RAPORT 3

Wykonany przy pomocy oprogramowania SPSS

PROBLEM: Szacowanie wartości zmiennej CreditScore

Sprawdzenie poprawności danych

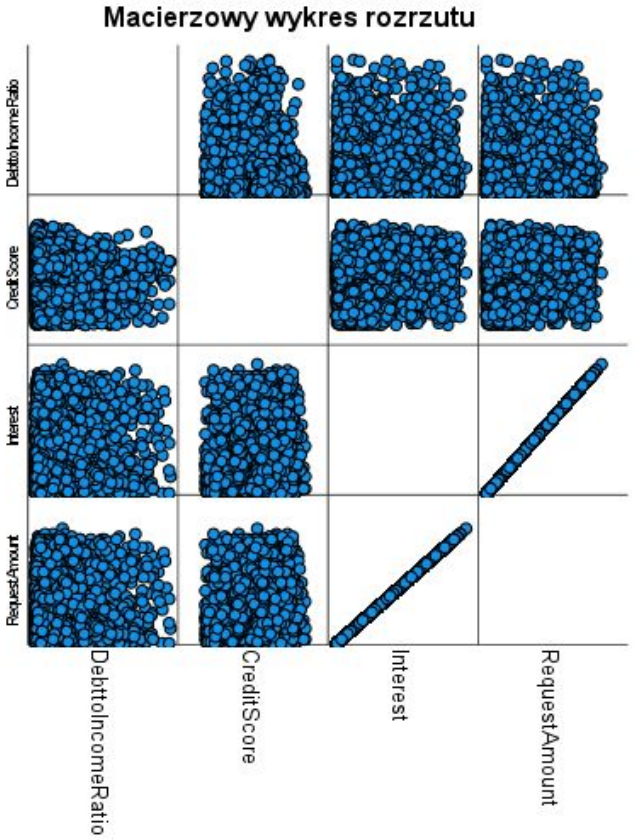
1. Brak braków danych
2. Zamiana zmiennych nominalnych łańcuchowych na nominalne binarne ‘T’ → 1, ‘N’ → 0
3. Szukanie punktów wysokiej dźwigni i usunięcie ich z danych. Dla każdej obserwacji wyliczono wartość dźwigni za pomocą wzoru $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

   Obserwacje uznaje się za wysoką, gdy $h_i > 2*(p+1)/n$. U nas p=4, n=10693. U nas punktami wysokiej dźwigni są 146 obserwacje. Zostały one usunięte w celu uzyskania dokładniejszych modeli.

Model 1 – regresja wielokrotna

Dla nas zmienną celu jest *‘CreditScore’*. Jako predyktory weźmiemy wszystkie pozostałe zmienne tj. *‘Approval’*, *‘DebttoIncomeRatio’*, *‘Interest’*, *‘RequestAmount’*.

Chcemy aby nasze predyktory były silnie skorelowane ze zmienną celu, a słabo ze sobą nawzajem. Można to podejrzeć na macierzowym wykresie rozrzutu:



Jako wniosek z macierzowego wykresu rozrzutu można uznać, że zmienne *'Interest'* oraz '*RequestAmount'* są ze sobą w korelacji. Można się więc spodziewać, że w ostatecznym równaniu regresji wielokrotnej będzie uwzględniona jedna z tych zmiennych.

Przejdźmy dalej. Aby poprawnie skorygować liczbę predyktorów użyjemy selekcji postępującej. Po wpisaniu wszystkich wymagań i warunków w programie SPSS otrzymujemy:

Model - podsumowanie[d,e]

Model	R		R-kwadrat	Skorygowane R-kwadrat	Błąd standardowy oszacowania	Statystyka Durbina-Watsona	
	nowy_podział = uczacy (Wybrane)	nowy_podział ~= uczacy (Nie wybrane)				nowy_podział = uczacy (Wybrane)	nowy_podział ~= uczacy (Nie wybrane)
1	,548[a]		,301	,301	52,236		
2	,573[b]		,329	,328	51,182		
3	,575[c]	,570	,330	,330	51,122	1,968	2,052

a. Predyktory: (Stała), Approval

b. Predyktory: (Stała), Approval, RequestAmount

c. Predyktory: (Stała), Approval, RequestAmount, DebttoIncomeRatio

d. Jeśli nie podano inaczej, statystyki oparte są tylko na obserwacjach, dla których nowy_podział = uczacy.

e. Zmienna zależna: CreditScore

Czyli najlepszy model powstanie, gdy jako predyktory weźmiemy *'Approval'*, *'RequestAmount'* oraz *'DebttoIncomeRatio'*. Opis podziału na zbiór testowy i uczący jest podany przy *Modelu 2* (sieci neuronowe).
Zatem jak wygląda sugerowane **równanie regresji wielokrotnej** powstałe na zbiorze uczącym? Odczytujemy to z tabeli współczynników:

Współczynniki[a,b]

Model		Współczynniki niestandaryzowane		Współczynniki standaryzowane		
		B	Błąd standardowy	Beta	t	Istotność
1	(Stała)	641,049	,854		750,412	,000
	Approval	68,592	1,176	,548	58,310	,000
2	(Stała)	625,416	1,200		520,968	,000
	Approval	69,996	1,155	,560	60,592	,000
	RequestAmount	,001	,000	,168	18,166	<,001
3	(Stała)	621,482	1,493		416,381	,000
	Approval	71,468	1,201	,571	59,514	,000
	RequestAmount	,001	,000	,163	17,524	<,001
	DebttoIncomeRatio	19,808	4,475	,043	4,427	<,001

a. Zmienna zależna: CreditScore

b. Wybrano tylko te obserwacje, dla których nowy_podział = uczacy

CreditScore = 621,482 + 71,468*Approval + 0,001*RequestAmount + 19,808*DebttoIncomeRatio

Zastosujmy to równanie dla wszystkich obserwacji, aby obliczyć wartość przewidywaną (y_pred).

Spójrzmy na jakość szacowania. W tym pomogą nam wyliczone błędy:

	Zbiór uczący	Zbiór testowy
RMSE	51,13	50,05
MSE	2614,11	2504,65
MAE	38,78	38,32
MAPE	5,94%	5,86%

Statystyki opisowe

nowy_podział		N	Średnia
testowy	AE	2636	38,3225
	ME	2636	2504,6526
	N Ważnych (wyłączanie obserwacjami)	2636	
uczacy	AE	7911	38,7838
	ME	7911	2614,1102
	N Ważnych (wyłączanie obserwacjami)	7911	

Statystyki opisowe

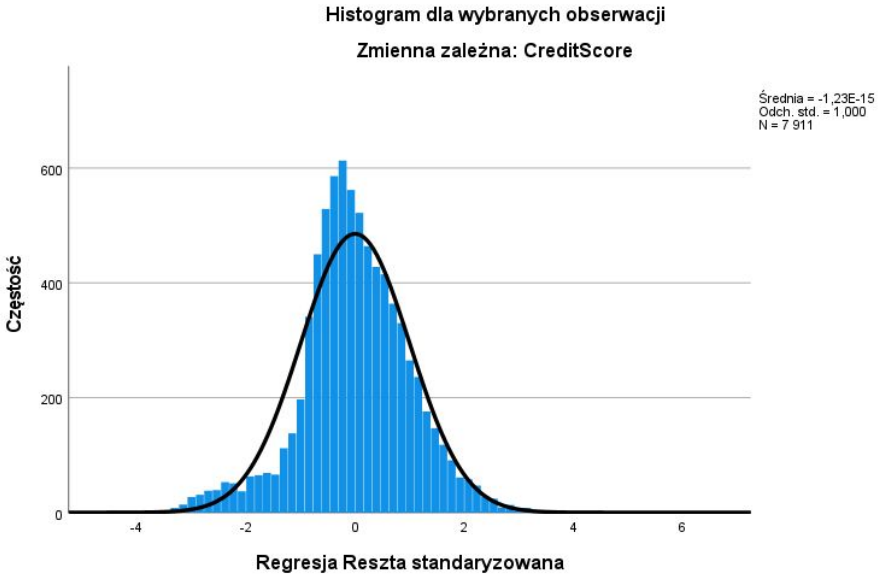
nowy_podział		N	Suma
testowy	APE	2636	154,49
	N Ważnych (wyłączanie obserwacjami)	2636	
uczacy	APE	7911	469,94
	N Ważnych (wyłączanie obserwacjami)	7911	

Aby uznać nasz model regresji wielokrotnej za poprawny musi on spełniać jej założenia.

Czy nasz model spełnia założenia wielokrotnej regresji liniowej?

1. NORMALNOŚĆ RESZT ✅

Możemy zbadać to założenie na podstawie histogramu, testu Shapiro-Wilka lub Kołmogorowa Smirnowa. W naszym przypadku:
Histogram sugeruje normalność. Test Kołmogorowa-Smirnowa zwraca p = 0,054 tzn. nie ma podstaw do odrzucenia H_0 potwierdzającej normalność próby.



Podsumowanie normalnego testu Kołmogorowa-Smirnowa dla jednej próby

Ogółem N			10547
Największe różnice	Wartość bezwzględna		,054
	Dodatnie		,023
	Ujemne		-,054
Statystyki testu			,054
Istotność asymptotyczna(test dwustronny)[a]			<,001
Istotność Monte Carlo (test dwustronny)[b]	Istotność		,000
	Przedział ufności 99%	Dolna granica	,000
		Górna granica	,000

a. Korekty Lillieforsa

b. Metoda Lillieforsa oparta na próbach Monte Carlo (10000) z wartością początkową 221623948.

2. NIEZALEŻNOŚĆ RESZT ✅

Niezależność reszt możemy zbadać testem Durbina-Watsona lub testem serii. W naszym przypadku użyjemy testu Durbina-Watsona. Aby test wykazał, że reszty są niezależne, statystyka Durbina-Watsona powinna znajdować się w przedziale (1,5; 2,5).

Model - podsumowanie[d,e]

Model	R		R-kwadrat	Skorygowane R-kwadrat	Błąd standardowy oszacowania	Statystyka Durbina-Watsona	
	nowy_podział = uczacy (Wybrane)	nowy_podział ~= uczacy (Nie wybrane)				nowy_podział = uczacy (Wybrane)	nowy_podział ~= uczacy (Nie wybrane)
1	,548[a]		,301	,301	52,236		
2	,573[b]		,329	,328	51,182		
3	,575[c]	,570	,330	,330	51,122	1,968	2,052

a. Predyktory: (Stała), Approval

b. Predyktory: (Stała), Approval, RequestAmount

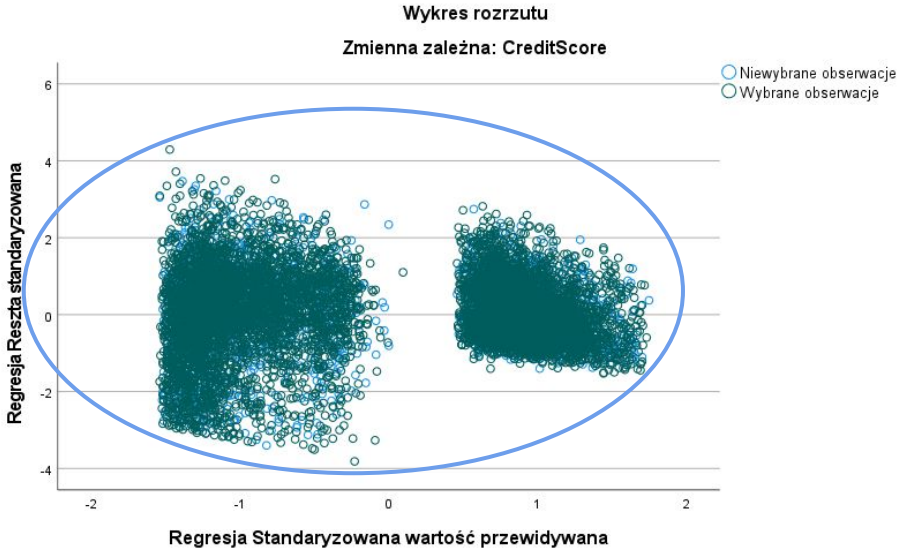
c. Predyktory: (Stała), Approval, RequestAmount, DebttoIncomeRatio

d. Jeśli nie podano inaczej, statystyki oparte są tylko na obserwacjach, dla których nowy_podział = uczacy.

e. Zmienna zależna: CreditScore

3. HOMOSKEDASTYCZNOŚĆ RESZT ✅

Sprawdzamy równość wariancji reszt. Badamy to na podstawie wykresu rozrzutu standaryzowanej reszty i standaryzowanej wartości przewidywanej. Jeśli da się je otoczyć okręgiem, owalem lub innym kształtem, nie mającym ostrych krańców to oznacza, że reszty mają równą wariancję.

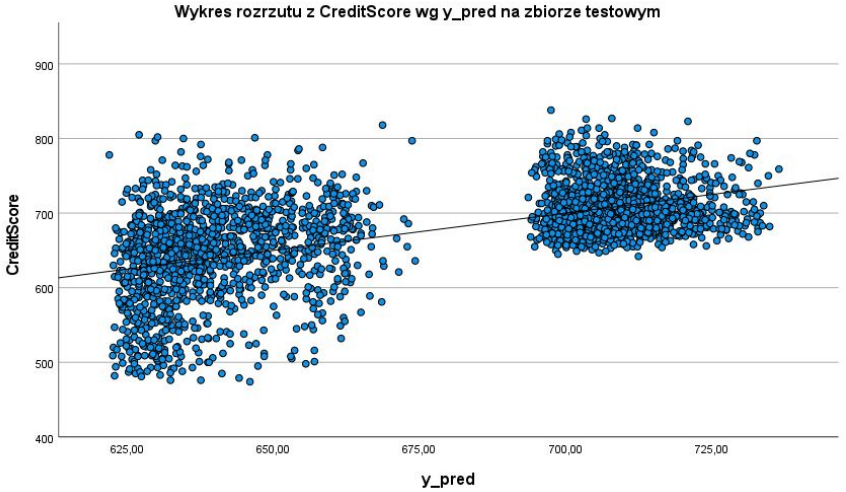


Wszystkie założenia są spełnione, a więc możemy uznać nasz model za wiarygodny!

Ostatecznie otrzymujemy następujące równanie regresji wielokrotnej:

$$Y = 621{,}482 + 71{,}468 \cdot X_1 + 0{,}001 \cdot X_2 + 19{,}808 \cdot X_3$$

- Y – wartość zmiennej celu CreditScore
- X_1 – wartość zmiennej Approval
- X_2 – wartość zmiennej RequestAmount
- X_3 – wartość zmiennej DebttoIncomeRatio



Model 2 – SIECI NEURONOWE

Również w tym modelu zmienną celu jest *'CreditScore'.* Predyktory zostały wykorzystane takie jak w regresji wielokrotnej, tzn. *'Approval', 'DebttoIncomeRatio',* '*RequestAmount*', biorąc pod uwagę że *'Interest'* oraz '*RequestAmount'* są ze sobą w korelacji.

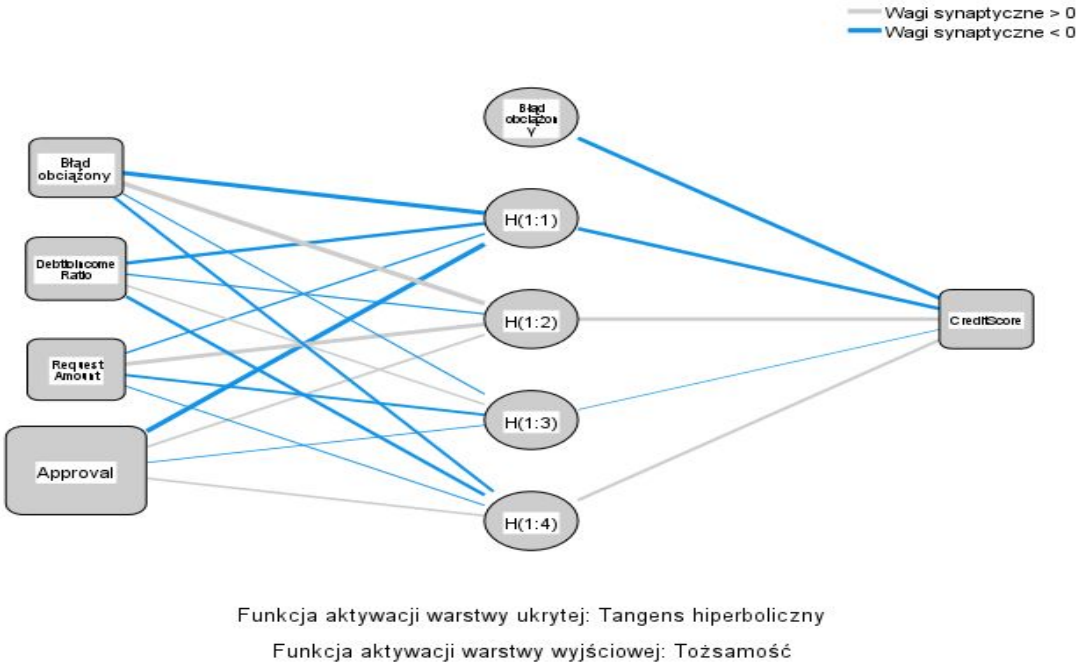
Zbiór danych został podzielony na próby uczący (50%), testowy (25%) oraz walidacyjny (25%). Dla potrzeb porównania dwóch modeli stworzono nową próbę uczącą i testową poprzez złączenie uczącego i testowego który stał się uczącym (75%) a walidacyjny - testowym (25%).

Informacja o analizowanych danych

		N	Procent
Próba	Uczący	5273	50,0%
	Testujący	2638	25,0%
	Walidacyjny	2636	25,0%
Ważnych		10547	100,0%
Wykluczone		0	
Ogółem		10547	

Model sieci neuronowej

W warstwie wejściowej znajdują się 3 neutrony (*‘Approval’, ‘DebttoIncomeRatio’,* ‘*RequestAmount’*), w warstwie wyjściowej 1 neutron (*‘CreditScore’)* oraz jest jedna warstwa ukryta.



Podsumowanie modelu i ocena ważności zmiennych

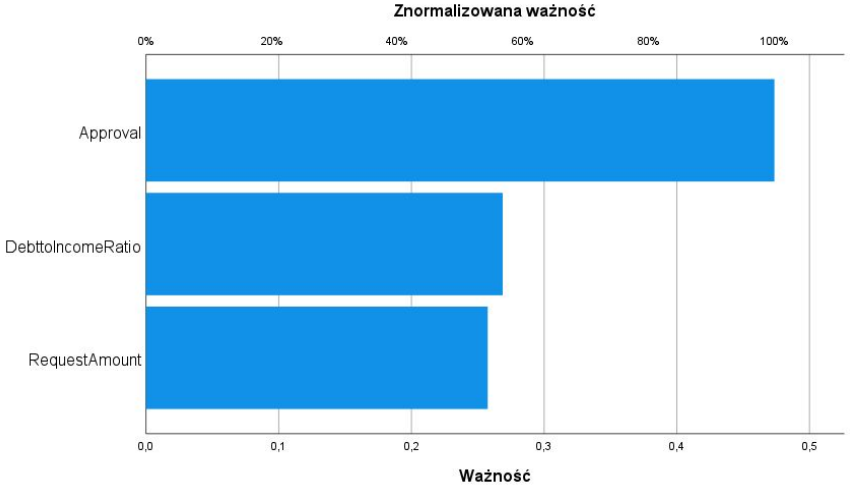
Błąd względny przy próbie uczącej wynosi 0,645 a dla próby testowej wynosi 0,636.

Podsumowanie modelu

Uczący	Suma kwadratów błędu	1699,182
	Błąd względny	,645
	Użyta reguła zatrzymywania	1 kolejnych kroków bez zmniejszenia wartości błędu[a]
	Czas uczenia	0:00:00,16
Testujący	Suma kwadratów błędu	841,265
	Błąd względny	,636
Walidacyjny	Błąd względny	,645

Zmienna zależna: CreditScore

a. Obliczenia błędów opierają się na próbie testowej.



Wyraźnie wynika, że *'Approval'* jest najlepszym wskaźnikiem do określenia *'CreditScore'.* Na następnym miejscu znalazła się zmienna *'DebttoIncomeRatio'.*

RMSE, MAE i MAPE

Spójrzmy na jakość szacowania. W tym pomogą nam wyliczone błędy:

	Zbiór uczący	Zbiór testowy
RMSE	50,03	48,88
MSE	2502,92	2389,25
MAE	37,75	37,04
MAPE	5,79%	5,67%

Statystyki opisowe

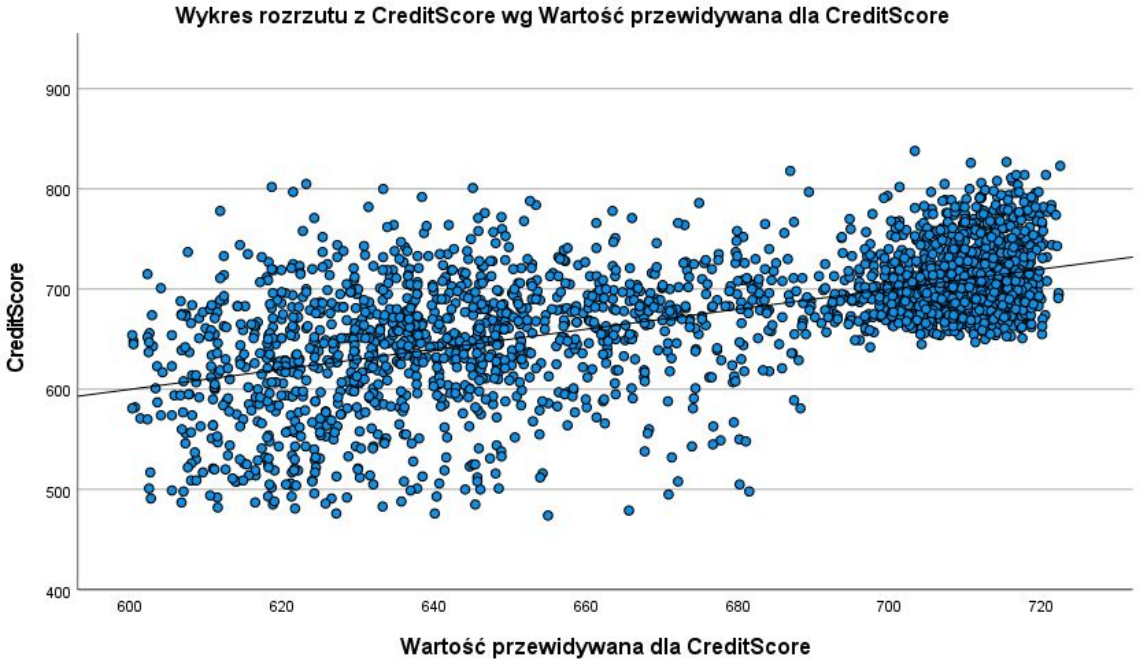
nowy_podział		N	Średnia
testowy	AE	2636	37,0373
	ME	2636	2389,2546
	N Ważnych (wyłączanie obserwacjami)	2636	
uczący	AE	7911	37,7533
	ME	7911	2502,9168
	N Ważnych (wyłączanie obserwacjami)	7911	

Statystyki opisowe

nowy_podział		N	Suma
testowy	APE	2636	149,51
	N Ważnych (wyłączanie obserwacjami)	2636	
uczący	APE	7911	457,95
	N Ważnych (wyłączanie obserwacjami)	7911	

Wykres rozrzutu na próbie testowej

Dla dobrego modelu powinny układać się one wzdłuż prostej y=x. Punkty znajdujące się poniżej tej prostej są niedoszacowane, a punkty powyżej są przeszacowane.



Podsumowanie

- wartości błędów (RMSE, MAPE, MAE) są zbliżone w obu modelach
- modele są wiarygodne, więc można uznać je za poprawne
- problem, z którym udało nam się wygrać, powstał przy podziale na zbiory. W przypadku MPL nastąpił podział na 3 zbiory (testowy, uczący i walidacyjny). Do regresji wielokrotnej potrzeba jedynie zbioru testowego i uczącego, więc zbiór testowy i uczący z MPL stał się uczącym (75%), a walidacyjny testowym (25%)
- wyrzucenie punktów wysokiej dźwigni ulepszyło oba modele

Raport wykonały:
Anna Cabaj i Aleksandra Grzegórska