**Odejście klientów z banku ABC Multinational Bank**

**Raport zaliczeniowy**

**Indukowane reguły decyzyjne**

Aliaksandr Mazur 109081

Aliaksandr Zaman 109163

Andrei Tezhyk 101032

Bogdan Yanovich 109072

Yaraslau Shemet 108887

Spis treści

[Podsumowanie 3](#_Toc125315989)

[Problem decyzyjny 3](#_Toc125315990)

[Zbiór danych 3](#_Toc125315991)

[Uzyskany wynik 3](#_Toc125315992)

[Opis głównych wyników 3](#_Toc125315993)

[Dane 4](#_Toc125315994)

[Zmienna objaśniana 4](#_Toc125315995)

[Analiza danych 4](#_Toc125315996)

[Zmienne objaśniające i ich przekształcenia 5](#_Toc125315997)

[Korelacja zmiennych objaśniających 14](#_Toc125315998)

[Siła predykcyjna 15](#_Toc125315999)

[Modele 15](#_Toc125316000)

[Regresja logistyczna 16](#_Toc125316001)

[Drzewo klasyfikacyjne 18](#_Toc125316002)

[Las losowy 20](#_Toc125316003)

[Ocena jakości modelu 21](#_Toc125316004)

[Wstęp 21](#_Toc125316005)

[Modele zbudowane na niezmodyfikowanym zbiorze danych 22](#_Toc125316006)

[Modele zbudowane na zbiorze danych z undersamplingiem 22](#_Toc125316007)

[Modele zbudowane na zmodyfikowanym zbiorze danych 22](#_Toc125316008)

[Wskaźniki oceny jakości 23](#_Toc125316009)

[Wyniki i ich interpretacja 24](#_Toc125316010)

[Wyniki 26](#_Toc125316011)

[Problem decyzyjny 26](#_Toc125316012)

[Istotne czynniki 26](#_Toc125316013)

[Uzyskane reguły decyzyjne 26](#_Toc125316014)

# Podsumowanie

## Problem decyzyjny

Problemem decyzyjnym opisanym w niniejszym raporcie jest kategoryzowanie klientów ze względu na podjęcie lub nie decyzji o rezygnacji z usług banku. Dzięki zaprezentowanym modelom ABC Multinational Bank będzie mógł sklasyfikować, którym klientom trzeba zaoferować lepsze warunki, żeby zapobiec ich odejściu i związane z tym straty finansowe.

## Zbiór danych

Do budowy modeli w naszej pracy wykorzystaliśmy dane pozyskane ze strony:

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

Dane pochodzą z ABC Multinational Bank. Celem klasyfikacji jest przewidzenie czy dany klient będzie skłonny zrezygnować z usług banku na podstawie różnych charakterystyk.

Zbiór składa się z 10000 obserwacji i 12 zmiennych, z których jedna jest zmienną o charakterze identyfikacyjnym, 10 zmiennych objaśniających i zmienna objaśniana „*Churn*”, która ma typ binarny i przyjmuje wartości 0 dla klientów, którzy pozostali i wartości 1 dla tych co zdecydowali się na odejście.

## Uzyskany wynik

Uzyskany wynik - końcowe drzewo po usunięciu wartości „*3*” i „*4*” ze zmiennej „*products\_number*” oraz wprowadzeniu ograniczenia na liść minimalny 100 obserwacji.

Diagram, timeline

Description automatically generated

Rysunek 1. Końcowe drzewo klasyfikacyjne (usunięto wartości "3" i "4" ze zmiennej "products\_number"; minbucket = 100)

## Opis głównych wyników

W ramach projektu zbudowano modele logitowe, drzewa klasyfikacyjne oraz lasy losowe dla danych bez zmian, dla danych z undersamplingiem oraz dla przypadków ze zmianą danych wejściowych. Ostatecznie las losowy okazał się najlepszym modelem, ale ze względu na niemożliwość zinterpretowania odpowiednich reguł decyzyjnych postanowiliśmy wybrać drzewo klasyfikacyjne.

# Dane

## Zmienna objaśniana

Dane pochodzą z baz danych banku ABC Multinational Bank. W zbiorze danych zawiera się 10 zmiennych objaśniających i zmienna objaśniana, udział kategorii której jest pokazany niżej.

Chart

Description automatically generated

Rysunek 2. Udział klasy negatywnej i pozytywnej dla zmiennej "churn"

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 ze zbiorem niezbilansowanym, a więc mamy podstawy do wykorzystania technik, aby uwzględnić możliwe niewystarczające wychwycenie klasy pozytywnej, czyli klientów rezygnujących.

## Analiza danych

W badanym zbiorze danych nie występuje braków danych. Gdyby braki występowały w niedużej ilości, rozważalibyśmy opcję usunięcia takich obserwacji z uwagi na dużą liczebność próby, pod warunkiem, że te braki danych nie wniosłyby istotnych informacji.

Nie obserwowano również wartości odstających dla zmiennych ciągłych.

Chart, box and whisker chart

Description automatically generated

Rysunek 3. Boxplot dla zmiennych ciągłych

Natomiast mieliśmy problem ze zmienną „*products\_number*” - zaledwie 3% całego zbioru stanowiły obserwacje z ilością produktów większą niż „*2*”, co spowodowało ich usunięcie.

## Zmienne objaśniające i ich przekształcenia

W pierwszej kolei postanowiono usunąć zmienna „*customer\_id*”, która jest unikalną i nie ma żadnego efektu na wyniki.

Zmienna „*credit\_score*”, która oznacza liczbę punktów nadanych bankiem klientom. Te punkty, z kolei, oznaczają czy klient jest „wiarygodny” w ocenie banku. Im więcej jest punktów, tym lepszy jest wynik i tym większym zaufaniem można obdarzyć klienta, oferując mu usługi. Poniższy wykres przedstawia histogram wartości przyznanych jednostek punktowych:

Chart, histogram

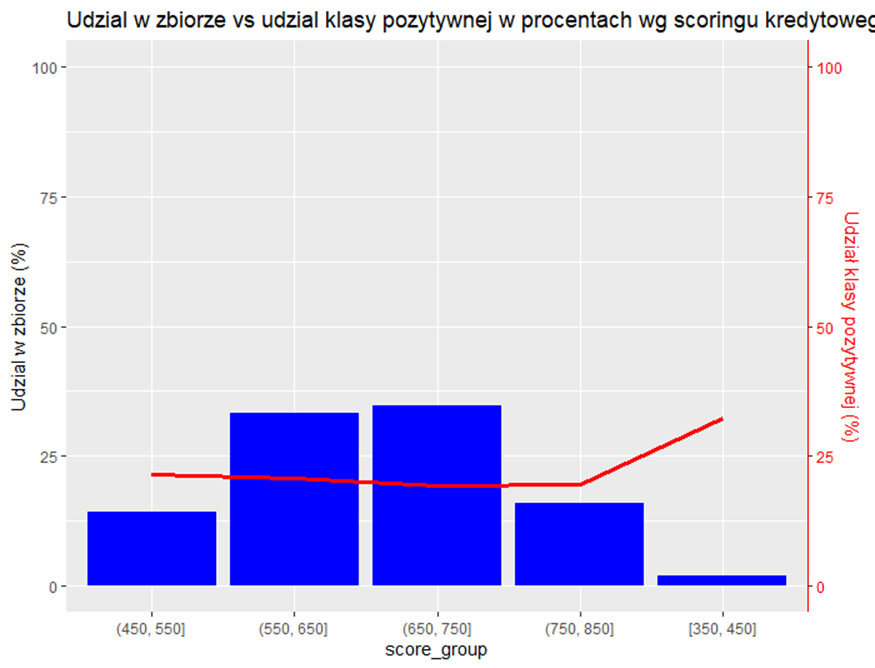
Description automatically generated

Rysunek 4. Histogram zmiennej "credit\_score"

Do budowy modeli ze zbioru po transformacji zmiennych zmienna „*credit\_score*” została przekształcona w zmienną nominalną, gdzie wydzielono 5 przedziałów:

1. 350 – 450 punktów - Very bad
2. 450 – 550 punktów – Bad
3. 550 – 650 punktów – Ok
4. 650 – 750 punktów – Good
5. 750 – 850 punktów – Great

Odpowiedni dla tego podziału wykres prezentuje się następująco:



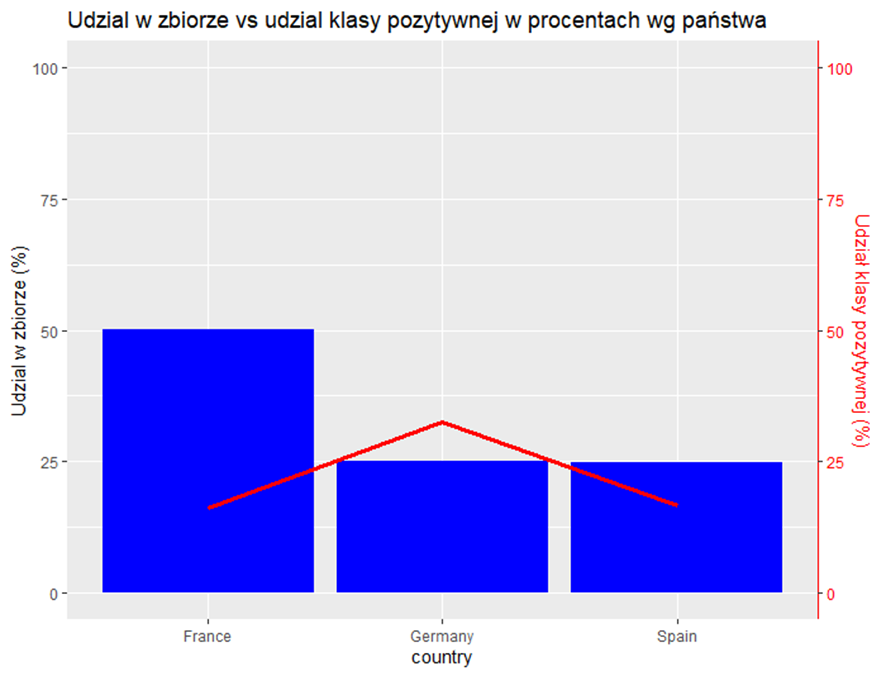
Rysunek 5. Histogram z udziałem klasy pozytywnej dla zmiennej "credit\_score"

Czerwona linia obecna na wykresie pokazuje udział klasy pozytywnej w pięciu grupach. Jak i było oczekiwane największą wartość (31.25%) przyjmuje ona dla grupy [350, 450], czyli dla klientów ocenionych jako „najmniej wiarygodnych”. Czyli te osoby najczęściej rezygnowały z korzystania z usług banku ABC Multinational. Natomiast klienci z innych grup scoringowych odchodziły z banku z mniejszym prawdopodobieństwem.

Przejdziemy teraz do następnej zmiennej, „*country*”. Jest to zmienna typu nominalnego, a w danych znajdziemy trzy kraje: Francję, Niemcy oraz Hiszpanię. Podział udziałów tych krajów wygląda następująco:

1. we Francji mają konta 50.14% z 10000 osób
2. w Niemczech znajdują się 24.77% obserwacji
3. w Hiszpanii 25.09% osób

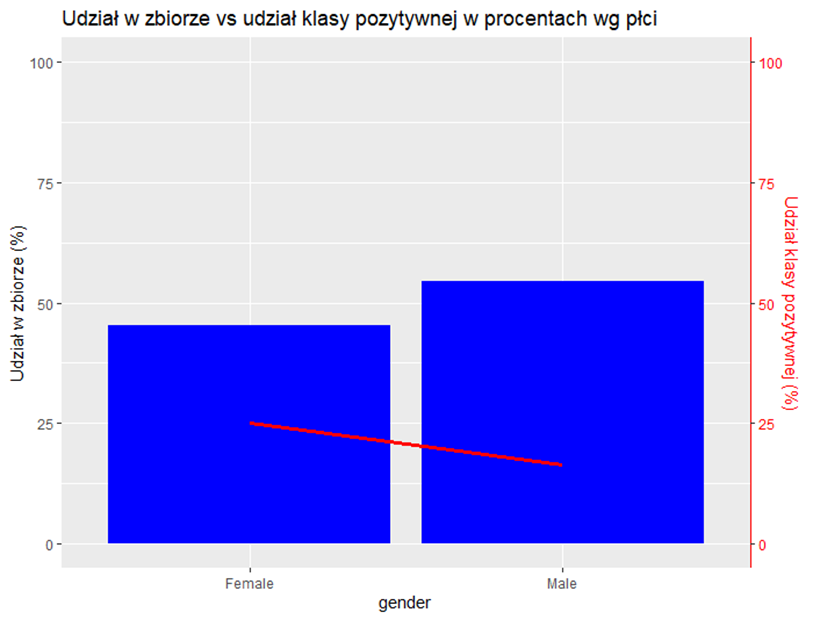
Zobaczymy to również na poniższym wykresie:



Rysunek 6. Histogram z udziałem klasy pozytywnej dla zmiennej "country"

Kolorem czerwonym są oznaczone klienci, którzy postanowili zrezygnować z usług ABC International bank. Warte zauważenia jest to, że przy około dwukrotnej różnicy w liczbie klientów we Francji i Niemczech, procent rezygnujących w Niemczech jest zdecydowanie wyższy, co może być wskazówką dla banku, w której lokalizacji trzeba przede wszystkim implementować rozwiązania “anti-churnowe”.

Dalej rozpatrzymy zmienną „*gender*”. Jest to zmienna binarna przyjmująca wartości “*Male*” lub “*Female*”. Wykres zmiennej znajduje się poniżej:



Rysunek 7. Histogram z udziałem klasy pozytywnej dla zmiennej "gender"

Na powyższym histogramie widać, że klientów deklarujących przynależność do płci męskiej jest więcej. Należy również zaznaczyć, że odpowiadający im udział klasy pozytywnej jest również mniejszy, co oznacza, że mężczyźni rzadziej rezygnują z usług ABC International.

Kolejną zmienną do omówienia jest „*age*”. Oryginalnie to była zmienna dyskretna z minimalną wartością „*15*” i maksymalną „*95*”. Odpowiedni histogram znajduje się poniżej:

Chart, histogram

Description automatically generated

Rysunek 8. Histogram zmiennej "age"

Została ona jednak przekształcona w zmienna nominalną dla zbioru z modyfikacją zmiennych. Grupy wiekowe wyglądają następująco:

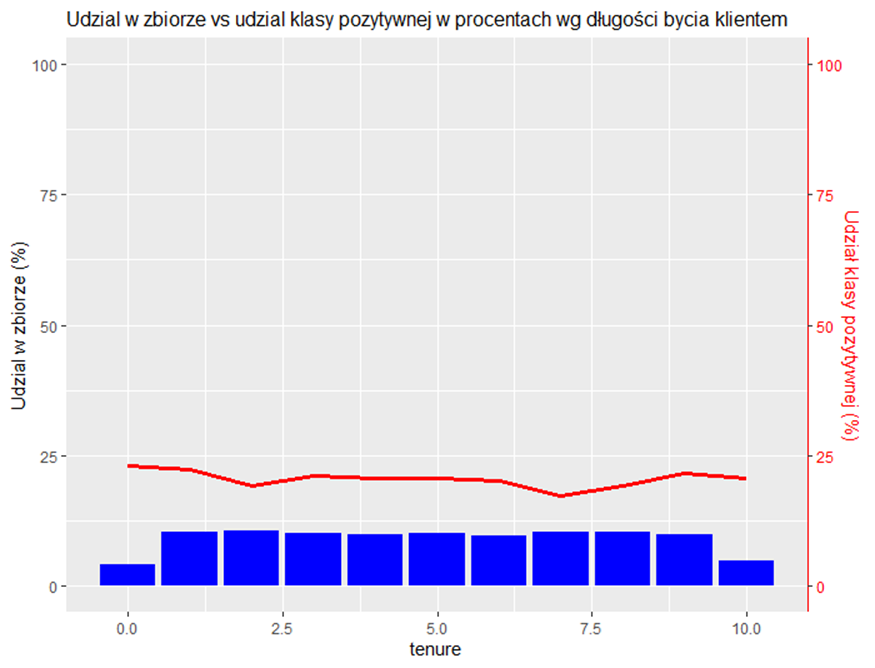
Chart

Description automatically generated

Rysunek 9. Histogram z udziałem klasy pozytywnej dla zmiennej "age"

Na powyższym wykresie możemy zauważyć, że z każdą kolejną grupą udział klasy pozytywnej ciągle wzrastał osiągając szczyt dla osób w wieku 50-55 lat. Oni decydują się na odejście z prawdopodobieństwem wynoszącym aż 58%. Bardzo niskie znaczenia wartości udziału klasy pozytywnej zobaczymy u najmłodszych i najstarszych przedstawicieli populacji. Grupy wiekowe [15, 20]; (60, 65] itd. stanowią mniej niż 5 procent całego zbioru, jednak mają istotne znaczenie dla predykcji odejścia klienta.

Przejdziemy do zmiennej „*tenure*”, która ma typ dyskretny i pokazuje jak długo (w latach) klient ma konto w tym banku. W oparciu na tą zmienną zbudowaliśmy następujący histogram:



Rysunek 10. Histogram z udziałem klasy pozytywnej dla zmiennej "tenure"

Pomijając pierwszą i ostatnią grupę, udziały klientów w pozostałych grupach są rozłożone mniej więcej równie. Udział klasy pozytywnej jest podobny dla wszystkich grup.

Przekształciliśmy więc tą zmienną dla zbioru z modyfikacją zmiennych:

1. 0-1 - Nowy klient
2. 1-5 - Stały klient
3. 5-10 - Lojalny klient

Chart, histogram

Description automatically generated

Rysunek 11. Histogram z udziałem klasy pozytywnej dla zmiennej "tenure" z nowymi przedziałami

Następną zmienną do omówienia jest „*balance*”. Jako zmienna ciągła przyjmuje ona wartości od „*0*” do „*250898*”. Histogram dla „*balance*” przedstawiamy poniżej:

Chart, histogram

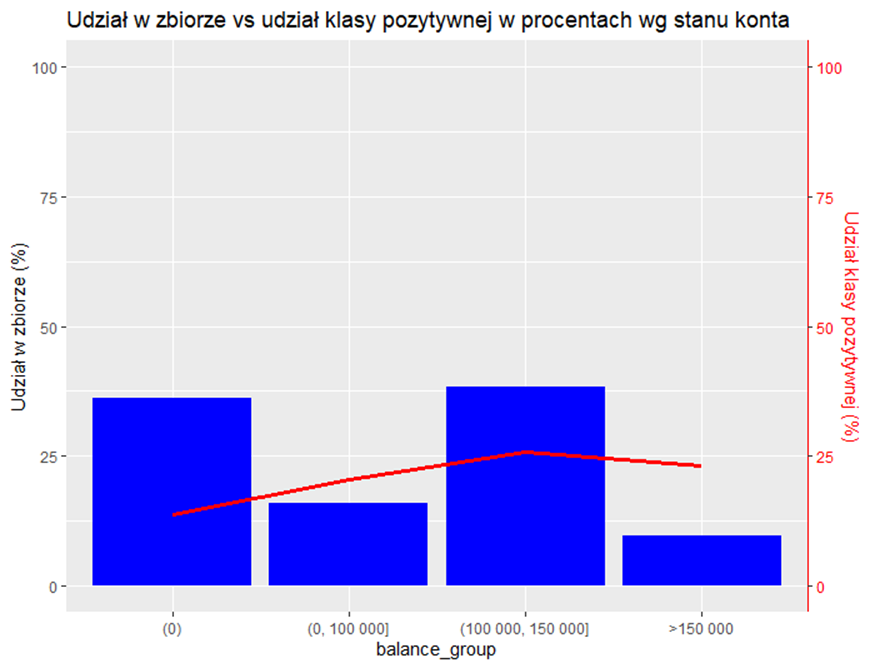
Description automatically generated

Rysunek 12. Histogram zmiennej "balance"

Przekształciliśmy ją w zm. typu nominalnego dla zbioru z modyfikacją zmiennych z następującym podziałem na grupy:

1. 0 – osoby nie mające środków na koncie,
2. 0 - 100000 jednostek pieniężnych na koncie,
3. 100000 - 150000 jednostek pieniężnych,
4. 150000 i więcej j. p.

Opierając się na wyżej opisany podział zbudowaliśmy histogram:



Rysunek 13. Histogram z udziałem klasy pozytywnej dla zmiennej "balance" z nowymi przedziałami

Na prezentowanym wykresie można zaobserwować, że największa grupa (ok. 37.5%) klientów banku ABC International posiada na koncie od „*100000*” do „*150000*” jednostek pieniężnych. Przedstawiciele tej grupy również najczęściej rezygnują z usług banku (osoba z grupy (100000, 150000] odejdzie z prawdopodobieństwem wynoszącym 25%). Warte uwagi jest to, że drugą najliczniejszą grupą są klienci z zerową wartością środków na koncie. Osoby z tej grupy mają jednak najmniejsze prawdopodobieństwo rezygnacji (ok. 13%). Osoby posiadające więcej niż 150000 j.p. na koncie stanowią najmniejszą grupę, a szansa, że odejdą z ABC International jest tylko o kilka procent mniejsza od poprzedniej grupy (100000, 150000]. Nieco więcej klientów zaliczają się do grupy (0, 100000] (ok. 15%). W ich przypadku udział klasy pozytywnej wynosi ok. 18%.

Nie zapomnijmy również o zmiennej „*products\_number*”, która odzwierciedla ile produktów banku posiada dany klient. „*Products\_number*” ma typ dyskretny i przyjmuje wartości {1, 2, 3, 4}. Zbudowano odpowiedni histogram:

Chart, line chart

Description automatically generated

Rysunek 14. Histogram z udziałem klasy pozytywnej dla zmiennej "products\_number"

Łatwe do zaobserwowania jest to, że największy udział klasy pozytywnej przypada na klientów posiadających 4 produkty. Prawdopodobieństwo ich odejścia wynosi 1, czyli oni zawsze rezygnują (grupa ta liczy 60 ludzi). Osoby posiadające o jeden produkt mniej też mają bardzo wysokie prawdopodobieństwo rezygnacji z usług (ok. 0.82). Mamy dwie hipotezy dlaczego obserwujemy taki efekt. Po pierwsze, oferty innych banków stają się coraz bardziej atrakcyjniejsze (klient skorzystał z wszystkich proponowanych ofert). Z innej strony, produkty te mogą zawierać w sobie pożyczki i kredyty, więc klient preferuje wycofać się już z banku.

Z uwagi na bardzo małą liczbę klientów (<5%) posiadających więcej niż 2 produkty, postanowiliśmy skasować te obserwacje ze zbioru do budowy wszystkich modeli, tym samym przekształcając zmienną „*products\_number*” w binarną - klient posiada 1 lub 2 produkty.

Chart, histogram

Description automatically generated

Rysunek 15. Histogram z udziałem klasy pozytywnej dla zmiennej "products\_number" (bez wartości "3" i "4")

Następną zmienną, którą wykorzystano jest „*credit\_card*”. Ma ona typ binarny, gdzie „*0*” (lewa kolumna) oznacza, że nie posiada karty kredytowej, a „*1*” (prawa kolumna), że posiada.

Na poniższym histogramie da się zaobserwować, że posiadaczy kart jest o około 40% więcej. Natomiast, jak widać, udział klasy pozytywnej różni się bardzo nieznacznie (różnica wynosi ok. 2%), co może powodować niską siłę predykcyjną.

Chart, histogram

Description automatically generated

Rysunek 16. Histogram z udziałem klasy pozytywnej dla zmiennej "credit\_card"

Zmienna „*active\_member*” ma taki sam tym jak poprzednia, binarny. W lewej kolumnie znajdują się obserwacje zaliczone do grupy nieaktywnej, a w prawej do grupy aktywnej. Ta zmienna oznacza zaangażowanie danego klienta w korzystanie z usług banku. Na poniższym wykresie są przedstawione liczebności obu grup oraz udziały klas pozytywnych:

Chart, histogram

Description automatically generated

Rysunek 17. Histogram z udziałem klasy pozytywnej dla zmiennej "active\_member"

Możemy zauważyć, że udział obu grup w zbiorze jest podobny (ok. 4% różnicy). Z kolei, nieaktywni klienci znacznie częściej decydują się na rezygnację (o około 14% częściej), co jest jak najbardziej przewidywalne. Udział klasy pozytywnej w grupie klientów aktywnych nie jest jednak niski, wynosi ok. 13%.

Przejdźmy teraz do ostatniej zmiennej, „*estimated\_salary*”. Jako zmienna ciągła przyjmuje ona wartości na przedziale pomiędzy „*11.58*” i „*199992.48*” i pokazuje przybliżoną płacę danego klienta. Prezentujemy histogram odpowiadający „*estimated\_salary*”:

Chart, histogram

Description automatically generated

Rysunek 18. Histogram zmiennej "estimated\_salary"

Możemy zaobserwować, że w zbiorze wynagrodzenia rosną ciągle bez dużych wzrostów lub spadków. To najprawdopodobniej jest celowym zabiegiem ABC International bank, który te dane udostępnił.

Dla zmiennej „*estimated\_salary*”, jak i dla innych zmiennych ciągłych, zmieniliśmy typ na nominalny dla zbioru z modyfikacją zmiennych. Podział wygląda następująco:

1. 0 - 50000 jednostek pieniężnych,
2. 50000 - 100000 j. p.,
3. 100000 - 150000 j. p.,
4. 150000 - 200000 j. p.

Chart, bar chart

Description automatically generated

Rysunek 19. Histogram z udziałem klasy pozytywnej dla zmiennej "estimated\_salary" z nowymi przedziałami

Na tym wykresie również obserwujemy prawie takie same udziały we wszystkich czterech grupach. Podobnie wygląda też sytuacja z udziałem obserwacji pozytywnych, który jest nieco wyższy tylko w trzeciej grupie, a ogólnie jest równy około 20%, co znowu sugeruje, że zmienna ma niską wartość predykcyjną.

## Korelacja zmiennych objaśniających

Na poniższym wykresie można zaobserwować, że pomiędzy zmiennymi „*products\_number*” a „*balance*” występuje korelacja, ale ona jest słaba.

Chart, treemap chart

Description automatically generated

Rysunek 20. Macierz korelacji zmiennych objaśniających

## Siła predykcyjna

|  |  |
| --- | --- |
| Rysunek 21. IV do przekształceń | Rysunek 22. IV po przekształceniach |

Tutaj możemy zaobserwować wartości Information Value dla zmiennych, z których korzystamy do przekształceń (po lewej) i po przekształceniach (po prawej). Największą różnicę możemy zaobserwować u zmiennej „*age*”, a najmniejszą u „*active\_member*”, „*country*” oraz „*credit\_card*”. W przypadku trzech najmniej wpływowych zmiennych („*tenure*”, „*estimated\_salary*”, „*credit\_card*”) czerwone linie na wykresach oznaczające udział klasy pozytywnej są prawie proste, w wyniku czego nie mamy możliwości prognozować odejście klienta ze względu na dane cechy. Widzimy, że dla niektórych zmiennych przekształcenie w typ nominalny zmniejszyło siłę predykcyjną, jednak modele mogą radzić sobie lepiej bez zaburzających wartości odstających.

# Modele

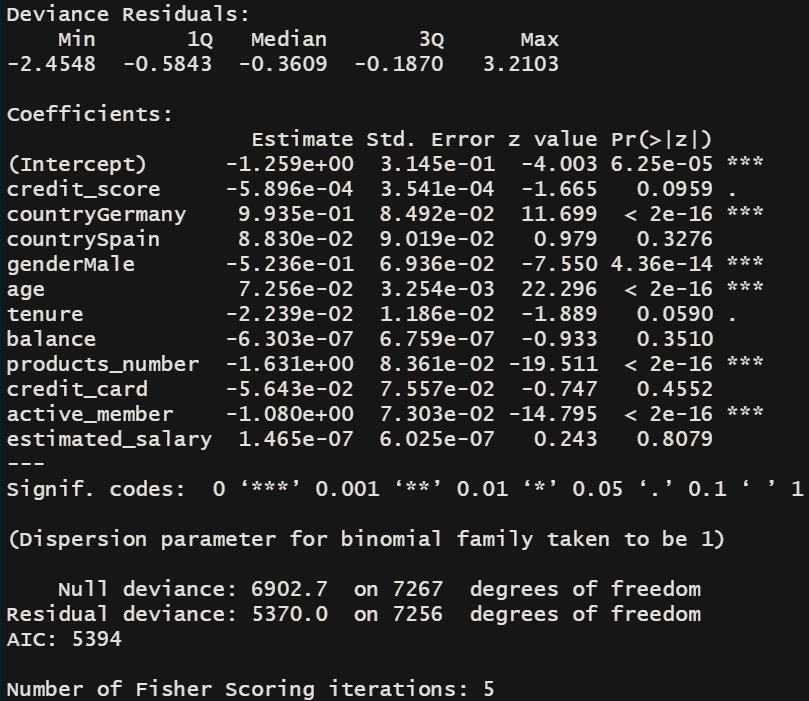
Do klasyfikacji wykorzystaliśmy regresję logistyczną (nie mamy braków danych), drzewo klasyfikacyjne oraz las losowy. Wszystkie modele zostały zbudowane przy pomocy zbioru uczącego i sprawdzone na podstawie zbioru testowego. Dla wszystkich modeli usunęliśmy obserwacje z wartościami “*3*” i “*4*” w zmiennej „*product\_number*”. Żeby znaleźć najlepszy model, wykorzystaliśmy trzy metody dla zmiany danych wejściowych: model bazowy, model po transformacji danych oraz model z undersampling’iem, mając na celu uwzględnienia niezbilansowanego zbioru.

W modelu bazowym podzieliliśmy zbiór danych losowo na część uczącą (75% obserwacji) i testową (25% obserwacji). Nie korzystamy tutaj z walidacji krzyżowej i bootstrapu, ponieważ mamy dużą ilość obserwacji. Undersampling polega na wprowadzeniu tendencyjności w celu wybrania większej liczby próbek z jednej klasy niż z innej, aby zrekompensować nierównowagę, która jest już obecna w danych lub która prawdopodobnie rozwinie się, jeśli zostanie pobrana czysto losowa próbka. Model z undersampling’iem polega na podziale zbioru danych losowo na część uczącą i testową. Po podziale sprawdziliśmy ile mamy rekordów z wartością „*1*” w kolumnie „*churn*” (jest to klasa mniejszościowa) oraz usunęliśmy losowe rekordy z próbki większościowej, aż obie klasy miały tę samą liczebność. Model po transformacji danych wyróżnia się tym, że zamieniliśmy wszystkie zmienne ciągłe na przedziałowe, podzieliliśmy zbiór danych losowo na część uczącą i testową.

## Regresja logistyczna

Z powodu tego, że zmienna zależna „*churn*” jest binarną zrobiliśmy model logitowy, który zostanie punktem odniesienia dla pozostałych modeli.

* Model na podstawie danych bez zmian (bazowy):



Rysunek 23. Podsumowanie modelu logitowego bazowego

* Model z undersampling’iem

Text

Description automatically generated

Rysunek 24. Podsumowanie modelu logitowego z undersamplingiem

* Model po transformacji danych:

Text

Description automatically generated

Rysunek 25. Podsumowanie modelu logitowego po transformacji danych

## Drzewo klasyfikacyjne

Z powodu tego, że otrzymaliśmy dużo liści z małym udziałem od całej populacji wprowadziliśmy ograniczenie na minimalną wielkość. Nasze dane mają dość dużo rekordów, więc zdecydowaliśmy się na wprowadzenie ograniczenia „minbucket = 100”. Małe liście są bardziej podatne na wychwytywanie szumu z danych oraz mogą wskazywać na przetrenowanie modelu.

* Model na podstawie danych bez zmian (bazowy)
  + Przed wprowadzeniem ograniczenia na minimalną wielkość liścia:

Timeline

Description automatically generated

Rysunek 26. Drzewo klasyfikacyjne bazowe bez "minbucket = 100"

* + Po wprowadzeniu ograniczenia na minimalną wielkość liścia:

Diagram

Description automatically generated

Rysunek 27. Drzewo klasyfikacyjne bazowe z "minbucket = 100"

Text

Description automatically generated

Rysunek 28. Statystyki dla oceny

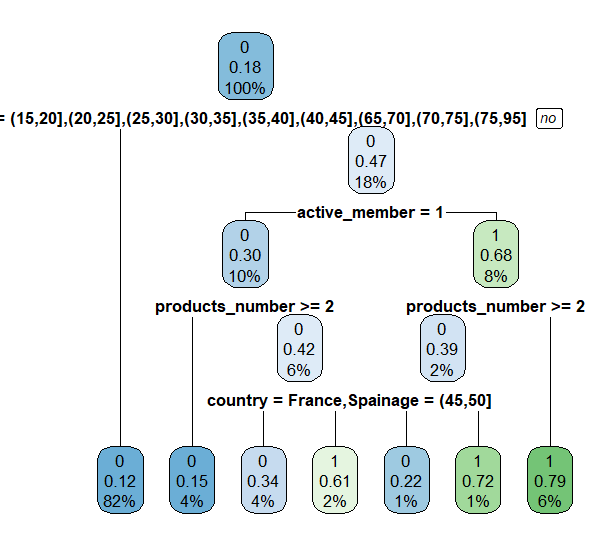
Jak widać, statystyki trochę się pogorszyły, natomiast możemy być pewni, że na innych danych jakość predykcji modelu nie będzie znacznie niższa.

* Model z undersampling’iem

Zbudowaliśmy model na podstawie danych z undersamplingiem i okazało się, że żaden z liści nie zawiera mniej niż 100 obserwacji.

|  |  |
| --- | --- |
| Timeline  Description automatically generated  Rysunek 29. Drzewo klasyfikacyjne z undersamplingiem z "minbucket = 100" | Text  Description automatically generated  Rysunek 30. Statystyki dla oceny |

* Model po transformacji zbioru danych
  + Do wprowadzenia ograniczenia na minimalną wielkość liścia:



Rysunek 31. Drzewo klasyfikacyjne po transformacji danych bez "minbucket = 100"

* + Po wprowadzeniu ograniczenia na minimalną wielkość liścia:

Timeline

Description automatically generated

Rysunek 32. Drzewo klasyfikacyjne po transformacji danych z "minbucket = 100"

Text

Description automatically generated

Rysunek 33. Statystyki dla oceny

Zgodnie z powyższymi statystykami, accuracy, precision i specificity są lepsze w modelu ze wprowadzonym ograniczeniem na minimalną wielkość liścia.

## Las losowy

Każdy las losowy został zbudowany na podstawie 500 drzew. Poniżej zamieściliśmy wykresy wpływu zmiennych. Współczynnik Giniego jest miarą nierównomierności rozkładu zmiennej losowej. Im wyższa wartość współczynnika Giniego, tym większe jest znaczenie danej zmiennej w modelu. Jak widać, najbardziej wpływową zmienną dla wszystkich modeli jest zmienna „*age*”.

* Model na podstawie danych bez zmian (bazowy)

|  |  |
| --- | --- |
| Rysunek 34. Las losowy bazowy | |
| Rysunek 35. Las losowy z undersamplingiem | Rysunek 36. Las losowy po transformacji danych |

# Ocena jakości modelu

## Wstęp

Na początku tej części naszego raportu przedstawiamy statystyki używane do oceny jakości modelu wraz z ich interpretacją:

Accuracy (trafność) – ACC - prawdopodobieństwo poprawnej klasyfikacji

Match Error Rate (MER) = 1- ACC - prawdopodobieństwo błędnej klasyfikacji

Precision (rzetelność) - udział prawidłowych klasyfikacji tych klientów, co odejdą.

Sensitivity (wrażliwość) - zdolność klasyfikatora do prawidłowego identyfikowania klientów, którzy zrezygnują z usług banku.

Specificity (specyficzność) - zdolność klasyfikatora do prawidłowego identyfikowania tych, kto pozostanie w banku.

### Modele zbudowane na niezmodyfikowanym zbiorze danych

Text

Description automatically generated

Rysunek 37. Statystyki dla modeli zbudowanych na niezmodyfikowanym zbiorze danych

Zgodnie z powyższymi wartościami statystyk, las losowy i drzewo klasyfikacyjne - to lepsze modele, ponieważ trafność, rzetelność i inne statystyki znajdują się na dobrym poziomie. Ale nie możemy wybrać najlepszego modelu z nich, dlatego zbudujemy również krzywą Lift i ROC później.

### Modele zbudowane na zbiorze danych z undersamplingiem

Text

Description automatically generated

Rysunek 38. Statystyki dla modeli zbudowanych na zbiorze danych z undersamplingiem

Otrzymane wyniki wskazują na to, że undersampling nie jest dobrą metodą na polepszenie modeli. Co więcej wartości statystyk dla lasu losowego w niektórych przypadkach przyjmują wartości 1, czyli te modele zbyt dobrze opisują nasz zbiór danych, co jest przetrenowaniem modelu, którego staramy się unikać. Zaznaczamy również, że wychwycenie klasy pozytywnej - czyli odejścia klienta - jest o 2,3 razy lepsze w porównaniu do modelu podstawowego.

### Modele zbudowane na zmodyfikowanym zbiorze danych

Również przetestowaliśmy trzy modele na naszym zbiorze danych z przekształconymi zmiennymi objaśniającymi. Otrzymaliśmy podane niżej statystyki:

Text

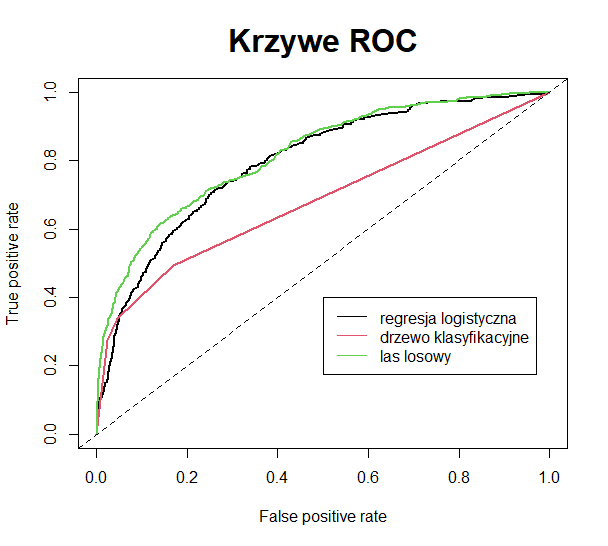
Description automatically generated

Rysunek 39. Statystyki dla modeli zbudowanych na zmodyfikowanym zbiorze danych

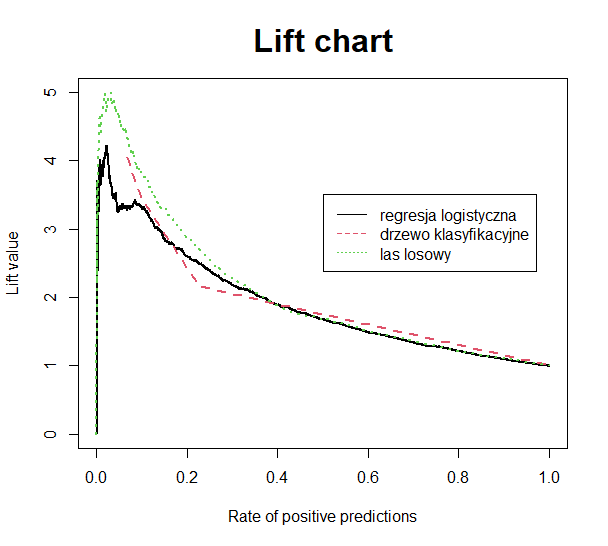
Zdecydowaliśmy, że pozostajemy ze zbiorem danych bez wprowadzonych zmian, ponieważ undersampling tylko pogarsza siłę predykcyjną naszego modelu, a różnica pomiędzy zmodyfikowanym zbiorem danych a zbiorem bez zmian jest minimalna i nie pozwala na wskazanie jaki zbiór danych pasuje do naszych modeli lepiej.

## Wskaźniki oceny jakości

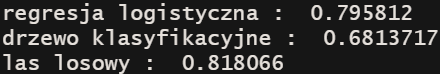
Teraz chcielibyśmy wrócić do naszego pierwotnego zbioru danych i zdecydować, jaki model jest najlepszy: las losowy czy regresja logistyczna. W tym celu zbudowaliśmy krzywe Lift i ROC i policzyliśmy pole pod krzywą (AUC).



Rysunek 40. Krzywe ROC dla regresji logistycznej, drzewa klasyfikacyjnego oraz lasu losowego



Rysunek 41. Krzywa Lift dla regresji logistycznej, drzewa klasyfikacyjnego oraz lasu losowego



Rysunek 42. Wartości AUC dla regresji logistycznej, drzewa klasyfikacyjnego oraz lasu losowego

Zgodnie z obliczoną statystyką AUC i wykresami powyżej, powinniśmy wybrać las losowy jako model najlepiej opisujący nasz zbiór danych. Chociaż warto zaznaczyć, że regresja logistyczna ma również bardzo dobrą siłę predykcyjną.

# Wyniki i ich interpretacja

Wybraliśmy las losowy jako najlepszy model dla klasyfikacji naszego zbioru danych, ale w tym rozdziale opiszemy drzewo klasyfikacyjne oparte na oryginalnym zbiorze danych, ponieważ nie możemy zinterpretować reguł decyzyjnych w lasach losowych.

Nasze drzewo decyzyjne:

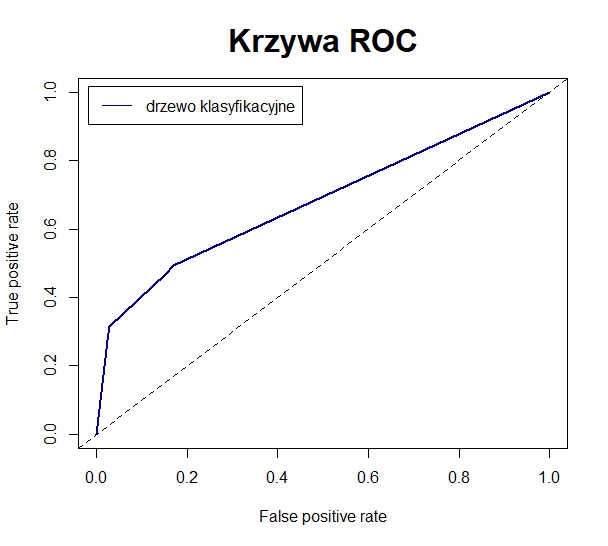
Diagram

Description automatically generated

Rysunek 43. Ostateczny model (drzewo klasyfikacyjne; zbior bazowym; "minbucket = 100"; "products\_number" bez wartości "3" i "4")

Naszym modelem ostatecznym jest drzewo klasyfikacyjne zbudowane na zbiorze bazowym, bez zmiennej „*customer\_id*” z minimalną wielkością liścia równą 100 obserwacji i usuniętymi obserwacjami “3 produkty” i 4 produkty” dla „*products\_number*”.

Krzywa ROC dla ostatecznego modelu:



Rysunek 44. Krzywa ROC dla ostatecznego modelu

Area Under the Curve (AUC) naszego modelu jest równa 0.6813717, czyli można uznać klasyfikator jest słaby.

Krzywa Lift dla ostatecznego modelu:

**Chart, line chart

Description automatically generated**

Rysunek 45. Krzywa Lift dla ostatecznego modelu

Zgodnie z krzywą lift, która pokazuje zysk z zastosowania badanego klasyfikatora względem klasyfikatora losowego, nasz klasyfikator jest lepszy niż losowy.

## Wyniki

### Problem decyzyjny

Problemem decyzyjnym opisanym w raporcie jest kategoryzowanie klientów ze względu na podjęcie lub nie decyzji o rezygnacji z usług banku. Dzięki zaprezentowanym modelom zaklasyfikowaliśmy, którym klientom trzeba zaoferować lepsze warunki, żeby zapobiec ich odejście i związane z tym straty finansowe.

### Istotne czynniki

* Age - liczba lat osoby, będącej klientem banku
* Products\_number - zmienna która odzwierciedla ile produktów banku posiada dany klient
* Active\_member - zmienna oznacza zaangażowanie danego klienta w korzystanie z usług banku

### Uzyskane reguły decyzyjne

Reguły decyzyjne, które otrzymaliśmy w modelu:

1. Age < 45 →True

Interpretacja: jeśli osoba korzystająca z usług banku ma mniej niż 45 lat, prawdopodobieństwo, że odejdzie z tego banku wynosi 11%,

1. Age < 45 → False

active\_member = 1 → True

Interpretacja: jeśli osoba korzystająca z usług banku ma więcej niż 45 lat i aktywnie korzysta z usług banku, że odejdzie z tego banku wynosi 26%

1. Age < 45 → False

active\_member = 1 → False

Products\_number >= 2 True

Interpretacja: jeśli osoba korzystająca z usług banku ma więcej niż 45 lat, aktywnie nie korzysta z usług banku i ma 2 produkty w tym banku, prawdopodobieństwo, że odejdzie z tego banku wynosi 37%.

1. Age < 45 → False

active\_member = 1 → False

Products\_number >= 2 → False

Interpretacja: jeśli osoba korzystająca z usług banku ma więcej niż 45 lat, nie korzysta aktywnie z usług banku i ma mniej niż 2 produkty w tym banku, to prawdopodobieństwo, że odejdzie z tego banku wynosi 76%. Zatem trzeba zaoferować lepsze warunki, żeby zapobiec ich odejściu i związanych z tym strat finansowych.