UNIWERSYTET GDAŃSKI

Projekt zaliczeniowy z przedmiotu Modele Nieparametryczne

Opracowanie:

Współautor
Hinc Błażej

SOPOT 2021

Spis treści

1. Modelowanie zmiennej jakościowej 2
1.1. Zbiór danych – informacje 2
1.2. Modelowanie 3
1.3. Drzewa klasyfikacyjne 3
1.3.1. Wnioski 7
1.4. Agregacja drzew – podejście wielomodelowe 8
1.4.1. Bagging 8
1.4.2. Boosting 11
1.4.3. Las losowy 14
1.4.4. Agregacja wszystkich modeli 16
1.5. Wnioski 16
2. Uogólnione modele addytywne 17
2.1. Dane 17
2.2. Model pierwszy 17
2.2.1. Podsumowanie modelu 18
2.3. Model liniowy 20
2.4. Model drugi 22
2.4.1. Podsumowanie modelu 23
2.5. Wnioski 23
Spis wykresów 25
Spis tabel 25

1.Modelowanie zmiennej jakościowej

1.1. Zbiór danych – informacje.

Zbiór danych wykorzystanych w projekcie dotyczy klientów pewnego banku. Zawarte w nim są niektóre cechy charakteryzujące klienta. Zbiór powstał w celu przewidywania w przyszłości czy klient opuści bank, czy w nim pozostanie na podstawie określonych danych. Każdy rekord w zbiorze to jeden klient. Dane pochodzą z https://www.kaggle.com/mathchi/churn-for-bank-customers.

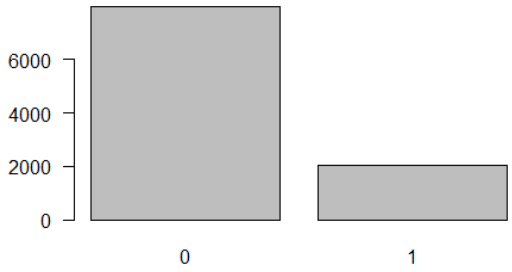
Zbiór danych wykorzystanych w projekcie zawiera następujące zmienne:

- *CreditScore* – ocena kredytowa, zmienna ilościowa wielowariantowa;
- *Geography* – informacja o kraju, w którym mieszka dana osoba, zmienna jakościowa;
- *Gender*– informacja o płci osoby badanej, zmienna nominalna dychotomiczna;
- *Age* – wiek osoby badanej, zmienna ilościowa ciągła;
- *Tenure* – liczba lat, przez które osoba była klientem banku, zmienna ilościowa wielowariantowa;
- *Balance* – wysokość salda, zmienna ilościowa ciągła;
- *NumOfProducts* – liczba produktów, które klient kupił za pośrednictwem banku, zmienna ilościowa;
- *HasCrCard* – informacja o tym czy dana osoba posiada kartę kredytową, zmienna nominalna dychotomiczna;
- *IsActiveMember* – informacja o tym czy dana osoba jest aktywnym klientem, zmienna nominalna dychotomiczna;
- *EstimatedSalary* – szacowane wynagrodzenie, zmienna ilościowa ciągła;
- *Exited* – informacja o tym czy dana osoba opuściła bank, zmienna nominalna dychotomiczna.

Zmienna *Exited* jest zmienną objaśnianą, a pozostałe są objaśniającymi. Poniżej jest przedstawiony rozkład tejże zmiennej. *0* – oznacza ilu klientów opuściło bank, a *1* – ile osób nadal pozostaje jego klientami.

Tablica 1. Warianty zmiennej *Exited*.

0	1
7963	2037



1.2. Modelowanie

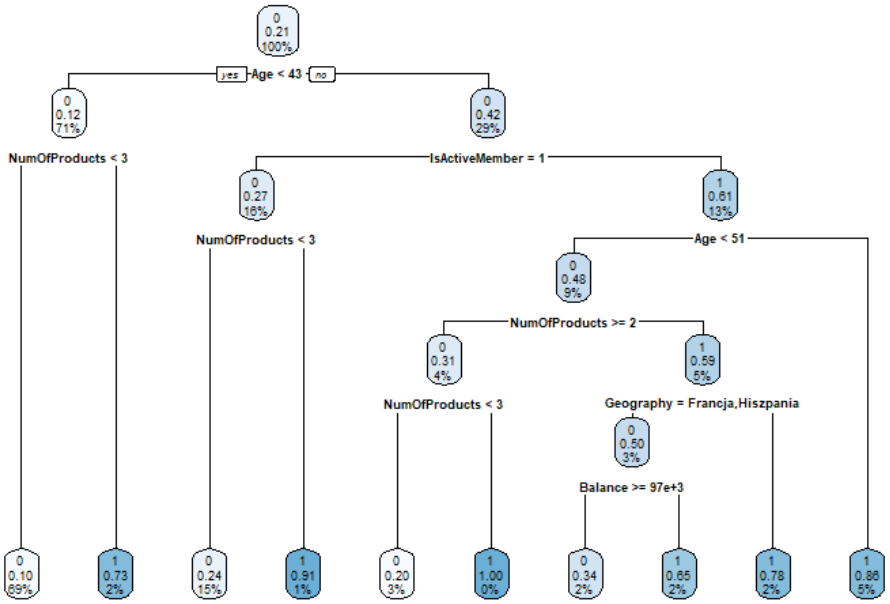
Podjętym przez nas problemem klasyfikacyjnym jest zaklasyfikowanie klienta, do którejś z dwóch klas, na podstawie pozostałych informacji zawartych w zbiorze danych. W tym celu wykorzystamy metodę drzewa klasyfikacyjnego, baggingu, boostingu oraz lasu losowego.

W przypadku pierwszej metody sposób postępowania będzie następujący, na początku dane zostaną losowo podzielone na zbiór uczący oraz zbiór testowy, aby w ostateczności poprzez to działanie przejść do budowy drzewa klasyfikacyjnego.

1.3. Drzewa klasyfikacyjne

Rysowanie wykresu drzewa

Wykres 1. Drzewo początkowe.



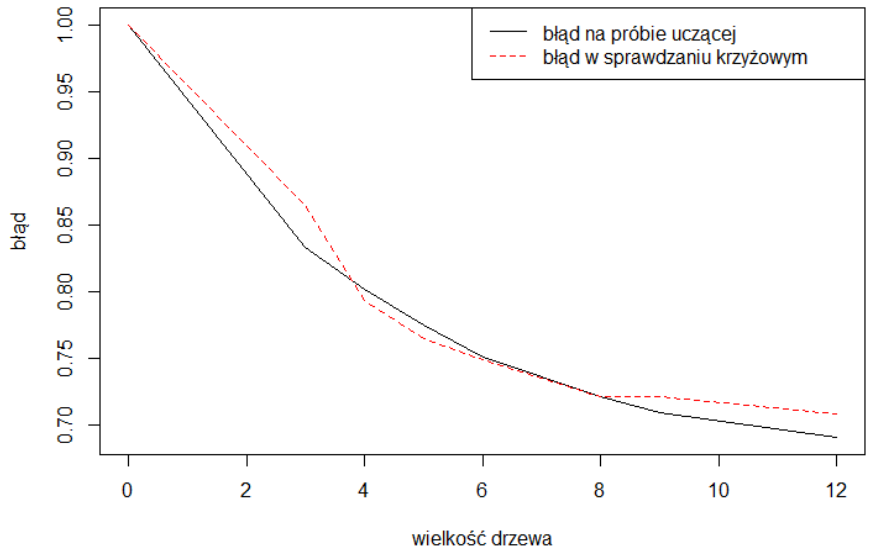
Opis wykresu.

Drzewo początkowe obejmuje 8 podziałów oraz ma 10 „liści". Pierwszą zmienną różnicującą na klasy jest zmienna wiek. Dzieli ona dane na zbiory- osoby poniżej 43 lat oraz w wieku lat 43 oraz powyżej. Jak widzimy oba nowo powstałe zbiory zostały zaklasyfikowane jako klasa 0- osoby, które nie opuściły banku. Następnie węzeł po lewej zostaje podzielony na podstawie zmiennej ***NumOfProducts***, w zależności od tego czy dana osoba nabyła za pomocą banku mniej niż 3 produkty. Po tym podziale stworzone zostają 2 „liście". Węzeł po prawej został podzielony na dwa kolejne przy pomocy zmiennej ***IsActiveMember***. Następnie węzeł powstały po lewej zostaje podzielony na podstawie zmiennej ***NumOfProducts***. Z podziału tego powstają 2 „liście". Węzeł z prawej ponownie dzielony jest za pomocą

wieku. Z podziału tego powstaje kolejny węzeł oraz 1 „liść". Nowopowstały węzeł dzielony jest za pomocą zmiennej ***NumOfProducts*** na 2 kolejne węzły. Z węzła po lewej po kolejnym podziale za pomocą zmiennej ***NumOfProducts*** powstają 2 „liście". Z węzła po prawej powstaje kolejny węzeł oraz „liść". Ostatni podział dokonany jest za pomocą zmiennej *Balance*. W klasyfikacji za pomocą drzewa nie pojawiają się zmienne ***CreditScore***, ***Gender***, ***Tenure***, ***HasCrCard*** oraz ***EstimatedSalary***.

Na podstawie drzewa klasyfikacyjnego można wnioskować, iż osobami opuszczającymi bank są osoby poniżej 43 roku życia, które nabyły za pomocą banku 3 lub więcej produktów; osoby mające 43 lata lubi więcej, które są aktywnymi członkami oraz nabyły za pomocą banku 3 lub więcej produktów; osoby, które nie są aktywnymi członkami mają mniej niż 51 lat oraz nabyły za pomocą banku mniej niż 3 produkty lub takie, które nabyły za pomocą banku mniej niż 2 produkty oraz mieszkają we Francji/Hiszpanii albo ich wysokość salda jest mniejsza niż 37e+3 ; osoby mające 51 lat lub powyżej.

Ustalenie optymalnych parametrów modelu.



Błąd klasyfikacji na zbiorze testowym.

Tablica 2. Błąd resubstytucji.

Błąd klasyfikacji	0.1476

14.76 % osób ze zbioru testowego zostało błędnie sklasyfikowanych.

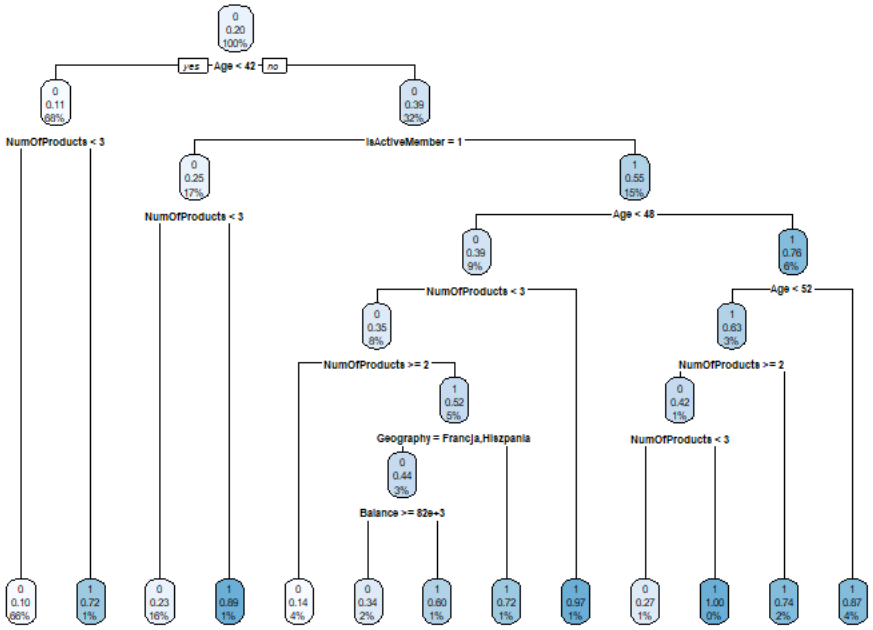
Tablica 3. Macierz błędnych klasyfikacji (*confiusion matrix*).

Rzeczywiste	Nie	Tak
Nie	1913	62
Tak	307	218

Na podstawie macierzy predykcji możemy stwierdzić, że **2 131** (1913+218) jednostek zostało poprawnie zaklasyfikowanych, **62** zostały zaklasyfikowane jako fałszywie pozytywne, a **307** jako fałszywie negatywne.

Ponowne rysowanie wykresu drzewa z optymalnymi parametrami.

Wykres 2. „Maksymalne" drzewo po przycięciu.



Tworzymy wykres dla „maksymalnego" drzewa. Argument *cp* oznacza tutaj o ile minimalnie musi zmniejszyć się błąd resubstytucji względem błędu korzenia, żeby podział w węźle był wykonywany. Powstałe drzewo ma 10 węzłów oraz 13 „liści". Pierwszą zmienną różnicującą na klasy jest zmienna wiek. Dzieli ona dane na zbiory- osoby poniżej 42 lat oraz osoby w wieku 42 lat lub powyżej. Jak widzimy oba nowo powstałe zbiory zostały zaklasyfikowane jako klasa 0- osoby, które nie opuściły banku. Następnie węzeł po lewej zostaje podzielony na podstawie zmiennej ***NumOfProducts***, w zależności od tego czy dana osoba nabyła za pomocą banku mniej niż 3 produkty. Po tym podziale stworzone zostają 2 „liście". Węzeł po prawej został podzielony na dwa kolejne przy pomocy zmiennej ***IsActiveMember***. Następnie węzeł powstały po lewej zostaje podzielony na podstawie zmiennej ***NumOfProducts***. Z podziału tego powstają 2 „liście". Węzeł z prawej ponownie dzielony jest za pomocą wieku. Z podziału tego powstają

2 kolejne węzły. Nowopowstały, po lewej, węzeł dzielony jest za pomocą zmiennej ***NumOfProducts*** na podstawie tego podziału powstaje węzeł oraz „liść". Powstały węzeł po raz kolejny dzielony jest za pomocą zmiennej ***NumOfProducts***. W wyniku tego podziału powstaje węzeł oraz „liść". Nowopowstały węzeł dzielony jest za pomocą zmiennej ***Geography***. Tworzony jest węzeł oraz „liść. Węzeł dzielony jest za pomocą zmiennej ***Balance***. Z tego podziału powstają 2 „liście". Węzeł po prawej, powstały po podziale za pomocą zmiennej wiek, dzieli się za pomocą zmiennej wiek na węzeł oraz „liść". Następnie nowoutworzony węzeł dzieli się za pomocą zmiennej ***NumOfProducts*** na „liść" i węzeł. Kolejny węzeł również dzieli się za pomocą zmiennej ***NumOfProducts*** na 2 „liście". W klasyfikacji za pomocą drzewa nie pojawiają się zmienne ***CreditScore***, ***Gender***, ***Tenure***, ***HasCrCard*** oraz ***EstimatedSalary***.

Na podstawie drzewa klasyfikacyjnego można wnioskować, iż osobami opuszczającymi bank są osoby poniżej 42 roku życia, które nabyły za pomocą banku 3 lub więcej produktów; osoby mające 42 lata lub więcej, które są aktywnymi członkami oraz nabyły za pomocą banku 3 lub więcej produktów; osoby, które nie są aktywnymi członkami mają mniej niż 48 lat oraz nabyły za pomocą banku 3 produkty lub więcej lub nabyły za pomocą banku mniej niż 2 produkty mieszkają we Francji/Hiszpanii, bądź w innym kraju, a ich wysokość salda jest mniejsza niż 82e+3; osoby nie będące aktywnymi członkami, mające 52 lata lub więcej; osoby nie będące aktywnymi członkami, mające mniej niż 52 lata, które nabyły za pomocą banku 1 produkt lub 3 oraz więcej produktów.

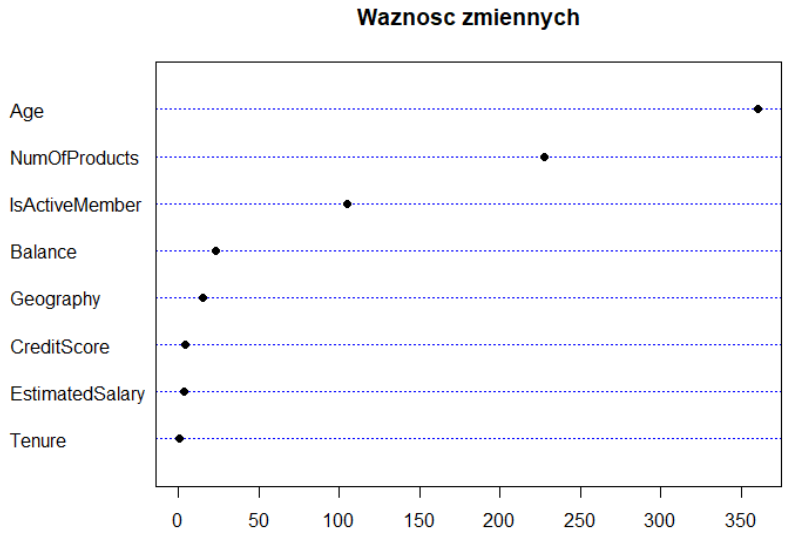
Odczytywanie wartości zmiennych

Tablica 4. Ważność zmiennych.

Zmienna	Ważność zmiennej
Age	360.587213
NumOfProducts	227.565690
IsActiveMember	105.188074
Balance	23.172363
Geography	15.350230
CreditScore	4.447260
EstimatedSalary	3.444354
Tenure	0.837469

Zmienną o największej zdolności dyskryminacyjnej jest zmienna ***Age***. Kolejną zmienną o dużej zdolności dyskryminacyjnej jest zmienna ***NumOfProducts***. Zmienne ***EstimatedSalary*** oraz ***Tenure*** mają bardzo małą zdolność dyskryminacyjną. Jedynie zmienne ***Gender*** oraz ***HasCrCard*** nie znalazły się w zestawieniu- nie mają dużej zdolności dyskryminacyjnej, takiej która będzie pozwalała uzyskać klasyfikację o wyższej jakości.

Wykres 3. Ważność zmiennych.



Największą moc predykcyjną ma zmienna *Age*, a następnie *NumOfProducts* oraz *IsActiveMember*.

Sprawdzanie dokładności na zbiorze testowym.

Tablica 5. Błąd resubstytucji.

Błąd klasyfikacji	0.1444

14.44 % osób ze zbioru testowego zostało błędnie sklasyfikowanych

Tablica 6. Macierz błędnych predykcji.

Rzeczywiste	**Nie**	**Tak**
Nie	1913	62
Tak	299	226

Natomiast na podstawie macierzy predykcji możemy stwierdzić, że **2139** (1913+226) jednostek zostało poprawnie zaklasyfikowanych, **62** zostały zaklasyfikowane jako fałszywie pozytywne, a **299** jako fałszywie negatywne.

1.3.1. Wnioski

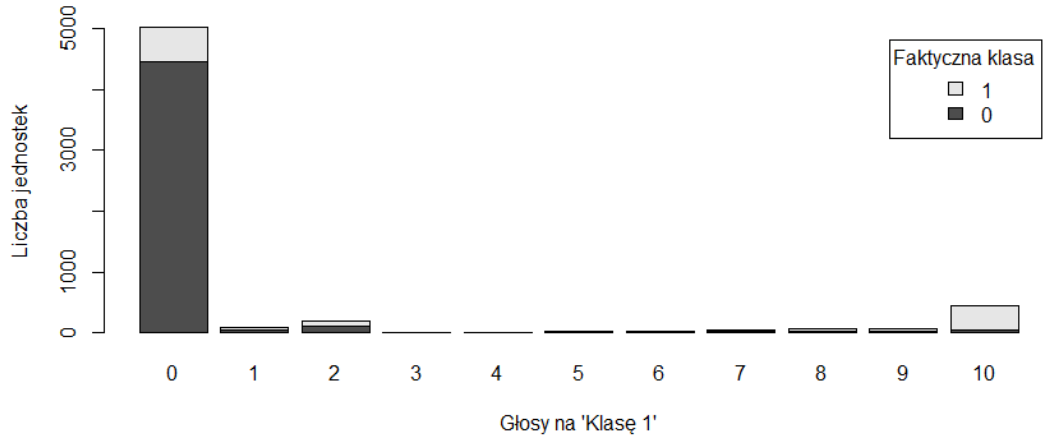
Drzewo powstałe po ustaleniu optymalnych parametrów i zastosowaniu ich jest większe od początkowego o 2 węzły oraz 3 „liście”. Po zmianie parametrów zmniejszył się błąd resubstytucji na zbiorze testowym. Zwiększyła się liczba poprawnie zaklasyfikowanych jednostek oraz zmniejszyła liczba jednostek zaklasyfikowanych jako fałszywie negatywne. Tabela ważności zmiennych pokazuje,

iż zmienne ***CreditScore***, ***EstimatedSalary*** oraz ***Tenure*** mają bardzo małą zdolność dyskryminacyjną, jest ona na tyle mała, iż nie zostały one uwzględnione w żadnym z drzew klasyfikacyjnych. Zmienne ***Gender*** oraz ***HasCrCard*** nie zostały uwzględnione w tabeli ze względu na brak zdolności dyskryminacyjnej. Na podstawie analizy drzew oraz ważności zmiennych można wnioskować, iż 5 zmiennych mogłoby zostać usuniętych z modelu.

1.4. Agregacja drzew – podejście wielomodelowe

1.4.1. Bagging

Wykres 4. Rozkład głosów.

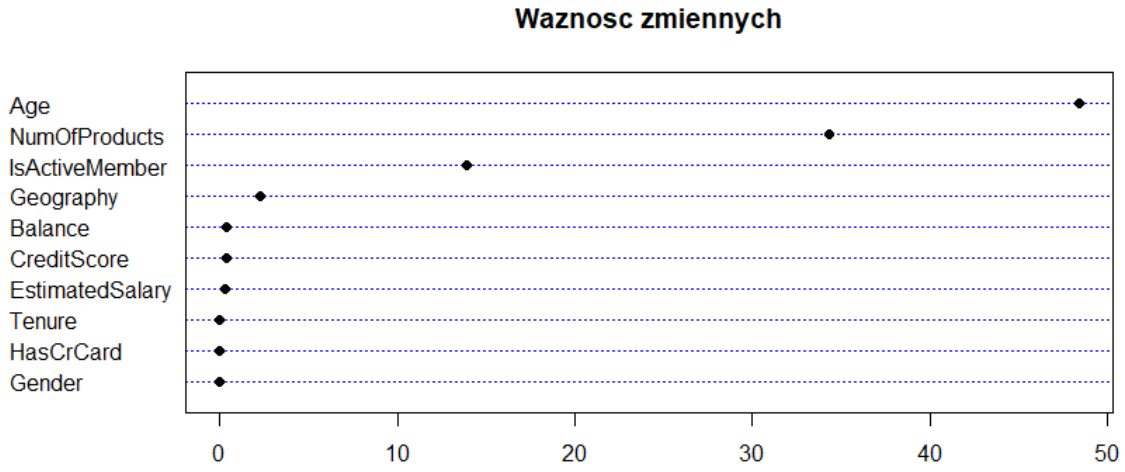


Rozkład głosów jest dosyć skrajny. Modele są mało zmienne. Większość jednostek (**4507**) została poprawnie sklasyfikowana przez wszystkie drzewa jako klasa 0. Natomiast **585** jednostek jest bardzo trudnych do zaklasyfikowania, żaden model nie zaklasyfikował ich do klasy 1. **398** jednostek zostało poprawnie zaklasyfikowanych jako klasa 1, natomiast **70** jednostek zostało błędnie zaklasyfikowanych jako klasa 1 przez wszystkie modele.

Tablica 7. Ważność zmiennych

Zmienna	Ważność zmiennej
Gender	0,00
HasCrCard	0,00
Tenure	0,00
EstimatedSalary	0,26
CreditScore	0,37
Balance	0,40
Geography	2,29
IsActiveMember	13,90
NumOfProducts	34,36
Age	48,41

Wykres 5. Ważność zmiennych.



Zmienną o największej zdolności dyskryminacyjnej jest zmienna ***Age***. Kolejną zmienną o dużej zdolności dyskryminacyjnej jest zmienna ***NumOfProducts***. Zmienne ***HasCrCard***, ***Gender***, ***EstimatedSalary*** oraz ***CreditScore*** oraz ***Tenure*** nie mają zdolności dyskryminacyjnej.

Tablica 8. Błąd resubstytucji.

Błąd klasyfikacji	0.1446

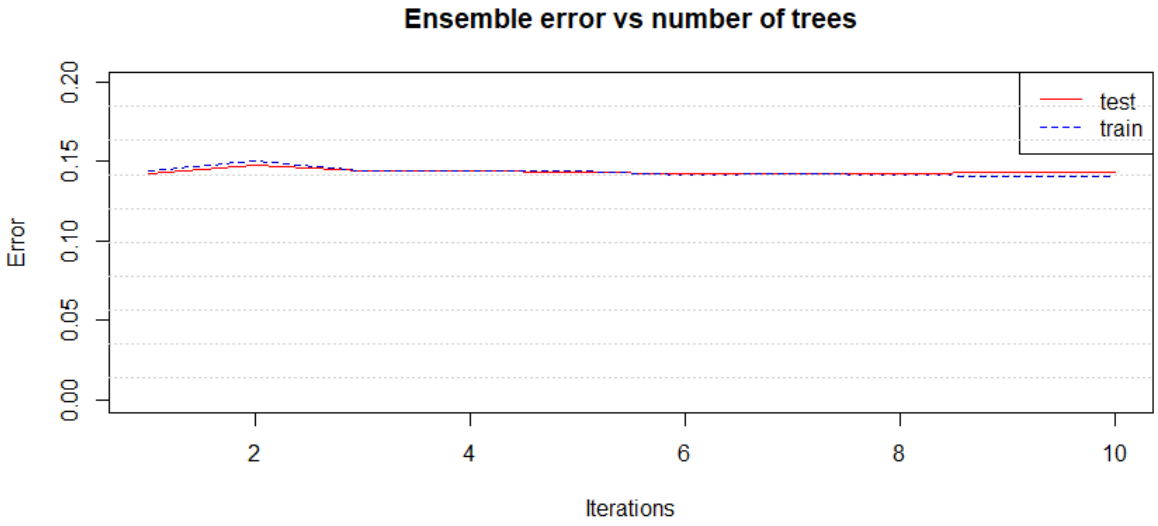
Na podstawie błędu klasyfikacji możemy ustalić, że **14.46 %** jednostek zostało źle zaklasyfikowanych

Tablica 9. Macierz błędnych predykcji.

	Obserwowane	
Przewidywane	**0**	**1**
0	4608	729
1	136	507

Natomiast na podstawie macierzy predykcji możemy stwierdzić, że **5 115** (4608+507) jednostek zostało poprawnie zaklasyfikowanych, **729** zostały zaklasyfikowane jako fałszywie pozytywne, a **136** jako fałszywie negatywne.

Wykres 6. Błędy na zbiorze uczącym i walidacyjnym.



Tablica 10. Macierz błędnych predykcji.

	Obserwowane	
Rzeczywiste	**0**	**1**
0	1527	221
1	57	147

Na podstawie macierzy predykcji, na zbiorze testowym, możemy stwierdzić, że **1 674** (1527+147) jednostek zostało poprawnie zaklasyfikowanych, **221** zostały zaklasyfikowane jako fałszywie pozytywne, a **57** jako fałszywie negatywne.

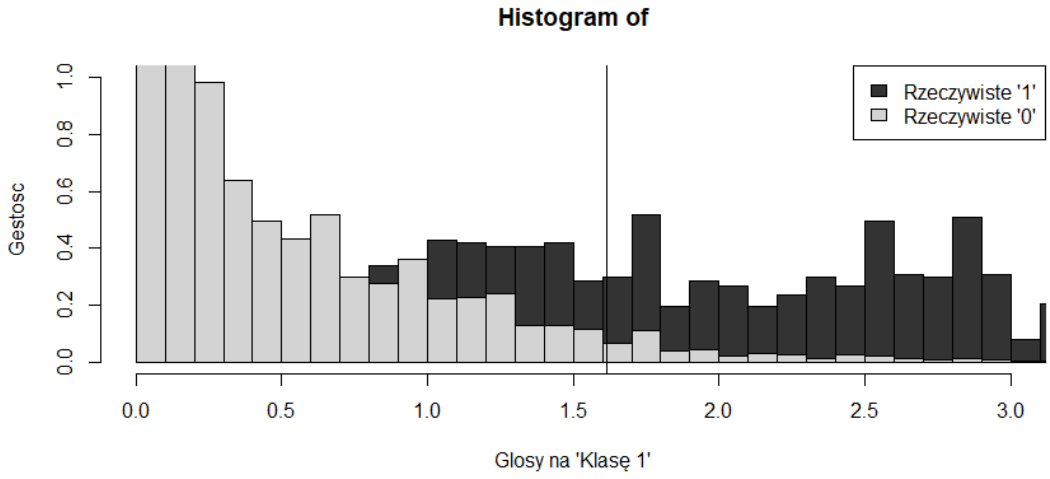
Tablica 11. Błąd resubstytucji.

Błąd klasyfikacji	0.1424

Na podstawie błędu klasyfikacji możemy ustalić, że **14.24 %** jednostek zostało źle zaklasyfikowanych w zbiorze testowym.

1.4.2. Boosting

Wykres 7. Rozkład głosów.

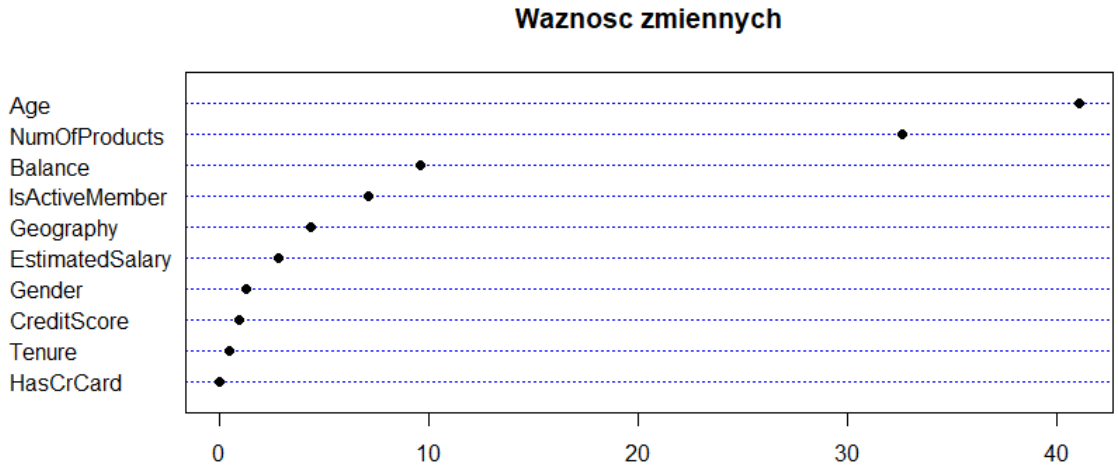


Model boosting nie jest do końca dobrym modelem. Jak widać część jednostek należących do klasy 1 została zaklasyfikowana jako klasa 0. Znacznie mniej jednostek z klasy 0 zostało zaklasyfikowanych jako jednostki z klasy 1, jednak jest ich na tyle dużo, iż można zaobserwować to na wykresie.

Tablica 12. Ważność zmiennych.

Zmienna	Ważność zmiennej
HasCrCard	0,00
Tenure	0,43
CreditScore	0,92
Gender	1,23
EstimatedSalary	2,79
Geography	4,36
IsActiveMember	7,07
Balance	9,61
NumOfProducts	32,56
Age	41,02

Wykres 8. Ważność zmiennych.



Zmienną o największej zdolności dyskryminacyjnej jest zmienna ***Age***. Kolejną zmienną o dużej zdolności dyskryminacyjnej jest zmienna ***NumOfProducts***. Zmienna ***HasCrCard*** nie ma zdolności dyskryminacyjnej.

Tablica 13. Błąd resubstytucji.

Błąd klasyfikacji	0.1351

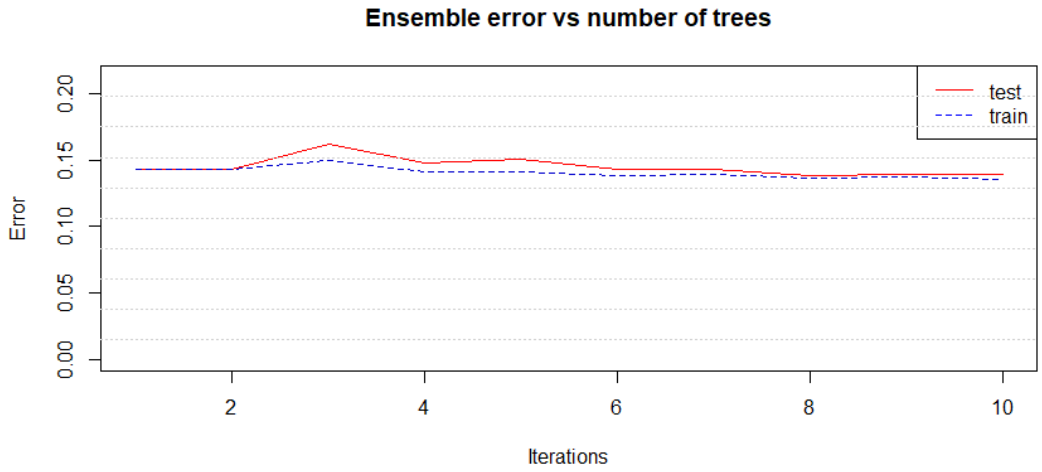
Na podstawie błędu klasyfikacji możemy ustalić, że **13.51 %** jednostek zostało źle zaklasyfikowanych.

Tablica 14. Macierz błędnych predykcji.

	Obserwowane	
Rzeczywiste	**0**	**1**
0	4546	610
1	198	626

Natomiast na podstawie macierzy predykcji możemy stwierdzić, że **5 172** (4546+626) jednostek zostało poprawnie zaklasyfikowanych, **610** zostały zaklasyfikowane jako fałszywie pozytywne, a **198** jako fałszywie negatywne.

Wykres 9. Błędy na zbiorze uczącym i walidacyjnym.



Tablica 15. Błąd resubstytucji.

Błąd klasyfikacji	0.1399

Na podstawie błędu klasyfikacji możemy ustalić, że **13.99 %** jednostek zostało źle zaklasyfikowanych w zbiorze testowym.

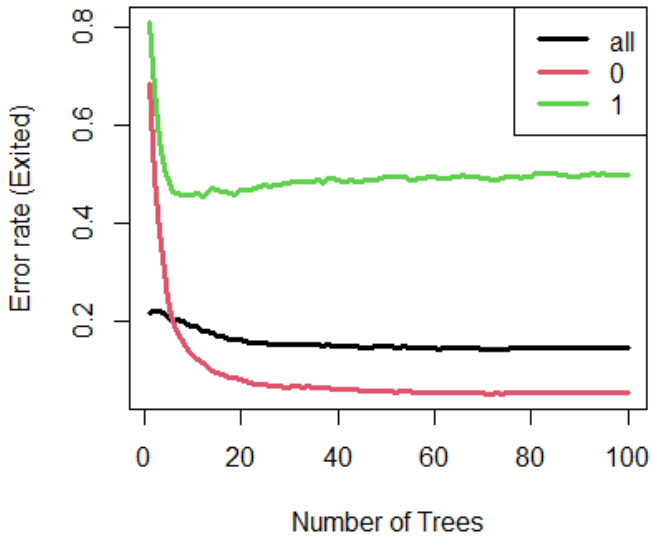
Tablica 16. Macierz błędnych predykcji.

	Obserwowane	
Rzeczywiste	0	1
0	1500	189
1	84	179

Na podstawie macierzy predykcji, na zbiorze testowym, możemy stwierdzić, że **1 679** (1500+179) jednostek zostało poprawnie zaklasyfikowanych, **189** zostały zaklasyfikowane jako fałszywie pozytywne, a **84** jako fałszywie negatywne.

1.4.3. Las losowy.

Wykres 10. Wykres błędu dla całego zbioru oraz poszczególnych klas.



Tablica 17. Ważność zmiennych.

Zmienna	Całkowita ważność
HasCrCard	-0.0001
EstimatedSalary	0.0004
Tenure	0.0012
CreditScore	0.0014
Gender	0.002
Geography	0.011
Balance	0,020
IsActiveMember	0.021
NumOfProducts	0,056
Age	0.058

Zmienną o największej zdolności dyskryminacyjnej jest zmienna ***NumOfProducts***. Kolejną zmienną o dużej zdolności dyskryminacyjnej jest zmienna ***Age***. Zmienna ***EstimatedSalary*** ma ujemną zdolność dyskryminacyjną. Zmienne ***CreditScore*** oraz ***HasCrCard*** nie mają zdolności dyskryminacyjnej.

Tablica 18. Błąd resubstytucji.

Błąd klasyfikacji	0.1462

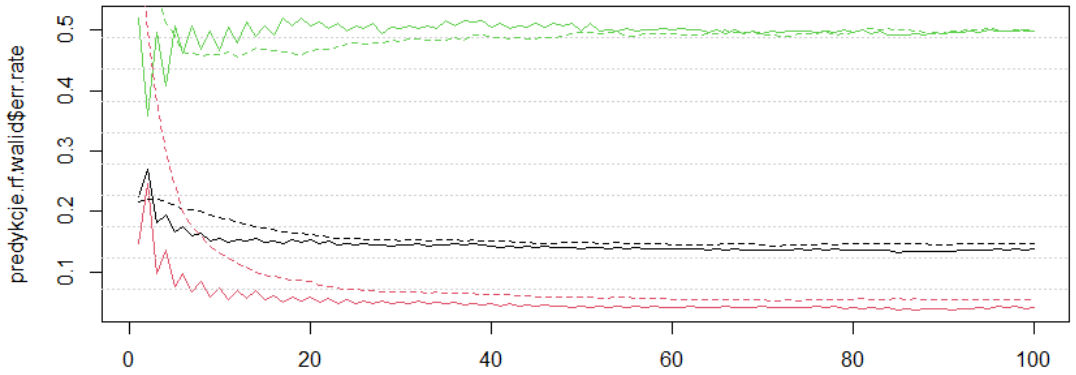
Na podstawie błędu klasyfikacji możemy ustalić, że **14.62 %** jednostek zostało źle zaklasyfikowanych.

Tablica 19. Macierz błędnych predykcji.

	Obserwowane	
Rzeczywiste	**0**	**1**
0	4497	247
1	626	610

Na podstawie macierzy predykcji możemy stwierdzić, że **5 107** (4497+610) jednostek zostało poprawnie zaklasyfikowanych, **247** zostały zaklasyfikowane jako fałszywie pozytywne, a **626** jako fałszywie negatywne.

Wykres 11. Ustalanie optymalnej liczby modeli bazowych.



Tablica 20. Błąd resubstytucji.

Błąd klasyfikacji	0.1414

Na podstawie błędu klasyfikacji możemy ustalić, że **14.14 %** jednostek zostało źle zaklasyfikowanych w zbiorze testowym.

Tablica 21. Macierz błędnych predykcji.

	Obserwowane	
Rzeczywiste	**0**	**1**
0	1504	80
1	195	173

Na podstawie macierzy predykcji, na **zbiorze testowym**, możemy stwierdzić, że **1 584** jednostek zostało poprawnie zaklasyfikowanych, **80** zostały zaklasyfikowane jako fałszywie pozytywne, a **195** jako fałszywie negatywne.

1.4.4. Agregacja wszystkich modeli

Tablica 22. Błąd resubstytucji.

Błąd klasyfikacji	0.1378

Na podstawie błędu klasyfikacji możemy ustalić, że 13.78 % jednostek zostało źle zaklasyfikowanych.

1.5. Wnioski

Najlepsza pod względem wielkości błędu resubstytucji okazała się agregacja modeli (13.78%). Jeśli chodzi o błąd na zbiorze testowym najlepszy okazał się model bagging (13.99%). Najmniej jednostek zaklasyfikowanych jako fałszywie pozytywne, na zbiorze testowym, było w modelu drzewa klasyfikacyjnego, a najmniej jednostek zaklasyfikowanych jako fałszywie w modelu boosting. Zmienna *HasCrCard* nie ma zdolności dyskryminacyjnej w modelu drzewa klasyfikacyjnego, baggingu, boostingu, a w modelu lasu losowego ma ujemną zdolność dyskryminacyjną. Jej usunięcie nie powinno mieć negatywnego wpływu na którykolwiek z modeli. Można by również rozważyć usunięcie zmiennej ***Tenure***, która w zależności od modelu nie ma zdolności dyskryminacyjnej albo ma bardzo małą zdolność dyskryminacyjną (mniejszą od 1) oraz zmiennej ***Gender***, która również nie ma zdolności dyskryminacyjnej lub ma małą zdolność dyskryminacyjną (w modelu boosting 1,23, a w modelu lasu losowego <1). We wszystkich modelach zmienną o największej zdolności dyskryminacyjnej jest zmienna ***Age***, a zmienną o drugiej największej zdolności dyskryminacyjnej jest zmienna ***NumOfProducts***.

2. Uogólnione modele addytywne

2.1. Dane

Ponownie wykorzystano dane opisujące klientów banku. Mając na uwadze specyfikę podjętego problemu regresji, zmienne użyte do modelowania zmiennej ciągłej to:

EstimatedSalary – jako zmienna objaśniana;

oraz zmienne objaśniające:

CreditScore – skoring kredytowy,

Balance – saldo na koncie bankowym,

Age – wiek klienta banku.

Powyższe zmienne są zmiennymi numerycznymi.

2.2. Model pierwszy

$$EstimatedSalary = f_1(CreditScore) + f_2(Balance) + \beta_0 + \boldsymbol{\beta_1} Age + \varepsilon$$

Modelowana jest estymowana wartość miesięcznej pensji w zależności od skoringu kredytowego, salda na koncie bankowym oraz wieku klienta banku. Zmienna ***CreditScore*** i ***Balance*** są składnikami wygładzanymi, natomiast zmienna ***Age*** określa jest składnikiem parametrycznym.

Polecenie `gam1` pozwala nam na otrzymanie miary *GCV*, służącą do oceny jakości modelu. Uogólnione sprawdzanie krzyżowe będzie służyć do porównania zdolności predykcyjnych między różnymi modelami. Im niższa wartość, tym lepsza jest ta zdolność.

Tablica 23. Wartość GCV modelu pierwszego.

GCV	3308970745

Wartość jest naprawdę duża, wręcz ogromna, tym samym świadczy o niskiej jakości modelu.

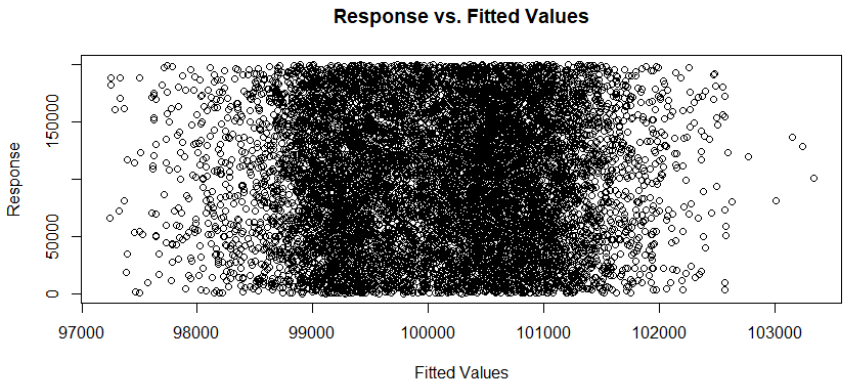
Charakterystyki zmiennych modelu umożliwiające sprawdzenie przestrzeni bazowej, prezentują się następująco.

Tablica 24. Charakterystyki zmiennych modelu pierwszego.

	k'	edf	k-index	p-value
s(CreditScore)	9.00	1.52	0.98	0.14
s(Balance)	9.00	1.00	1.01	0.69

Zmienne mają niskie wartości *p-value*, najniższe ma zmienna ***CreditScore***. Dodatkowo k-index tej zmiennej jest niższy niż 1, co może wskazywać na to, że *k*[1] jest za małe. Mimo wszystko edf (*efektywna liczba parametrów*) nie jest bliska wartości k. Dla zmiennej ***Balance*** wartość *p-value* to 0.69. Stwierdza się, że wartość *k* jest możliwa do akceptacji, parametr został dobrze dobrany.

Wykres 12. Wartości dopasowane względem wartości rzeczywistych modelu 1.



Jeśli model byłby bardzo dobrze dopasowany wykres wyginałby się pod kątem 45 stopni względem linii prostej. Jednakże takiej linii brakuje, a także widoczny jest duży rozrzut na szerokości obu oś, ze skupiskiem na osi X między wartościami 99000 a 101000. Model nie będzie dobry, a prawdopodobnie nawet będzie zły.

2.2.1. Podsumowanie modelu

Tablica 25. Współczynniki parametryczne modelu pierwszego.

	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	101712.97	2211.40	45.99	<2e-16	***
Age	-41.69	54.86	-0.76	0.447	-

Do modelu weszła zmienna ***Age***, zmienna numeryczna. Wynik **–41,69** należy intepretować w taki sposób, że wraz wiekiem estymowana pensja zmniejsza się średnio o **41,69 $** w stosunku do średniej pensji. Zmienna ta nie jest statystycznie istotna. Zmienna parametryczna cechuje się stosunkowo niskim błędem standardowym, zwłaszcza w odniesieniu do całego modelu.

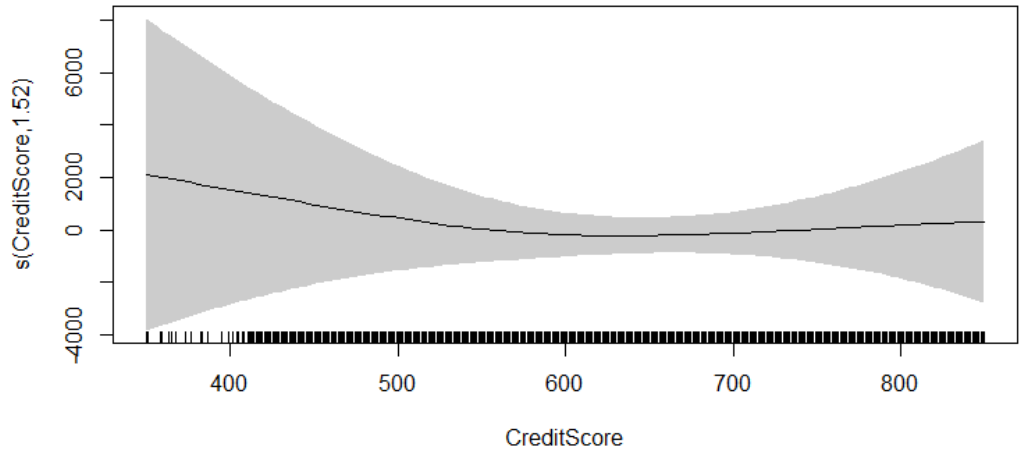
[1] k – wymiar bazy użytej do wygładzania, pochodna liczby węzłów. Im wyższą przyjmie wartość, tym bardziej nasza funkcja będzie mogła się „wyginać”.

Tablica 26. Części wygładzane modelu pierwszego.

	edf	Ref.df	F	p-value	Signif.
s(CreditScore)	1.524	1.892	0.289	0.764	-
s(Balance)	1.000	1.000	1.682	0.195	-

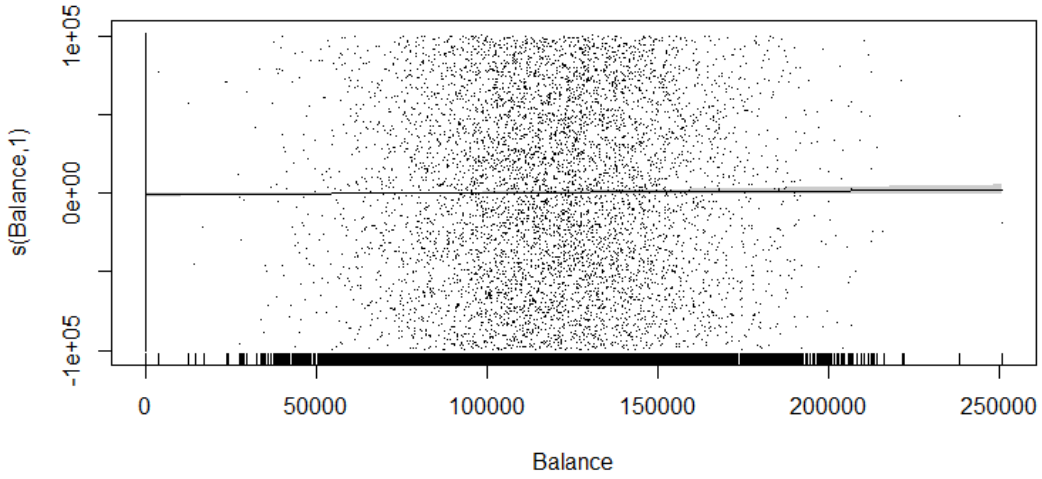
Posługując się testem *F*, możemy sprawdzić czy zmienne ***CreditScore*** oraz ***Balance*** ogólnie istotnie wpływają na zmienną objaśnianą. Wartość przy teście *F* dla zmiennej ***CreditScore*** wskazuje na niższą istotność, niż zmiennej ***Balance***. Warto zwrócić uwagę na *edf*, czyli domyślną wartość parametrów. Z tego względu, że metoda wygładzania zmniejsza liczbę parametrów, *edf* = **1.524** dla zmiennej ***CreditScore*** oznacza sytuację, którą można rozumieć tak jakbyśmy w modelu posiadali **półtora** dodatkowych zmiennych. Współczynnik *R-kwadrat* wynosi **0.0347 %**, jest to bardzo niska wartość wyjaśniania zmienności modelu. Tym samym można stwierdzić, że aż **99.9653 %** zmienności wynagrodzenia wyjaśnione jest innymi czynnikami niż saldo konta bankowego wraz ze skoringiem kredytowym.

Wykres 13. Oszacowana relacja wpływu skoringu kredytowego na pensje.



Oszacowana relacja sugeruje następującą zależność - jeśli wartość oceny kredytowej rośnie, tym wielkość pensji spada. Jednak warto zwrócić uwagę na „rozstrzał" od wartości 300 do 550 – świadczy to o tym, że wpływ ***CreditScore*** jest nieistotny, indywidualnie zmienna ta nieistotnie wpływa na zmienność pensji. Wpływ powoli traci na sile, utrzymując się na względnie podobnym poziomie w okolicy 600 punktacji skoringowej, mając również największy wpływ na pensję w przedziale od 600 punktów skoringowych. Mimo wszystko oddziaływanie ***CreditScore*** zobrazowane przez względnie prostą linię na całej długości wykresu oznacza, że wartość teoretyczna modelu nie powinna się poprawić.

Wykres 14. Relacja wpływu salda na koncie bankowym na pensje, z resztami.



Linia wykazuje się minimalnym nachylaniem dopiero od wartości salda w okolicach 200 000 $. Wpływ tej zmiennej na zmienną objaśnianą tym samym jest minimalny.

2.3. Model liniowy

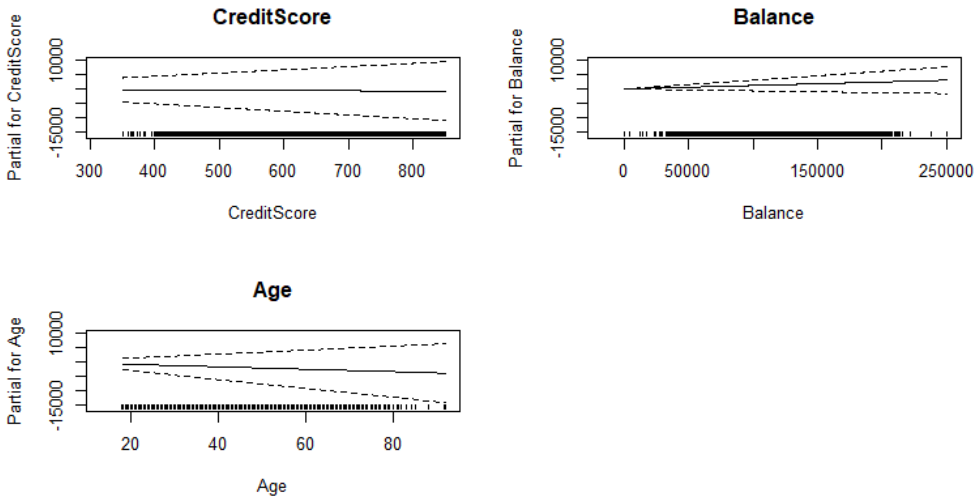
$$EstimatedSalary = CreditScore + Balance + \beta_0 + \boldsymbol{\beta_1} Age + \varepsilon$$

Tablica 27. Współczynniki parametryczne modelu liniowego.

	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	1.014e+05	4.508e+03	22.486	<2e-16	***
CreditScore	-8.901e-01	5.951e+00	-0.150	0.881	-
Balance	1.200e-02	9.222e-03	1.301	0.193	-
Age	-4.154e+01	5.486e+01	-0.757	0.449	-

Zdecydowanie zmalał wpływ zmiennej ***Age***. Generalnie stwierdzić należy, że wpływ każdej zmiennej jest niewielki. Jedynie zmienna ***Balance*** wykazuje wpływ pozytywny na wartość estymowanej pensji – jeśli saldo na koncie wzrośnie o 1 to pensja sumarycznie wzrośnie o **0.012 $**.

Wykres 15. Wpływ poszczególnych zmiennych na zmienną objaśnianą.



Powyższe wykresy wydatnie obrazują nieznaczny wpływ na zmienność modelu.

Tablica 28. Wartości GCV modelu liniowego.

GCV	3309034377

Porównując oba modele w stosunku do wartości *GCV*, model liniowy jest jedynie nieznacznie gorszy od modelu poprzedniego.

Tablica 29. Porównanie średniego błędu kwadratowego dla obu modeli.

gam 1	liniowy
3288778232	3288884397

Wyniki potwierdzają poprzednie – model liniowy jest gorszym modelem, zarówno pod względem predykcyjnym, a także możliwości objaśnienia wpływu zmiennych objaśniających na zmienną objaśnianą. Mimo to, *R-kwadrat* modelu liniowego wynosi 0.0434 %, tym samym jest wyższy niż modelu pierwszego, co utrudnia jednoznaczne wskazanie który model jest lepszym.

2.4. **Model drugi** (interakcja między zmiennymi wygładzanymi).

$$EstimatedSalary = f_1(CreditScore, Balance) + \beta_0 + \boldsymbol{\beta_1} Age + \varepsilon$$

Tablica 30. Wartość GCV modelu drugiego.

GCV	3309679240

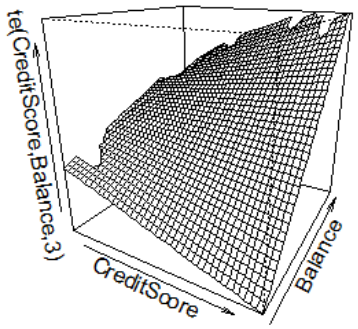
Zdolności predykcyjne modelu drugiego są niższe, niż modelu liniowego.

Tablica 31. Charakterystyki zmiennych modelu drugiego.

	k'	edf	k-index	p-value
te(CreditScore,Balance)	24	3	1.01	0.90

Zmienne przedstawione poprzez wzajemną interakcję wykazują wysoką wartość *p-value*. *K-index* przyjmuje wartość wyższą niż 1. Mimo różnicy między *k-index* a wartością 1 na poziomie 0.01, sytuacja ta może wskazywać , że *k* jest właściwe. *Efektywna liczba parametrów* znacząco różni się od wartości *k*. Parametr *k* został dobrany prawidłowo.

Wykres 16. Wartości dopasowanych względem wartości rzeczywistych modelu 2.



Najwyższy wpływ na wielkość pensji ma sytuacja w której zmienna ***CreditScore*** oraz ***Balance*** przyjmują wysokie wartości.

2.4.1. Podsumowanie modelu.

Tablica 32. Współczynniki parametryczne modelu drugiego.

	Estimate	Std. Error	t value	Pr(>|t|)	Signif.
(Intercept)	101704.38	2211.59	45.987	<2e-16	***
Age	-41.47	54.87	-0.756	0.45	-

Zmienna ***Age*** nadal pozostaje statystycznie nieistotna. Jedynie jej wartość uległa niewielkiej zmianie o **0.12 $**.

Tablica 33. Części wygładzane modelu drugiego.

	edf	Ref.df	F	p-value	Signif.
te(CreditScore,Balance)	3	3	0.589	0.622	-

Wartość testu *p* dla testu *F* jest wyższa od alfa, zatem nie ma podstaw do odrzucenia hipotezy zerowej. Domyślna wartość parametrów to **3**. Współczynnik R-kwadrat zmalał względem modelu pierwszego, wynosi **0.0228 %**, co oznacza, że stopień wyjaśniania zmienności modelu jest jeszcze niższy.

2.5. Wnioski.

Każdy z modeli wykorzystany został w celu oszacowania wartości miesięcznej pensji w zależności od skoringu kredytowego, salda na koncie bankowym oraz wieku klienta banku.

Jednocześnie wszystkie modele cechują się bardzo wysokimi wartościami uogólnionego sprawdzania krzyżowego. Sytuację tę obrazuje poniższa tablica.

Tablica 34. Porównanie modeli.

	GCV	Stopień wyjaśnienia zmienności
model pierwszy	3288778232	0.0471 %
model liniowy	3288884397	0.0434 %
model drugi	3309679240	0.0228 %

Pierwszy analizowany model przedstawiający medianę wartości mieszkania w bloku w zależności od różnych czynników okazał się nienajlepszy jakościowo. Charakteryzował się bardzo wysokimi wartościami uogólnionego sprawdzania krzyżowego oraz błędów standardowych, jednakże to model drugi przyjmuje najwyższe wartości *GCV*.

Dokonując porównania modeli między sobą rozważając je ze względu na stopień wyjaśnienia zmienności, to model pierwszy okazuje się być najlepszym z wartością **0.0471 %** wyjaśnienia zmienności całkowitej modelu.

Warto nadmienić, że żadna ze zmiennych w każdym z modeli nie była statystycznie istotna. Świadczy to o tym, że zmienne objaśniane *CreditScore*, *Balance*, jak również *Age* nie mają istotnego wpływu na zmienną objaśnianą *EstimatedSalary*.

Porównanie modeli między sobą wykazuje, że model drugi znacząco upraszcza zależność występującą między zmiennymi. Jeżeli o takowej zależności w ogóle możemy w tym przypadku mówić. Bazując jedynie na dostępnych danych, to model pierwszy okazuje się być najlepszym, zarówno jakościowo oraz pod kątem wyjaśnienia zmienności.

Spis wykresów

Wykres 1. Drzewo początkowe 3
Wykres 2. Maksymalne drzewo po przycięciu 5
Wykres 3. Ważność zmiennych 7
Wykres 4. Rozkład głosów 8
Wykres 5. Ważność zmiennych 9
Wykres 6. Błędy na zbiorze uczącym i walidacyjnym 10
Wykres 7. Rozkład głosów 11
Wykres 8. Ważność zmiennych 12
Wykres 9. Błędy na zbiorze uczącym i walidacyjnym 13
Wykres 10. Wykres błędu dla całego zbioru oraz poszczególnych klas 14
Wykres 11. Ustalenie optymalnej liczby modeli bazowych 15
Wykres 12. Wartości dopasowane względem wartości rzeczywistych modelu 1 18
Wykres 13. Oszacowana relacja wpływu skoringu kredytowego na pensje 19
Wykres 14. Relacja wpływu salda na koncie bankowym na pensje, z resztami 20
Wykres 15. Wpływ poszczególnych zmiennych na zmienną objaśnianą 21
Wykres 16. Wartości dopasowane względem wartości rzeczywistych modelu 2 22

Spis tablic

Tablica 1. Warianty zmiennej Exited 2
Tablica 2. Błąd resubstytucji 4
Tablica 3. Macierz błędnych klasyfikacji (confiusion matrix) 5
Tablica 4. Ważność zmiennych 6
Tablica 5. Błąd resubstytucji 7
Tablica 6. Macierz błędnych predykcji 7
Tablica 7. Ważność zmiennych 8
Tablica 8. Błąd resubstytucji 9
Tablica 9. Macierz błędnych predykcji 9
Tablica 10. Błąd resubstytucji 10
Tablica 11. Macierz błędnych predykcji 10

Tablica 12. Ważność zmiennych 11
Tablica 13. Błąd resubstytucji 12
Tablica 14. Macierz błędnych predykcji 12
Tablica 15. Błąd resubstytucji 13
Tablica 16. Macierz błędnych predykcji 13
Tablica 17. Ważność zmiennych 14
Tablica 18. Błąd resubstytucji 14
Tablica 19. Macierz błędnych predykcji 15
Tablica 20. Błąd resubstytucji 15
Tablica 21. Macierz błędnych predykcji 15
Tablica 22. Błąd resubstytucji 16
Tablica 23. Wartość GCV modelu pierwszego. 17
Tablica 24. Charakterystyki zmiennych modelu pierwszego 17
Tablica 25. Współczynniki parametryczne modelu pierwszego 18
Tablica 26. Części wygładzane modelu pierwszego 19
Tablica 27. Współczynniki parametryczne modelu liniowego 20
Tablica 28. Wartość GCV modelu liniowego. 21
Tablica 29. Porównanie średniego błędu kwadratowego dla obu modeli 21
Tablica 30. Wartość GCV modelu drugiego 22
Tablica 31. Charakterystyki zmiennych modelu drugiego 22
Tablica 32. Współczynniki parametryczne modelu drugiego 23
Tablica 33. Części wygładzane modelu drugiego 23
Tablica 34. Porównanie modeli 23