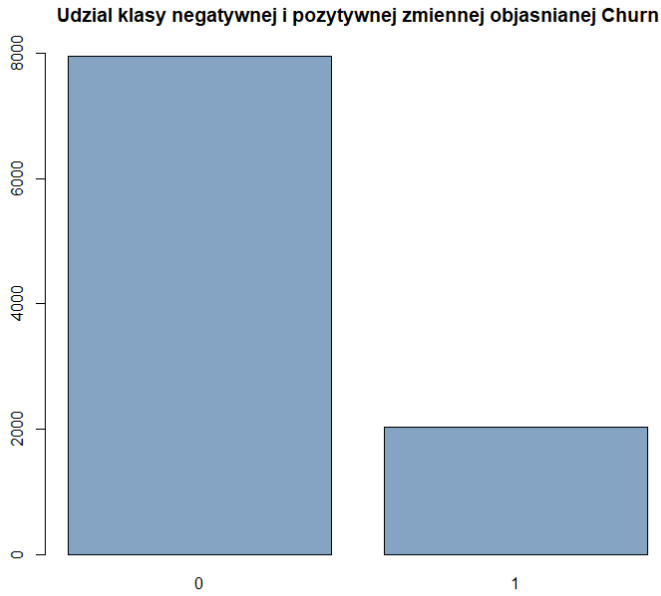
Nazwa zmiennej	Objaśnienie	Typ zmiennej	Przyjmowane wartości
Credit_score	Liczba punktów przypisana przez bank poszczególnym klientom w celu oceny ich zdolności kredytowej	Zmiennoprzecinkowy (num)	Od 350 do 850
Country	Państwo	Znakowy (char)	„France”, „Germany”, „Spain”
Gender	Płeć	Znakowy (char)	„Male”, „Female”
Age	Wiek	Zmiennoprzecinkowy (num)	Od 15 do 95
Tenure	Reprezentuje czas, wyrażony w latach, przez jaki klient utrzymuje konto w danym banku.	Zmiennoprzecinkowy (num)	Od 0 do 10

Balance	Ilość pieniędzy znajdujących się na koncie klienta w danym banku.	Zmiennoprzecinkowy (num)	Od 0 do 250898
Products_number	Informuje o liczbie produktów, jakie klient posiada w danym banku.	Zmiennoprzecinkowy (num)	Od 1 do 4
Credit_card	Posiadanie karty kredytowej	Zmiennoprzecinkowy (num)	1 – „Posiada” 0 – „Nie posiada”
Active_member	Zaangażowanie danego klienta w korzystanie z usług banku	Zmiennoprzecinkowy (num)	1 – „Jest aktywny” 0 – „Nie jest aktywny”
Estimated_salary	Przybliżona płaca danego klienta	Zmiennoprzecinkowy (num)	Od 11.58 do 199992.48
Churn	Czy klient zdecydował się na rezygnację z usług banku	Zmiennoprzecinkowy (num)	1 – „Zrezygnował” 0 – „Pozostał”

Źródło: opracowanie własne

Z powyższej tabeli można wywnioskować, że większość zmiennych w modelu ma typ zmiennoprzecinkowy. Dla zmiennych takich jak 'churn', 'active_member' czy 'credit_card', przyjmujących wartości 0 lub 1, ten typ danych może być niezbyt odpowiedni. W kolejnych częściach raportu przeprowadzono przekształcenie niektórych zmiennych na inny typ. Dla innych zmiennych, które będą wykorzystywane do zbudowania regresji logistycznej, utworzono tzw. dummy variables. W analizowanym zbiorze danych nie stwierdzono braków danych.

Poniżej przedstawiono wykres przedstawiający udział klasy negatywnej i pozytywnej zmiennej objaśnianej ‘Churn’.



Źródło: opracowanie własne

Wartość „0” jest przyjmowana przez 7963 obserwacji - klient pozostał w banku w prawie 80% przypadków. Wartość „1” jest przyjmowana przez 2037 obserwacji - klient zrezygnował z usług banku w prawie 20% przypadków.

Wykres sugeruje, że mamy do czynienia z niezbilansowanym zbiorem danych, co oznacza, że może być konieczne skorzystanie z pewnych technik w celu uwzględnienia potencjalnych wyzwań związanych z niewystarczającym wykrywaniem klasy pozytywnej, tj. klientów rezygnujących z usług banku w następnych częściach raportu.

Przekształcenia zmiennych

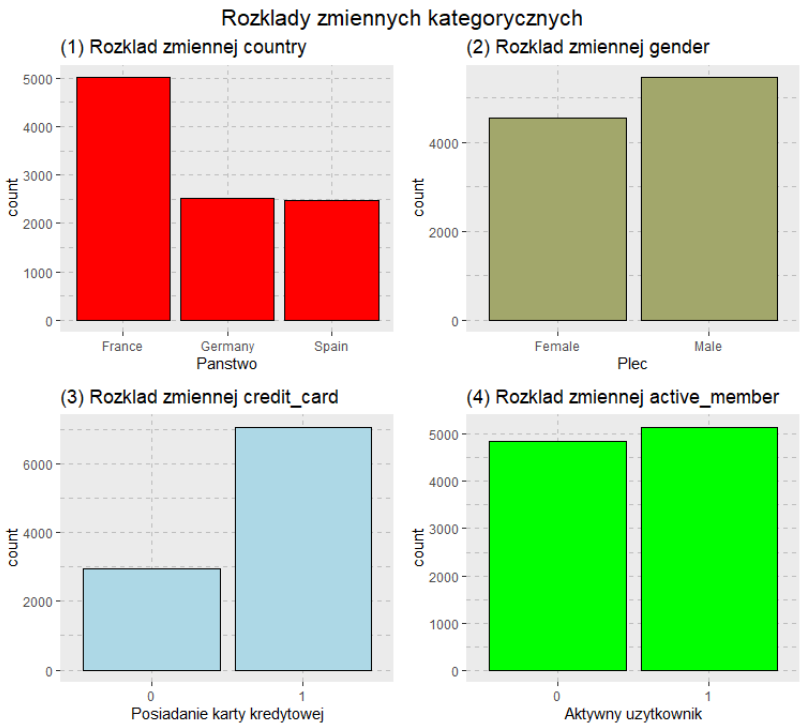
W pierwszej kolejności postanowiono usunąć zmienna „customer_id”, która nie jest wykorzystywana w części analizy danych oraz modeli uczenia maszynowego. W celu zniwelowania niezbilansowania w zbiorze danych zastosowano technikę undersampling, polegającą na losowym usunięciu w zbiorze testowym części obserwacji z klasy o większej liczbie wystąpień, tj. '0'.

Zdecydowano się również przekształcić zmienną objaśnianą 'churn' oraz zmienne objaśniające 'active_member', 'credit_card', ‘gender’ ‘country’ na typ kategoryczny (faktorowy). Zmienne 'products_number', 'tenure' zostały przekształcone na typ całkowity (int).

Wykresy w części analizy danych zostały stworzone na podstawie oryginalnego zbioru danych. Natomiast do treningu modeli uczenia maszynowego wykorzystano zbiór danych z zastosowanym undersamplingiem oraz zmiennymi dummy variables.

Analiza zbioru danych

W tej części analizy przedstawiono wykresy rozkładu wszystkich zmiennych ze zbioru danych. Wykresy słupkowe zostały użyte do prezentacji rozkładów zmiennych kategorycznych, natomiast wykresy pudełkowe zostały zastosowane dla zmiennych numerycznych.



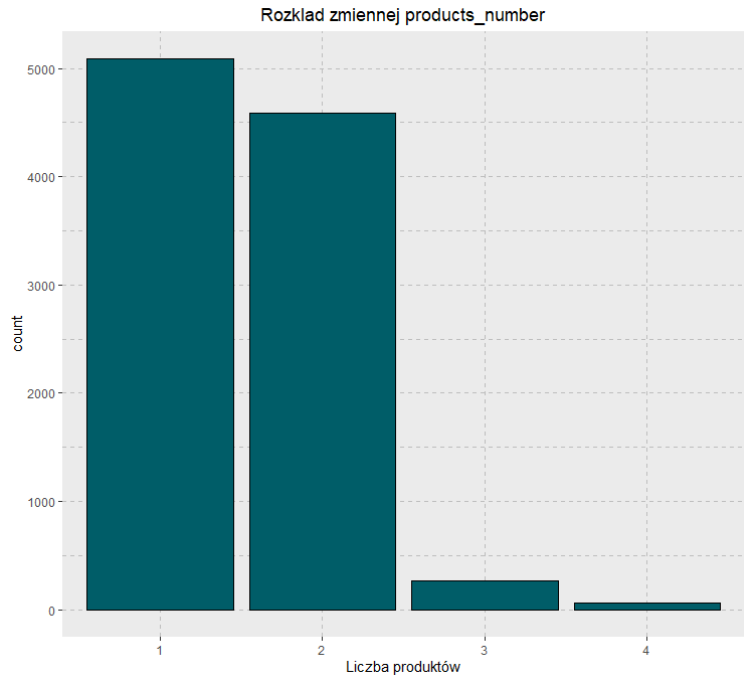
Źródło: opracowanie własne

1) Zmienna 'country' (państwo) przyjmuje wartości kategoryczne, reprezentujące trzy kraje: Francję, Niemcy oraz Hiszpanię. Podział udziałów tych krajów w zbiorze danych jest następujący:
    - We Francji zarejestrowanych jest 50.14% klientów spośród 10,000 osób.
    - Niemieccy klienci stanowią 24.77% wszystkich klientów ABC Multinational Bank.
    - W Hiszpanii udział klientów wynosi 25.09% osób.
2) Zmienna 'gender' (płeć) jest zmienną kategoryczną, gdzie kategorie to 'Male' (mężczyzna) lub 'Female' (kobieta). Na przedstawionym wykresie można zauważyć, że

liczba klientów deklarujących przynależność do płci męskiej jest trochę większa niż liczba klientów-kobiet.

3) Zmienna 'credit_card' (karta kredytowa) przyjmuje wartości kategoryczne, gdzie '0' oznacza brak posiadania karty kredytowej, a '1' wskazuje, że klient banku jest posiadaczem karty kredytowej. Analizując powyższy wykres, można zauważyć, że zdecydowana większość klientów ABC Multinational Bank posiada kartę kredytową.
4) Zmienna 'active_member' jest kategoryczna, gdzie wartość '1' oznacza zaangażowanie klienta w korzystanie z usług banku, a '0' wskazuje na brak zaangażowania. Jak można zauważyć na przedstawionym wykresie, liczba klientów zaangażowanych oraz niezaangażowanych jest prawie taka sama.

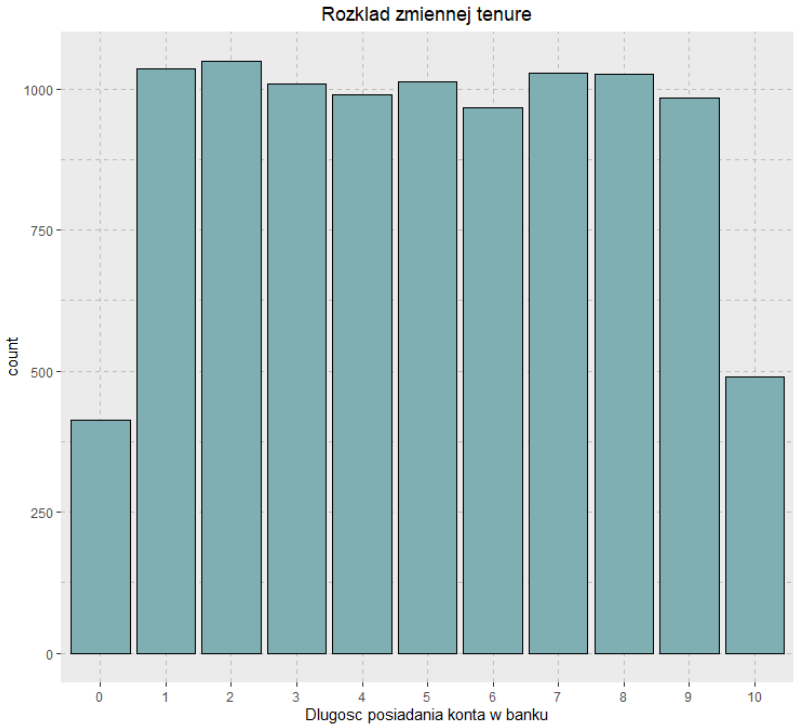
Naszkicowano również wykres rozkładu zmiennej products_number, który znajduje się poniżej.



Źródło: opracowanie własne

Zmienna 'products_number' odzwierciedla liczbę produktów, które dany klient posiada w banku. Jest to zmienna całkowita, przyjmująca wartości {1, 2, 3, 4}. Z analizy przedstawionego wykresu wynika, że zdecydowana większość klientów tego banku posiada 1 lub 2 produkty. Ponadto, jedynie 60 klientów (0.6%) korzysta z czterech różnych produktów oferowanych przez ten bank.

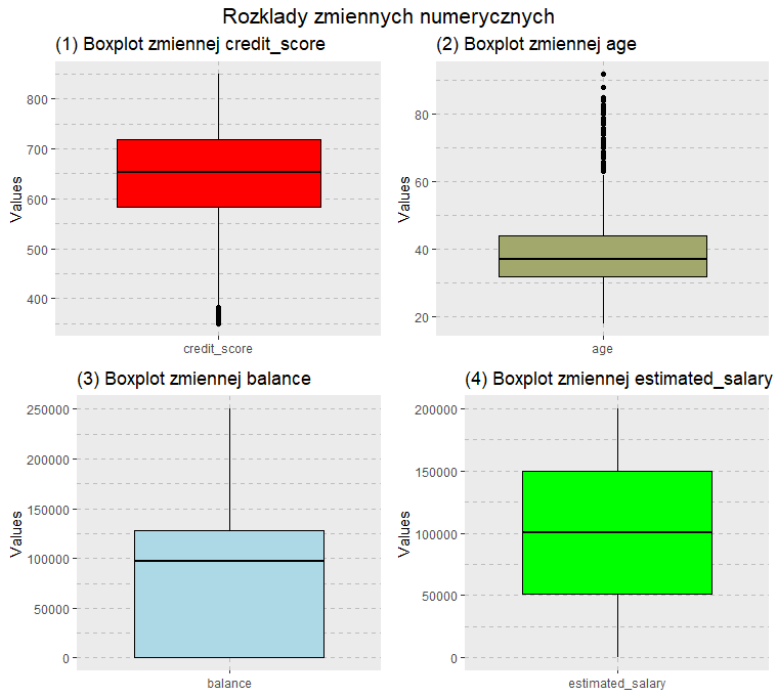
Poniżej zamieszczono wykres przedstawiający rozkład zmiennej 'tenure'.



Źródło: opracowanie własne

Zmienna 'tenure' ma typ całkowity i przedstawia, jak długo (w latach) klient posiada konto w banku ABC International. Analiza przedstawionego wykresu wskazuje, że, pomijając pierwszą i ostatnią grupę, udziały klientów w pozostałych grupach są rozłożone mniej więcej równomiernie.

W kolejnej części raportu skupiono się na wizualizacji rozkładów zmiennych numerycznych za pomocą wykresów pudełkowych. Wykres pudełkowy (boxplot) przedstawia kilka istotnych statystyk opisowych, takich jak mediana, kwartyle oraz potencjalne wartości odstające. Poniżej przedstawiane są wyniki tej analizy.

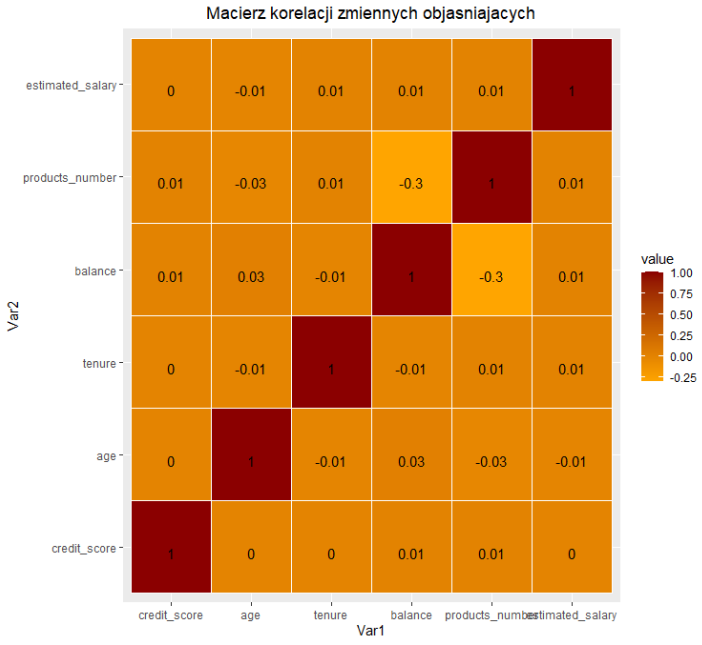


Źródło: opracowanie własne

1) Zmienna "credit_score" ma typ zmiennoprzecinkowy i reprezentuje liczbę punktów przyznanych klientom przez bank które są używane do oceny wiarygodności klienta. Bank może wykorzystać tą ocenę, aby określić poziom zaufania do klienta i zaoferować mu różne usługi finansowe. Im większa jest liczba punktów, tym większe zaufanie i lepsze warunki mogą być dostępne dla klienta. Mediana wynosi 652 punkty, co oznacza, że połowa klientów ma wynik mniejszy lub równy 652 punktów, a druga połowa ma wynik większy lub równy 652 punkty. Pierwszy i trzeci kwartyle wynoszą odpowiednio 584 i 718 punktów. Na wykresie można również zauważyć wartości odstające, gdzie 'credit_score' przyjmuje wartości mniejsze niż 400 punktów.
2) Kolejną zmienną do omówienia jest zmienna 'age', reprezentująca wiek klientów. Jest to zmienna zmiennoprzecinkowa o zakresie od 15 do 95 lat. Mediana wynosi 37 lat. Warto zauważyć, że są obserwowane wartości odstające (outliery), które występują dla wieku 60 lat i więcej.
3) Zmienna 'balance' reprezentuje liczbę pieniędzy na koncie bankowym klienta. Jest to zmienna zmiennoprzecinkowa, przyjmująca wartości od 0 do 250898. Mediana tej zmiennej wynosi 76486, a pierwszy kwartyl to 0, a trzeci kwartyl to 127644. To oznacza, że 25% klientów ma saldo równe 0, a 75% klientów ma saldo równe lub mniejsze niż 127644 jednostki pieniężne.

4) Ostatnią zmienną do omówienia jest 'estimated_salary', reprezentująca przybliżoną płacę danego klienta. Jest to zmienna zmiennoprzecinkowa, przyjmująca wartości na przedziale od 11.58 do 199992.48. Mediana tej zmiennej wynosi 100193.91. To oznacza, że połowa klientów ma szacowaną płacę mniejszą lub równą 100193.91 jp., a druga połowa ma płacę większą lub równą tej wartości.

W trakcie analizy danych przeprowadzono badanie korelacji między zmiennymi zmiennoprzecinkowymi i całkowitymi w zbiorze. Celem było zrozumienie stopnia wzajemnego powiązania między tymi zmiennymi. Wartości korelacji mieszczą się w zakresie od -1 do 1, gdzie wartość bliska 1 wskazuje silną dodatnią korelację, wartość bliska -1 wskazuje silną ujemną korelację, a wartość bliska 0 oznacza brak lub bardzo słabą korelację między danymi zmiennymi.



Źródło: opracowanie własne

Analizując powyższą macierz korelacji, można zauważyć, że prawie wszystkie zmienne numeryczne wykazują bardzo słabe wzajemne powiązania. Tylko pomiędzy zmiennymi 'products_number' a 'balance' występuje ujemna korelacja o wartości -0.3. Choć jest to zależność ujemna, to jednak jej siła jest niewielka, co sugeruje, że zmienne te nie są silnie skorelowane ze sobą. Wartości korelacji bliskie zeru wskazują na brak wyraźnych wzajemnych zależności między analizowanymi zmiennymi.

W ramach raportu obliczono również wskaźnik Information Value (IV), który dostarcza cennych informacji dotyczących istotności poszczególnych zmiennych w kontekście

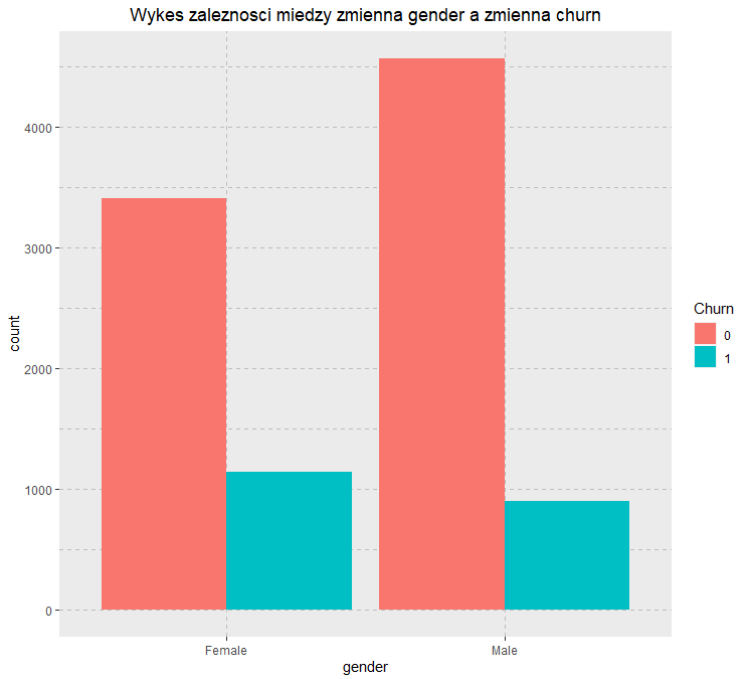
predykcji. Im większa wartość IV, tym bardziej istotna jest zmienna w przewidywaniu churnu klienta.

	Variable	IV
4	age	0.7840697826
7	products_number	0.2215301088
2	country	0.1684189706
9	active_member	0.1532326369
6	balance	0.0977340885
3	gender	0.0697748761
1	credit_score	0.0109597678
5	tenure	0.0085964872
10	estimated_salary	0.0023099724
8	credit_card	0.0003126349

Źródło: opracowanie własne

Spośród analizowanych zmiennych największy wpływ na model mają zmienne takie jak 'age' (IV = 0.784), 'products_number' (IV = 0.2215) oraz 'country' (IV = 0.1684). Wartości te sugerują, że te zmienne posiadają znaczącą zdolność predykcyjną w kontekście zdarzenia docelowego. Z drugiej strony, zmienne 'estimated_salary' (IV = 0.0023) oraz 'credit_card' (IV = 0.000312) wykazują minimalną zdolność predykcyjną. Ich niskie wartości IV sugerują, że te zmienne mają ograniczony wpływ na predykcje modelu.

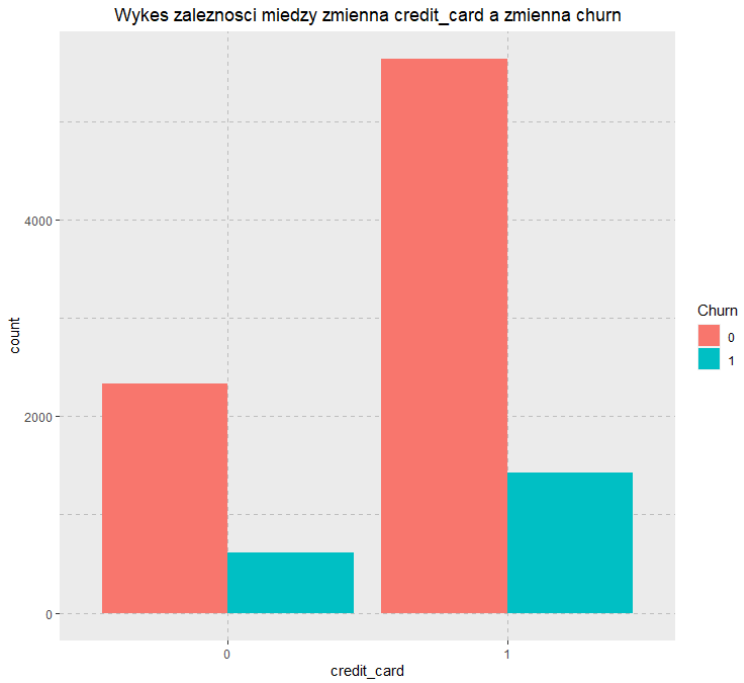
W następnej części raportu przedstawiono wykresy zależności między zmienną objaśnianą 'churn' a zmiennymi objaśniającymi: 'gender', 'credit_card', 'products_number' oraz 'country'.



Źródło: opracowanie własne

Powyższy wykres pokazuje, że liczba mężczyzn, którzy zrezygnowali z usług banku wśród ogółu mężczyzn, jest niższa niż liczba kobiet, które zrezygnowały z usług wśród ogółu kobiet. Co więcej, mimo że liczba mężczyzn w zbiorze danych jest większa, to jednak liczba kobiet, które zrezygnowały z usług banku, przewyższa liczbę mężczyzn którzy podjęli decyzję o rezygnacji z usług banku.

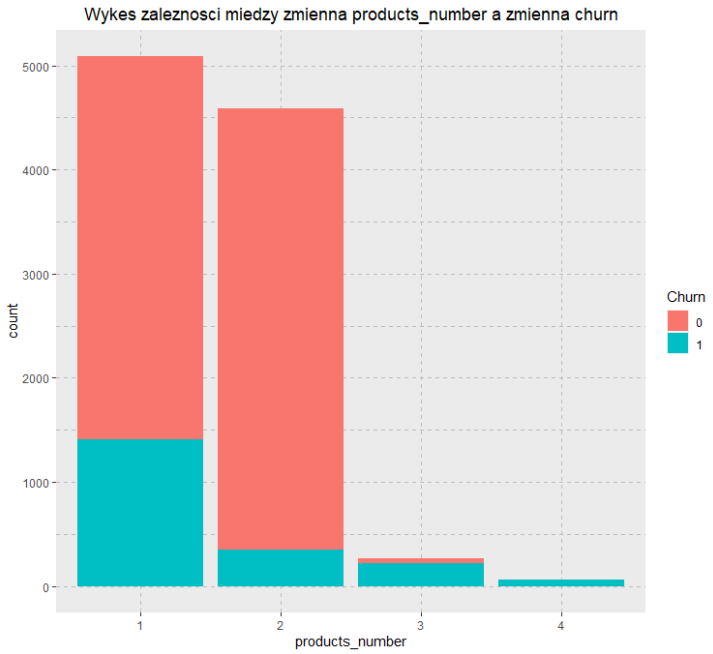
Na poniższym wykresie przedstawiono zależność między zmienną credit_card a zmienną churn.



Źródło: opracowanie własne

Z analizy powyższego wykresu nie można wyciągnąć jednoznacznych wniosków dotyczących zależności między zmiennymi 'credit_card' a 'churn'. Co więcej, przeprowadzone obliczenia wykazały, że udział klasy pozytywnej (klienci odchodzący z banku) różni się minimalnie (różnica wynosi około 2%) między klientami posiadającymi a nieposiadającymi kartę kredytową. Oznacza to, że ta zmienna, czyli posiadanie karty kredytowej, prawdopodobnie nie ma znaczącego wpływu na zjawisko odchodzenia klientów.

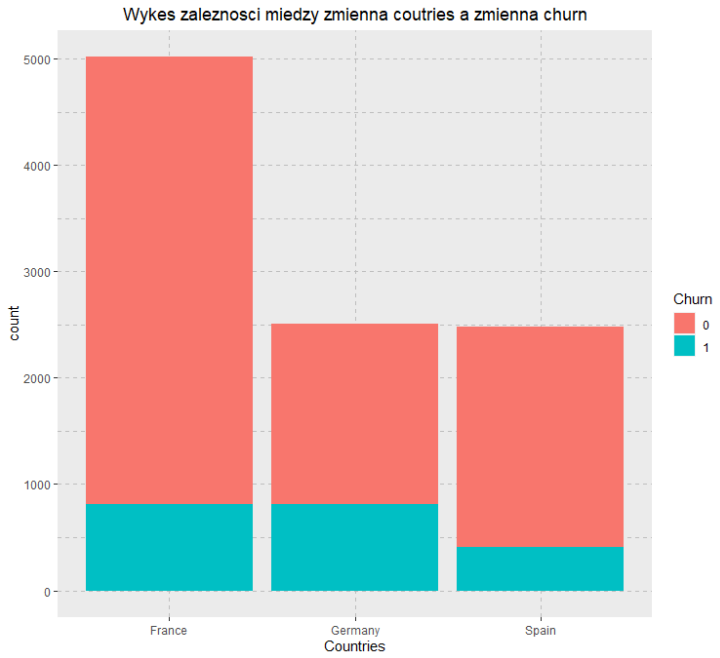
Na poniższym wykresie przedstawiono zależność między zmienną products_number a zmienną churn za pomocą skumulowanego wykresu słupkowego.



Źródło: opracowanie własne

Na podstawie analizy powyższego wykresu można wyciągnąć kilka istotnych wniosków. Pierwszym spostrzeżeniem jest, że zdecydowana większość klientów korzysta z 1 lub 2 produktów oferowanych przez bank ABC International. Co więcej, największy odsetek klientów, którzy zrezygnowali z usług banku (klasa pozytywna), przypada na tych posiadających 4 produkty. Prawdopodobieństwo rezygnacji z banku dla tej grupy wynosi 1, co oznacza, wszystkie 60 osób odeszły z banku. Możliwe jest, że rosnąca atrakcyjność ofert innych banków stanowi główny czynnik decydujący o odejściu klientów. W takim scenariuszu można przypuszczać, że klienci, skorzystając już z pełnej gamy oferowanych produktów przez obecny bank, stają się bardziej podatni na korzyści płynące z atrakcyjnych ofert konkurentów.

Poniżej przedstawiono kolejny skumulowany wykres słupkowy ilustrujący zależność między zmienną 'countries' a zmienną 'churn'.



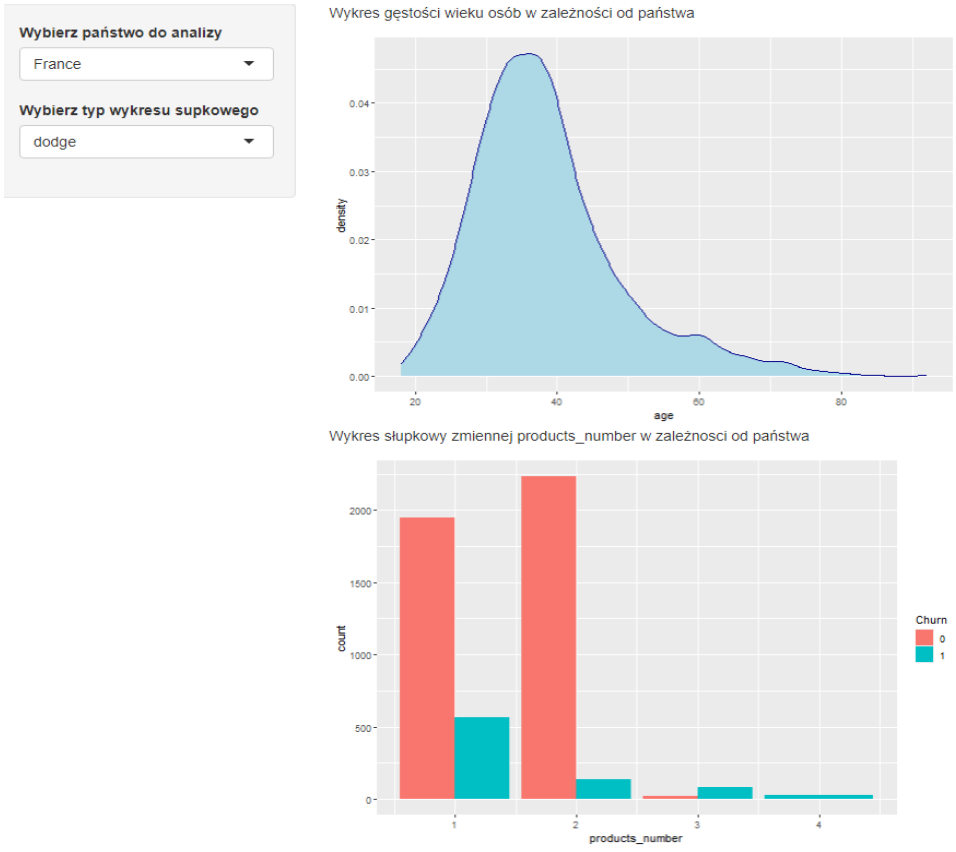
Źródło: opracowanie własne

Kolorem niebieskim na powyższym wykresie oznaczeni są klienci, którzy podjęli decyzję o zrezygnowaniu z usług ABC International Bank. Warto zauważyć, że pomimo około dwukrotnej różnicy w liczbie klientów we Francji i Niemczech, liczba klientów rezygnujących w Niemczech jest taka sama jak we Francji. To istotne spostrzeżenie może stanowić wskazówkę dla banku, sugerując, że implementacja rozwiązań anti-churnowych może być szczególnie istotna w lokalizacji o wyższym odsetku rezygnacji, czyli w Niemczech.

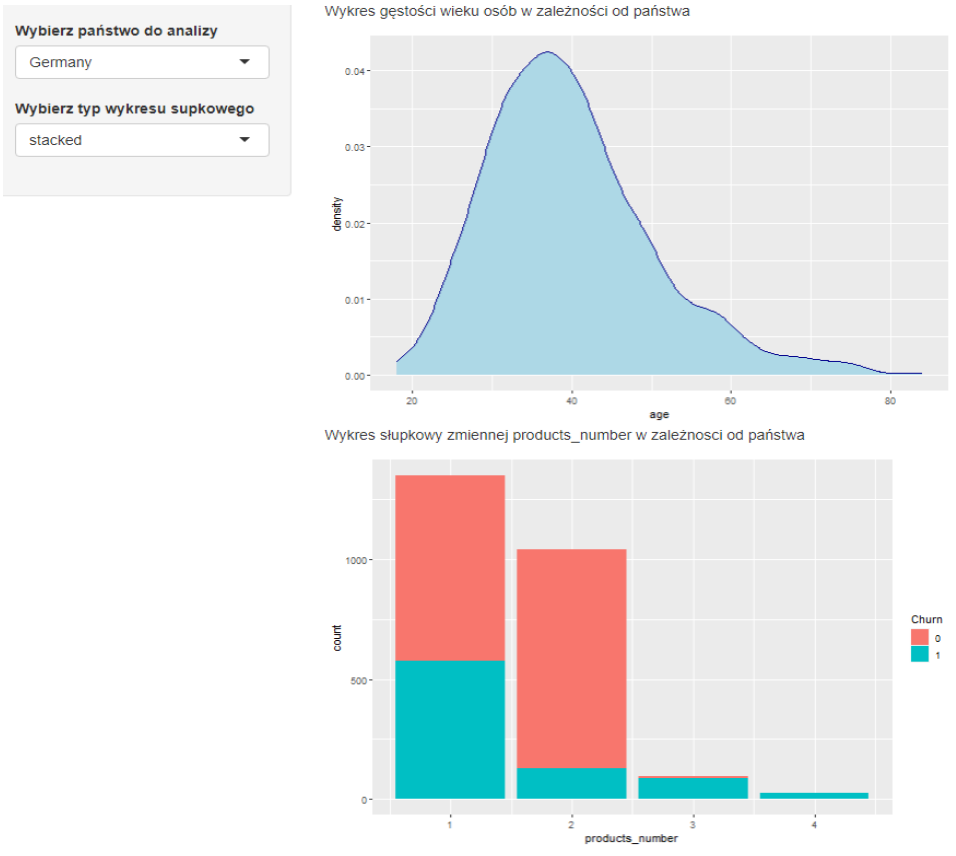
Wykorzystanie pakietu Shiny

Za pomocą pakietu Shiny zbudowano interaktywne wykresy gęstości wieku osób oraz wykresy słupkowe zmiennej products_number w zależności od państwa. Istnieje również możliwość wyboru typu histogramu ("dodge", "stacked", "fill", "identity"). Poniżej znajdują się 3 przykłady wykresów. Wykresy są dostępne pod poniższym linku:

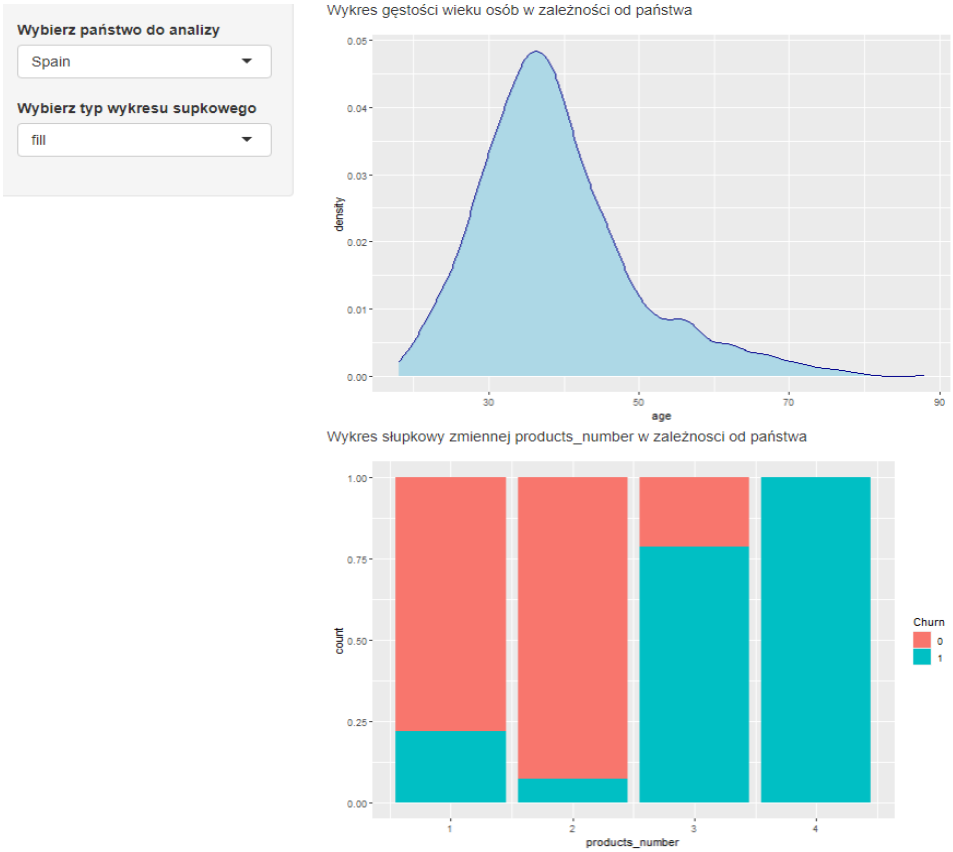
https://bogdanyanovich.shinyapps.io/PIWD/



Źródło: opracowanie własne



Źródło: opracowanie własne



Źródło: opracowanie własne

Z powyższych wykresów gęstości można wywnioskować, że wykresy gęstości wieku osób ze zbioru danych w przypadku trzech państw prawie się nie różnią. Warto jednak zauważyć, że wśród osób we Francji są klienci banku w wieku 20 lat, podczas gdy najmłodsi klienci w Niemczech i Hiszpanii mają odpowiednio 23 i 25 lat. Wykresy słupkowe zmiennej products_number są również podobne: we wszystkich trzech histogramach liczba klientów, które posiadają 1 lub 2 produkty w tym banku jest znacznie wyższa niż liczba klientów posiadających 3 lub 4 produkty. Warto jednak zaznaczyć, że w przypadku Niemiec odsetek klientów posiadających 2 produkty w banku wśród wszystkich klientów z tego kraju jest niższy niż we Francji oraz Hiszpanii.

Modele uczenia maszynowego

W celu klasyfikacji rezygnacji klientów z usług banku zastosowano regresję logistyczną oraz drzewo klasyfikacyjne. Zbiór danych został losowo podzielony na część treningową (75% obserwacji) i testową (25% obserwacji).

W celu zniwelowania niezbilansowania zbioru danych, zastosowano technikę undersampling na zbiorze uczącym, która polega na usunięciu losowo wybranych rekordów z

próbki większościowej (churn = 0), aż obie klasy uzyskały równą liczebność. W zmodyfikowanym zbiorze treningowym 1553 osoby nie zrezygnowały, a 1525 osób zrezygnowało z usług banku ABC International bank. Przeprowadzenie Undersamplingu tylko na zbiorze uczącym było świadomym wyborem, mającym na celu zachowanie rzeczywistego rozkładu klas w zbiorze testowym. Dzięki temu chciano zapewnić, że zbiór testowy odzwierciedli prawdziwy rozkład klas, z którym model będzie się spotykał w warunkach rzeczywistych. Undersampling zbioru testowego mógłby wprowadzić zniekształcenia i prowadzić do błędnej oceny rzeczywistej wydajności modelu.

Modele zostały zbudowane na zmodyfikowanym zbiorze treningowym. Skuteczność modeli została oceniona na zbiorze testowym. Dla zmiennych 'country', 'gender', "products_number" oraz "tenure" utworzono dummy variables.

Regresja logistyczna

Pierwszym modelem, który został zbudowany, jest regresja logistyczna. Poniżej przedstawiono tabelę oszacowanych parametrów.

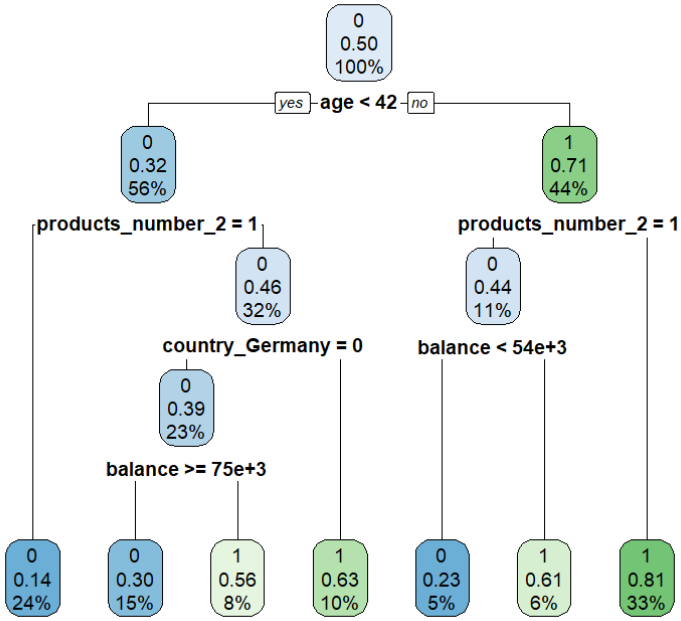
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9377  -0.8014  -0.2389   0.8128   2.8925

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.379e+00  4.252e-01  -3.244  0.00118 **
credit_score    -7.008e-04  4.543e-04  -1.542  0.12298
age              7.534e-02  4.662e-03  16.161  < 2e-16 ***
balance         -2.088e-07  8.106e-07  -0.258  0.79677
credit_card     -1.727e-01  9.790e-02  -1.764  0.07770 .
active_member   -9.905e-01  9.068e-02 -10.924  < 2e-16 ***
estimated_salary 4.791e-08  7.767e-07   0.062  0.95082
country_Germany  1.026e+00  1.138e-01   9.017  < 2e-16 ***
country_Spain    4.936e-02  1.111e-01   0.444  0.65692
gender_Male     -5.455e-01  8.958e-02  -6.089 1.13e-09 ***
products_number_2 -1.529e+00  1.005e-01 -15.218  < 2e-16 ***
products_number_3  2.733e+00  3.797e-01   7.199 6.08e-13 ***
products_number_4  1.579e+01  3.289e+02   0.048  0.96172
tenure_1        -3.529e-01  2.443e-01  -1.444  0.14867
tenure_2        -4.630e-01  2.417e-01  -1.916  0.05542 .
tenure_3        -3.258e-01  2.487e-01  -1.310  0.19022
tenure_4        -1.562e-01  2.486e-01  -0.628  0.52996
tenure_5        -3.293e-01  2.486e-01  -1.325  0.18528
tenure_6        -2.198e-01  2.434e-01  -0.903  0.36639
tenure_7        -4.372e-01  2.475e-01  -1.766  0.07735 .
tenure_8        -4.226e-01  2.449e-01  -1.726  0.08444 .
tenure_9        -4.707e-01  2.464e-01  -1.910  0.05607 .
tenure_10       -1.613e-01  2.916e-01  -0.553  0.58014
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4266.8  on 3077  degrees of freedom
Residual deviance: 3058.3  on 3055  degrees of freedom
AIC: 3104.3

Number of Fisher Scoring iterations: 15
```

Źródło: opracowanie własne

Z powyższego oszacowania możemy wywnioskować, że zmienne age, active_member, country_Germany, gender_Male, products_number_2 i products_number_3 są istotne statystycznie na poziomie istotności mniejszym niż 0.001. Zmienne credit_card, tenure_2, tenure_7, tenure_8, tenure_9 są istotne statystycznie na poziome istotności 0.05.

Drzewo klasyfikacyjne

W związku z dużą liczbą liści o niewielkim udziale w całej populacji w pierwotnie zbudowanym drzewie klasyfikacyjnym, zdecydowano się wprowadzić ograniczenie na minimalną wielkość liścia. Z uwagi na obfitość rekordów w danych, ustalono ograniczenie 'minbucket = 100'. Małe liście są bardziej podatne na wychwytywanie szumu z danych i mogą być przyczyną przetrenowania modelu.



Źródło: opracowanie własne

Podano również 2 interpretacje reguł decyzyjnych wychodzących z drzewa klasyfikacyjnego:

1) Jeśli klient korzystający z usług banku ma mniej niż 42 lata i posiada 2 produkty w tym banku, to prawdopodobieństwo, że zdecyduje się na opuszczenie tego banku, wynosi 14%.
2) Jeśli klient korzystający z usług banku ma 42 lata lub więcej, posiada 2 produkty w tym banku, a saldo na koncie wynosi 54,000 (54e+3) jednostek pieniężnych lub więcej, to prawdopodobieństwo, że zdecyduje się na opuszczenie tego banku, wynosi 61%.

Poniżej znajduje się tabela istotności zmiennych w modelu drzewa decyzyjnego.

age	products_number_2	balance	country_Germany	products_number_3
232.4952604	157.8939797	71.3207623	35.1657040	6.8200544
products_number_4	country_Spain	tenure_10	estimated_salary	credit_score
3.3210188	2.9726283	1.6514602	1.3537050	1.2808002
active_member	tenure_9			
0.6605841	0.5108175			

Źródło: opracowanie własne

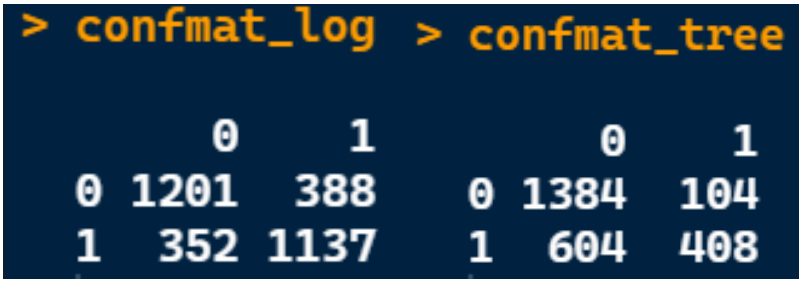
Najbardziej istotnymi zmiennymi w modelu drzewa decyzyjnego są zmienne 'age' oraz 'products_number_2'.

Ocena jakości modeli

Po trenowaniu modeli uczenia maszynowego należy sprawdzić jakość predykcji każdego modelu i wybrać jeden który ma największą siłę predykcyjną. W tym celu porównano modele, używając statystyk pochodnych. W raporcie zostały wykorzystane następujące statystyki wraz z ich interpretacją:

- Accuracy (dokładność) – stosunek prawidłowo zaklasyfikowanych obserwacji do wszystkich obserwacji.
- Match Error Rate (MER) – stosunek nieprawidłowo zaklasyfikowanych obserwacji do wszystkich obserwacji.
- Precision (rzetelność) – stosunek prawidłowo zaklasyfikowanych obserwacji do klasy pozytywnej do wszystkich obserwacji zaklasyfikowanych do klasy pozytywnej.
- Sensitivity (czułość) – stosunek prawidłowo zaklasyfikowanych obserwacji do klasy pozytywnej do sumy obserwacji prawidłowo zaklasyfikowanych do klasy pozytywnej i obserwacji nieprawidłowo zaklasyfikowanych do klasy negatywnej.
- Specificity (swoistość) – stosunek prawidłowo zaklasyfikowanych obserwacji do klasy negatywnej do wszystkich obserwacji zaklasyfikowanych do klasy negatywnej.
- F1 Score – średnia harmoniczna pomiędzy rzetelnością (precision) i czułością (sensitivity).

W ramach raportu zostały stworzone tablice pomyłek dla każdego modelu. Poniżej znajdują się tablice pomyłek dla modeli regresji logistycznej oraz drzewa klasyfikacyjnego.



Źródło: opracowanie własne

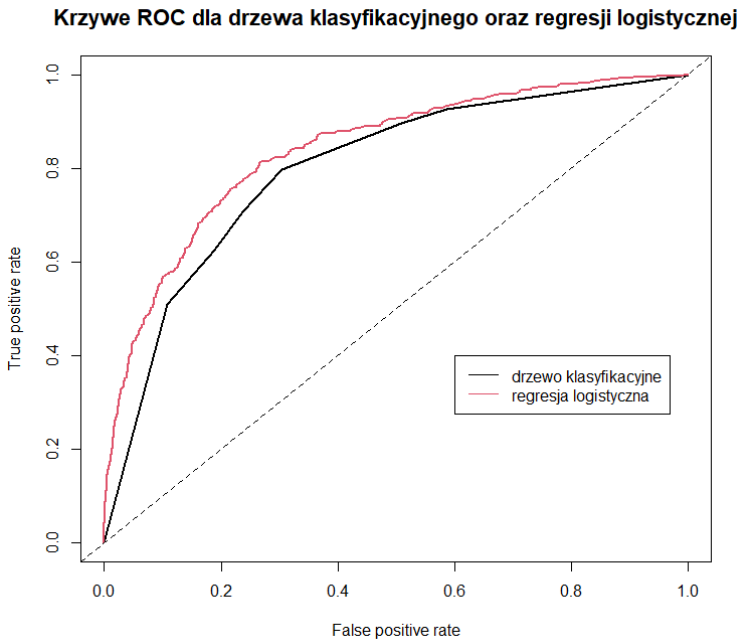
Po trenowaniu modeli na zmodyfikowanym zbiorze treningowym przeprowadzono predykcję zmiennej churn na zbiorze testowym i obliczono statystyki pochodne dla każdego modelu. W poniższej tabeli są przedstawione wartości statystyk pochodnych dla każdego modelu.

	tree	under_reg_log
accuracy	0.7168	0.7595841
MER	0.2832	0.2404159
precision	0.4031621	0.7635997
sensitivity	0.796875	0.7455738
specificity	0.6961771	0.7733419
F1	0.5354331	0.7544791

Źródło: opracowanie własne

Z analizy danych przedstawionych w powyższej tabeli wynika, że model regresji logistycznej zbudowany na zbiorze z undersamplingiem charakteryzuje się lepszą jakością predykcji w niemal wszystkich statystykach w porównaniu do modelu drzewa klasyfikacyjnego. Jedynym wyjątkiem jest czułość, gdzie model drzewa klasyfikacyjnego osiąga lepsze wyniki.

Innym narzędziem do oceny poprawności klasyfikatora jest krzywa ROC, która pokazuje jak zmienia się True Positive Rate w zależności od False Positive Rate. Poniżej są przedstawione wykresy krzywych ROC dla każdego modelu.



Źródło: opracowanie własne

W raporcie obliczono również statystykę AUC (Area Under the Curve) dla krzywych ROC każdego modelu w celu sprawdzenia, jaki model jest najlepszy pod względem pola pod krzywą ROC. W poniższej tabeli znajdują się wartości statystyk AUC ROC dla każdego modelu.

Drzewo klasyfikacyjne	Regresja logistyczna
0.79833	0.8388

Źródło: opracowanie własne

Z powyższej tabeli wynika, że największą wartością Area under the Curve cechuje się model regresji logistycznej (0.8388).

Wyniki i podsumowanie

Raport miał na celu wizualizację rozkładu zmiennych ze zbioru danych, wizualizację zależności między zmiennymi objaśniającymi a zmienną 'churn', stworzenie interaktywnych wizualizacji z wykorzystaniem pakietu Shiny oraz prognozowanie wystąpienia zjawiska churnu klienta za pomocą metod uczenia maszynowego. W tym celu przedstawiono wykresy rozkładu zmiennych, wykresy zależności między zmiennymi, wykresy z zastosowaniem pakietu Shiny oraz skonstruowano dwa modele: drzewo klasyfikacyjne oraz regresję logistyczną. Warto zaznaczyć, że w pracy wykorzystano zmodyfikowany zbiór treningowy z undersamplingiem i dummy variables.

Po przetrenowaniu modeli na zmodyfikowanym zbiorze treningowym i prognozowaniu na zbiorze testowym, porównano siłę predykcyjną modeli za pomocą tablicy pomyłek, statystyk pochodnych, krzywej ROC oraz wartości AUC. W zdecydowanej większości stosowanych metod oceny jakości, model regresji logistycznej okazał się być najlepszy, osiągając najwyższe statystyki pochodne oraz największe pole pod krzywą ROC.

Analizując istotność zmiennych w obu modelach, najważniejszymi zmiennymi okazały się: age, products_number_2, oraz balance. W przypadku modelu drzewa klasyfikacyjnego, oprócz wyżej wymienionych, istotne okazały się również zmienne 'country_Germany', 'products_number_3', 'products_number_4', oraz 'country_Spain'. W modelu regresji logistycznej innymi istotnymi zmiennymi okazały się 'active_member', 'country_Germany', 'gender_Male', oraz 'products_number_3'.