Eksploracja Danych – Raport 3 Klaudia Pełczyńska

Analizowany zbiór danych znajduje się w pliku *bank_reg_training.csv*. Predyktorami są zmienne: **Approval, Debt_to_Income_Ratio, Request_Amount, Interest**, a zmienną celu jest zmienna **Credit Score**. Ziarno generatora wynosi 308285.

Podczas wczytywania danych, a także późniejszej analizy nie zostały zauważone braki danych. Poniższa tabelka pokazuje podstawowe miary statystyczne: średnia, odchylenie standardowe, kwantyle, wartości najmniejsze i największe dla każdej zmiennej ilościowej.

Zmienna	Średnia	Odchylenie std.	Min	25%	50%	75%	Max
Credit Score	674.06	66.953	393	650	685	714	844
Debt_to_Income_Ratio	0.1815	0.13457	0.00	0.0900	0.1600	0.2400	1.00
Request_Amount	13334.99	9416.033	500	6000	11000	19000	43000
Interest	6000.74	4237.215	225	2700	4950	8550	19350

Tabela 1. Tabela podstawowych miar statystycznych

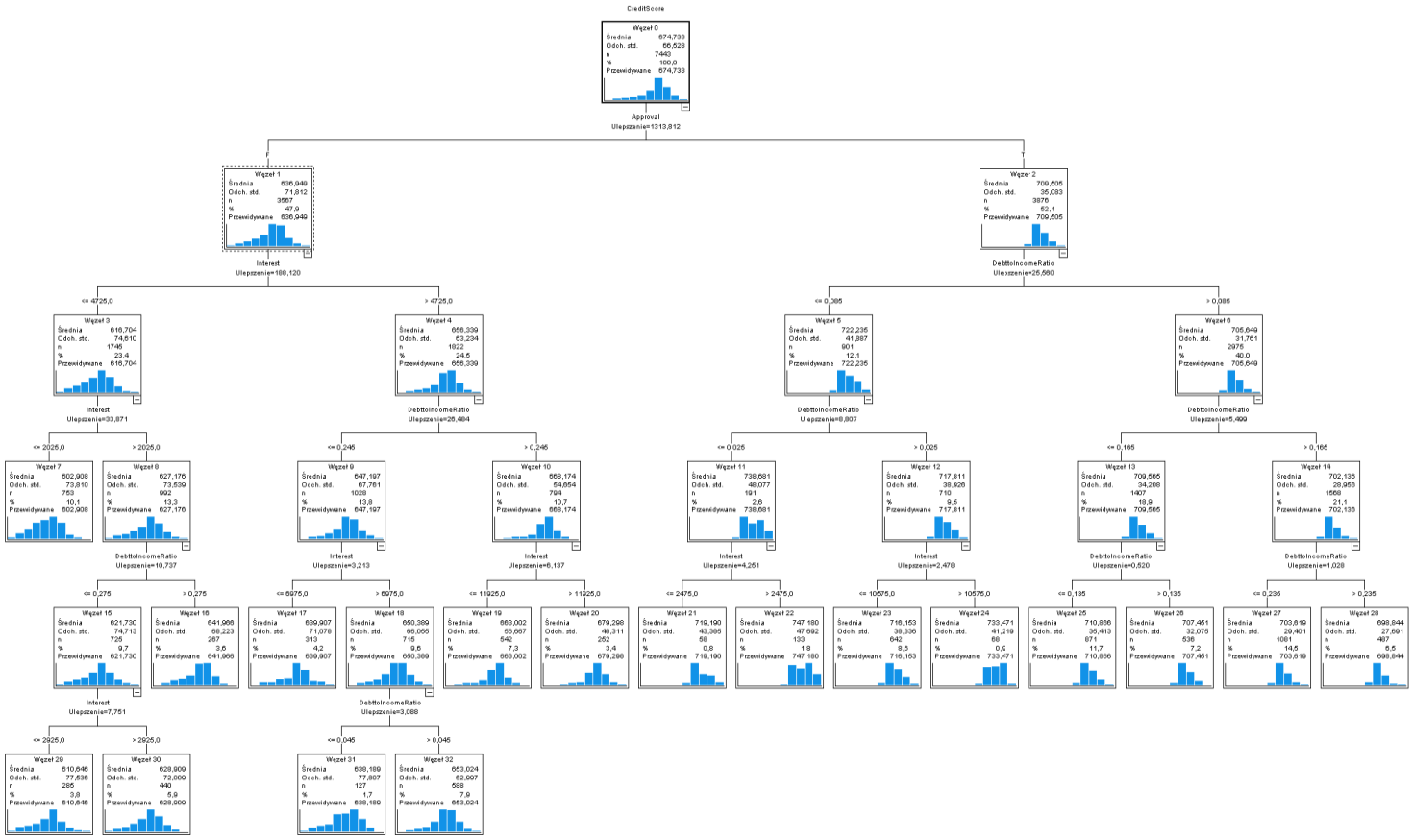
	Credit Score	Debt_to_Income_Ratio	Interest	Request_Amount	Approval
Credit Score	1	- 0.077	0.139	0.139	0.548
Debt_to_Income_Ratio	- 0.077	1	0.131	0.131	- 0.269
Interest	0.139	0.131	1	1,000	- 0.061
Request_Amount	0.139	0.131	1,000	1	- 0.061
Approval	0.548	- 0.269	- 0.061	- 0.061	1

Tabela 2. Tabela korelacji (istotność każdej korelacji była mniejsza niż 0.005)

Patrząc na tabelę 2 możemy odrzucić jedną ze zmiennych **Interest** lub **Request_Amount**, ponieważ ich korelacja jest równa 1.000. Podczas budowy wybranych przeze mnie modeli będę używać zmiennych **Approval**, **Debt_to_Income_Ratio** i **Interest**. Wybrane modele to drzewo CART i sieć MLP i zostały one zbudowane w programie SPSS. Pierwotnie chciałam wykonać również model regresji liniowej jednak nie zostały spełnione wszystkie założenia.

1. Drzewo CART

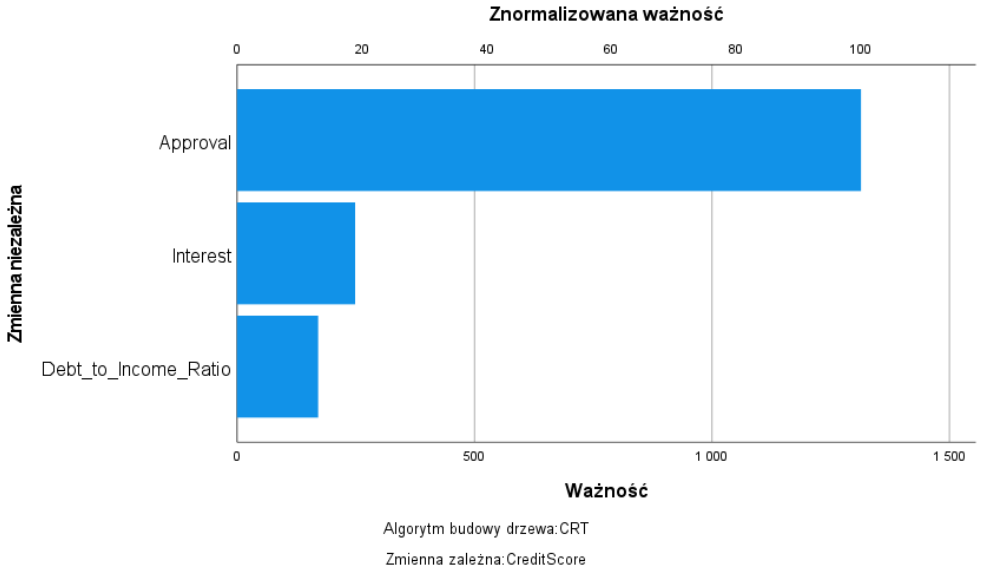
Pierwszym wytrenowanym modelem jest drzewo CART o maksymalnej głębokości drzewa 5, minimalnej liczba obserwacji w węźle nadrzędnym 100 i minimalnej liczba obserwacji w węźle podrzędnym 50 (wartości wybrane ze względu na liczbę wszystkich obserwacji – jest ich ponad 10 000). Otrzymane drzewo ma 33 węzły, w tym 17 liści.



Rysunek 1. Drzewo CART

Analizując przebieg drzewa, można zauważyć, że po wprowadzeniu danych model sprawdza czy zmienna **Approval** ma wartość F lub T tzn., czy wniosek o pożyczkę/kredyt został zaakceptowany czy nie. Jeśli wniosek został zaakceptowany to porównuje kolejno zmienne **Interest**, **Debt_to_Income_Ratio** itd. Model zwraca przewidywaną wartość dla zmiennej **Credit Score**.

Rysunek 2 przedstawia ważność zmiennych dla powyższego modelu. Najważniejszą zmienną okazała się zmienna **Approval** pomimo, że została użyta tylko raz.

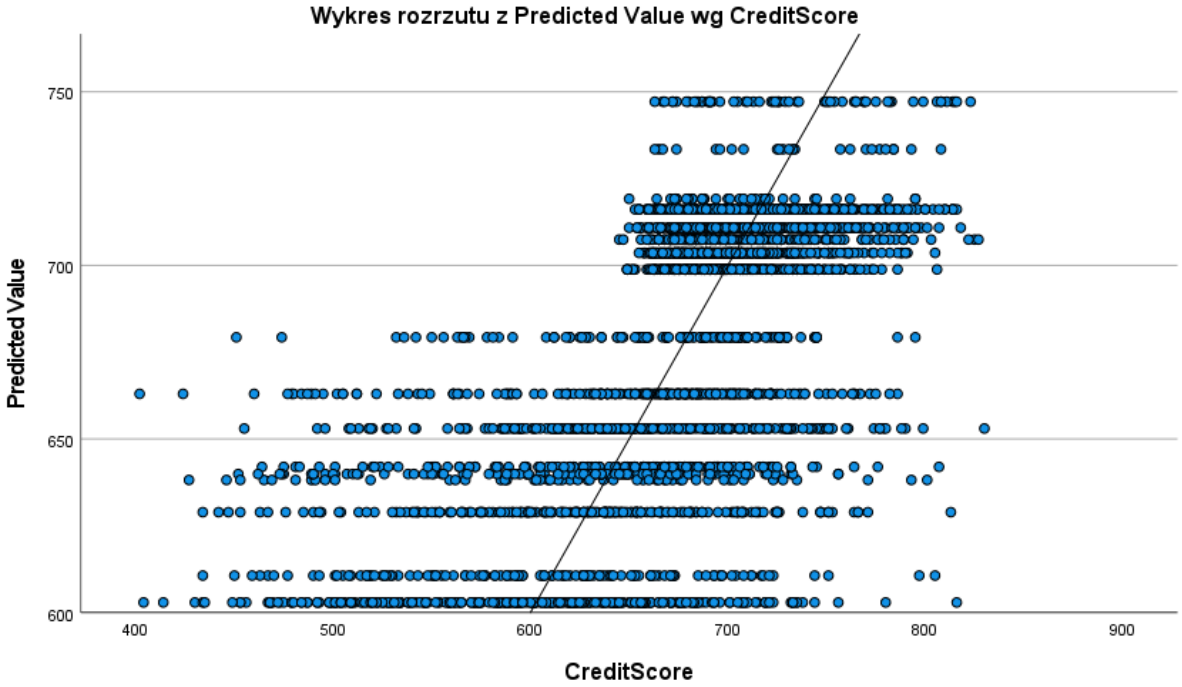


Rysunek 2. Ważność predyktorów dla drzewa CART.

Funkcja	Dane uczące	Dane testowe
RMSE	52.76	53.92
MAE	39.12	39.64
MAPE	6.12%	6.26%

Tabela 3. Wartości funkcji służących do określenia jakości szacowania w podziale na zbiory uczące i testowe

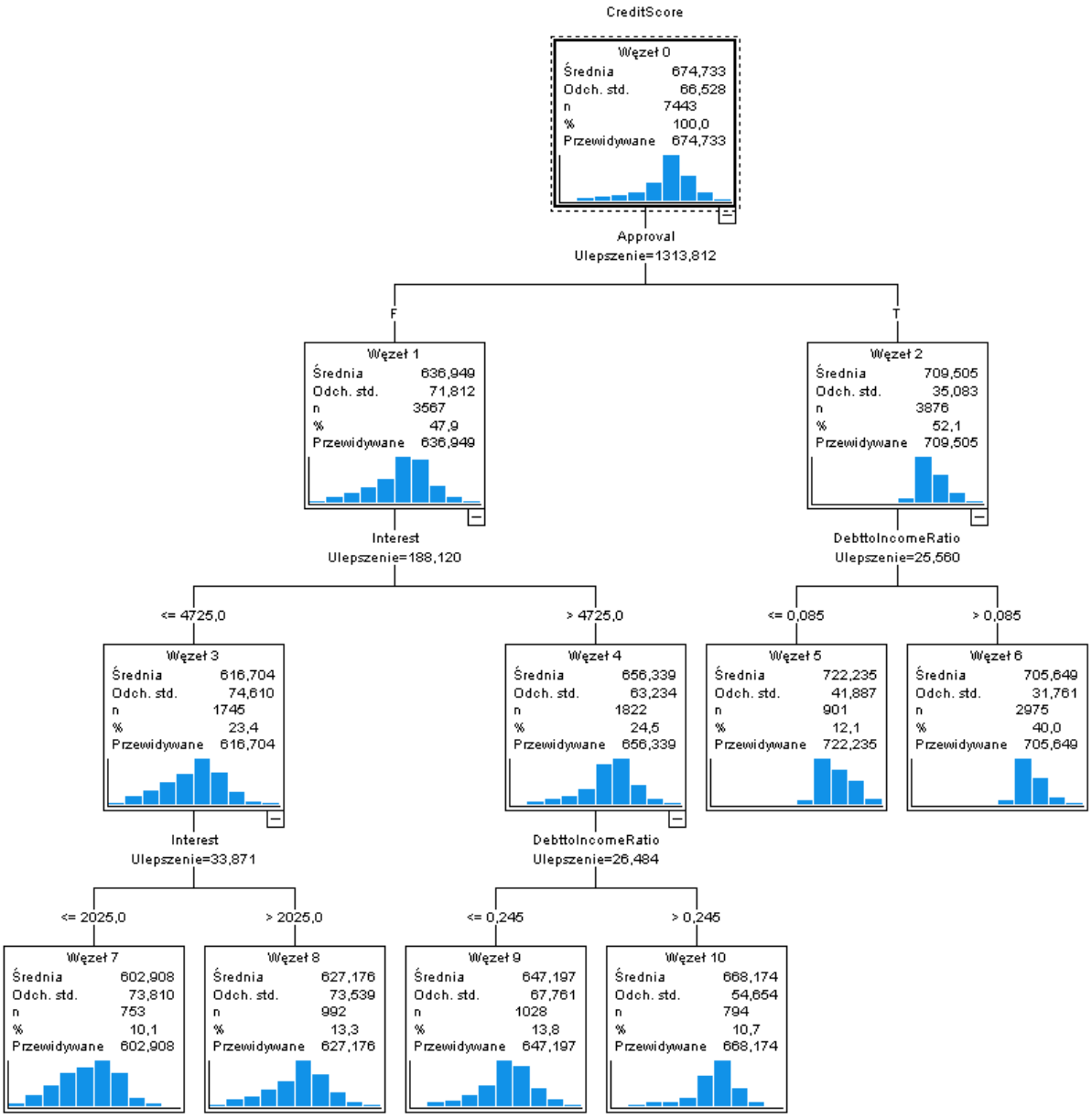
Wartości funkcji RMSE, MAE i MAPE są niższe dla danych treningowych niż dla danych testowych, co świadczy o delikatnym przetrenowaniu modelu. Jednak mimo wszystko MAE jest dość niskie w porównaniu do wartości rzeczywistych zmiennej celu.



Rysunek 3. Wykres wartości przewidywanych względem obserwowanych na zbiorze testowym

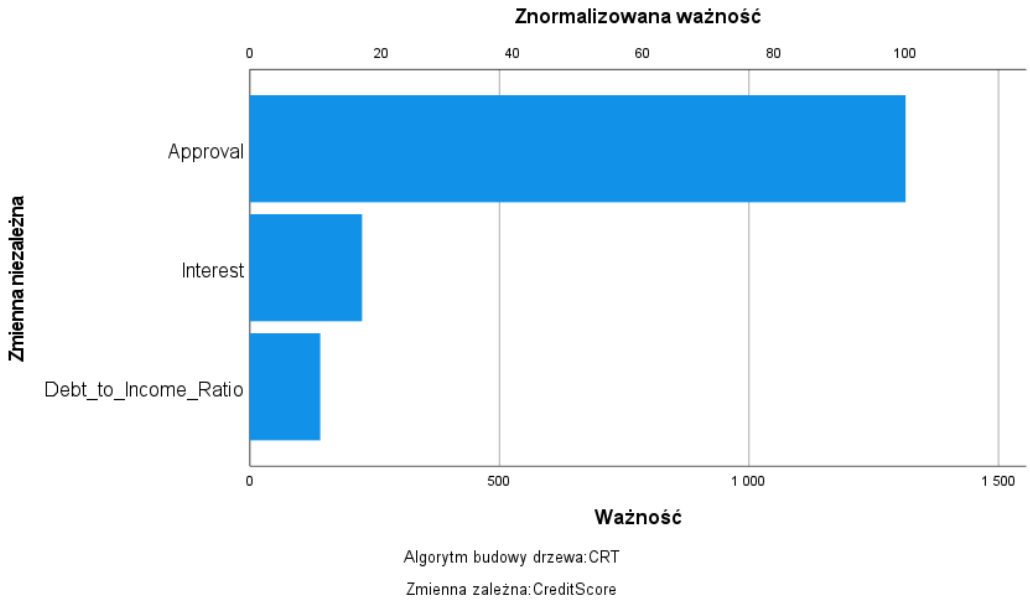
Na rysunku 3 jest przedstawiony wykres wartości przewidywanych względem obserwowanych wykonany na zbiorze testowym. Dodatkowo jest na nim zaznaczona prosta y = x, dzięki której możemy stwierdzić, że nasz model najlepiej przewiduje w rzeczywistości przyjmują wartości około 700. Każde piętro punktów na wykresie odpowiada jednemu liściu drzewa.

2. Drzewo CART z przycinaniem



Rysunek 4. Przycięte drzewo CART

Drzewo CART przedstawione na rysunku 4 podobnie jak poprzednie ma maksymalną głębokość 5, minimalna liczba obserwacji w węźle nadrzędnym to 100 i minimalna liczba obserwacji w węźle podrzędnym to 50. Powyższe drzewo ma 11 węzłów, w tym 6 liści i jest głębokości 3.



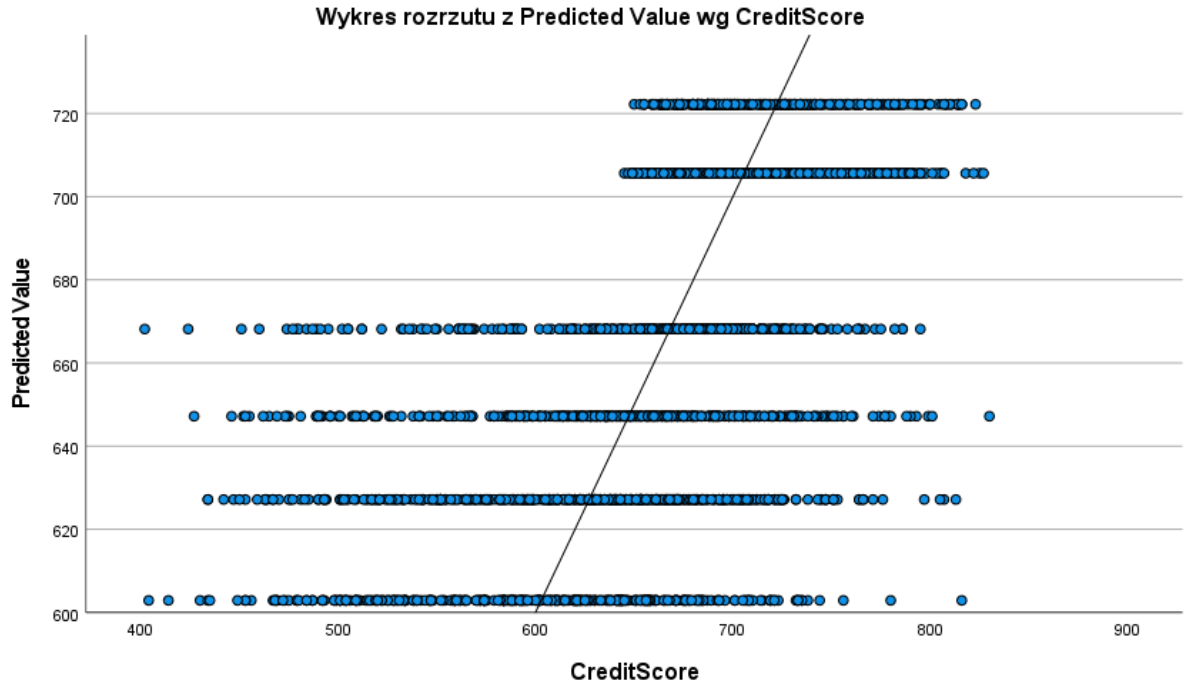
Rysunek 5. Ważność predyktorów dla przyciętego drzewa CART

W przypadku tego modelu podobnie jak poprzednio zmienna **Approval** jest najważniejszą zmienną.

Funkcja	Dane uczące	Dane testowe
RMSE	53.27	54.44
MAE	39.54	39.97
MAPE	6.18%	6.31%

Tabela 4. Wartości funkcji służących do określenia jakości szacowania w podziale na zbiory uczące i testowe

Porównując ten model do pierwszego drzewa ten wypada trochę gorzej. Podobnie jak pierwsze drzewo CART ten model jest delikatnie przeuczony.

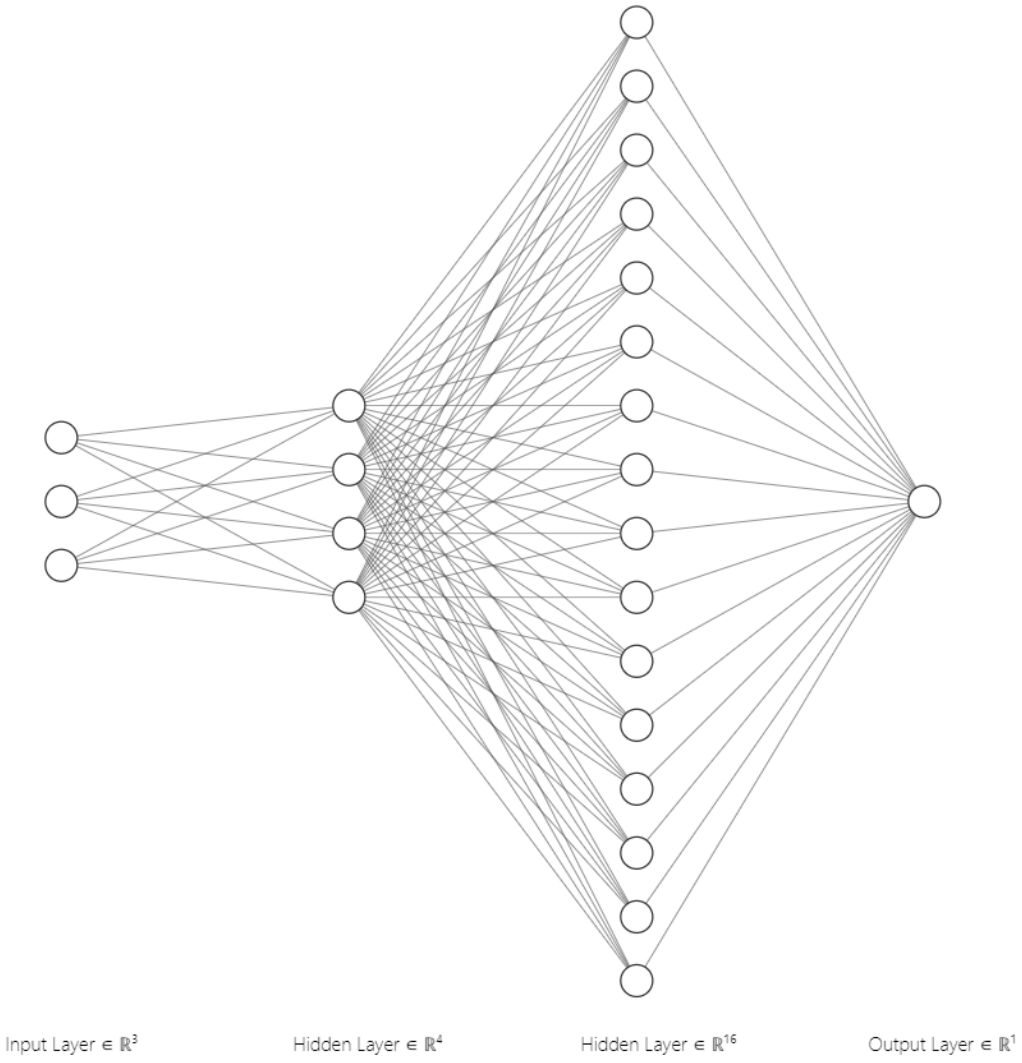


Rysunek 6. Wykres wartości przewidywanych względem obserwowanych na zbiorze testowym

Rysunek 6 podobnie jak rysunek 3 przedstawia wykres wartości przewidywanych względem obserwowanych na zbiorze testowym. Ten wykres również posiada prostą y = x. Na powyższym rysunku można zauważyć ma mniej ‘’pięter’’ obserwacji. Jest to spowodowane tym, że nasze drugie drzewo zostało przycięte i posiada tylko 6 liści, a pierwsze miało ich aż 17. Najlepiej zostały przewidziane wartości, które w rzeczywistości znajdują się w przedziale od 600 do 700.

3. Sieć MLP

Ostatnim zbudowanym modelem jest sieć MLP. Otrzymana się posiada warstwę wejściową, a w niej 3 neurony, dwie warstwy ukryte, w których są kolejno 4 neurony i 16 i warstwę wynikową, w której jest 1 neuron. Funkcją aktywacji warstw ukrytych jest tangens hiperboliczny, a warstwy wynikowej jest tożsamość.

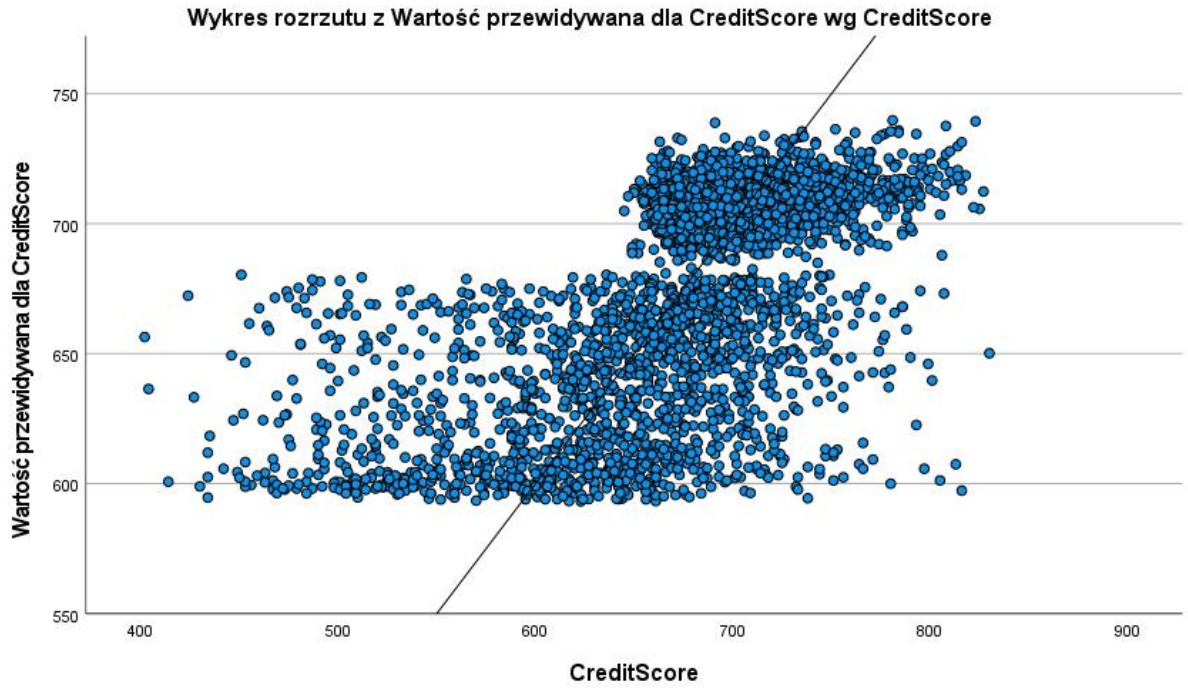


Rysunek 7. Sieć neuronowa

Funkcja	Dane uczące	Dane testowe
RMSE	53.84	54.10
MAE	39.99	39.66
MAPE	6.24%	6.25%

Tabela 5. Wartości funkcji służących do określenia jakości szacowania w podziale na zbiory uczące i testowe

Analizując dane w tabeli 5 możemy zauważyć delikatne przeuczenie modelu, jednak porównując to z danymi dla drzew CART przeuczenie tutaj jest najmniejsze.



Rysunek 8. Wykres wartości przewidywanych względem obserwowanych na zbiorze testowym

Rysunek 8 przedstawia wykres wartości przewidywanych względem obserwowanych dla sieci neuronowej na zbiorze testowym. Na tym wykresie podobnie jak na wykresach dla drzew CART znajduje się prosta y = x. Z wykresu można odczytać, że najlepiej przewidzianymi wartościami są te, które w rzeczywistości przyjmują wartości z przedziału 600 – 750.

4. Wnioski

Wszystkie otrzymane wyniki dały podobne wartości. Jednak najlepszym modelem okazała się sieć neuronowa, ponieważ to dla niej przeuczenie było najniższe. Niestety jest to model najbardziej skomplikowany, który najtrudniej wytłumaczyć i jest najkosztowniejszy obliczeniowo. Dlatego jeżeli chcemy wiedzieć dlaczego otrzymaliśmy daną wartość lepiej byłoby użyć pierwszego modelu tj. drzewa CART, ponieważ w porównaniu z drugim modelem (przyciętym drzewem) on dawał lepsze wyniki.