Raport 2 - MSI

Tomasz Sewastynowicz

Uniwersytet im. Adama Mickiewicza w Poznaniu

10.12.2023 r.

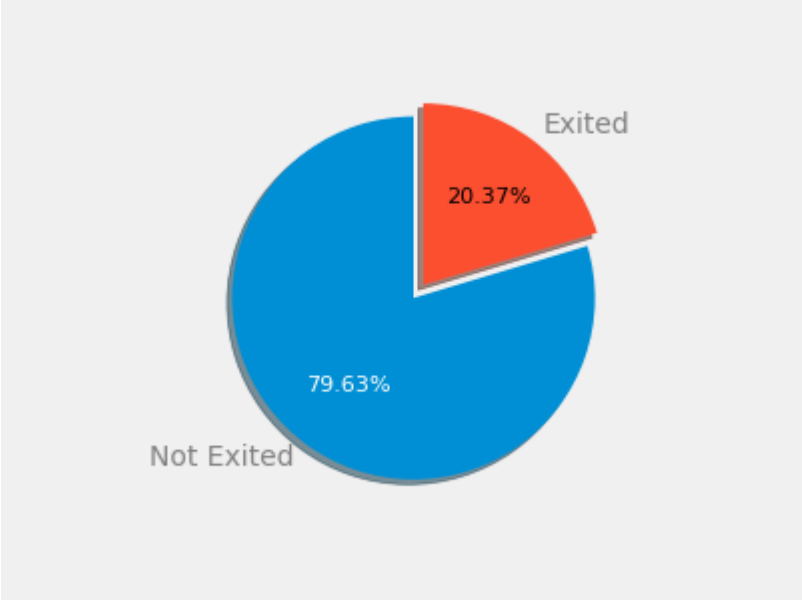
1. Cel pracy

Raport skupia się na zastosowaniu klasyfikacji do prognozowania, czy dana osoba zdecyduje się pozostać klientem firmy, czy też podjąć decyzję o rezygnacji z usług. Głównym celem badania jest ocena skuteczności sieci neuronowych w klasyfikowaniu klientów, co umożliwi precyzyjne przewidywanie, czy klient zachowa swoje związki z firmą, czy też podejmie decyzję o odejściu.

W ramach zadania został wykorzystany zbiór danych "Churn Modelling", z platformy Kaggle *(https://www.kaggle.com/datasets/shubh0799/churn-modelling/data)*. Zbioru tego użyto do trenowania i testowania modeli klasyfikacyjnych, które mają na celu zrozumienie kluczowych czynników wpływających na lojalność klientów oraz przewidywanie, jakie decyzje podejmą w odniesieniu do korzystania z usług firmy.

2. Zbiór danych

Zbiór ten obejmuje informację o 13 cechach o 10 000 klientów, takie jak unikalne identyfikatory klientów *(CustomerId)*, nazwisko klienta *(Surname)*, punkt kredytowy *(CreditScore)*, kraj zamieszkania *(Geography)*, płeć *(Gender)*, wiek *(Age)*, okres korzystnia z usług firmy *(Tenure)*, saldo na koncie *(Balance)*, liczba posiadanych produktów *(NumOfProducts),* posiadanie karty kredytowej *(HasCrCard)*, aktywność członkowska *(IsActiveMember)*, szacowany dochód *(EstimatedSalary)*, oraz zmienna docelowa *(Exited)*, która określa, czy klient zdecydował dalej korzystać z usług firmy czy też zrezygnować. Do zbudowania sieci neuronowej kolumny z danymi o numerze identyfikatora i nazwiska klientów nie zostały wykorzystane.



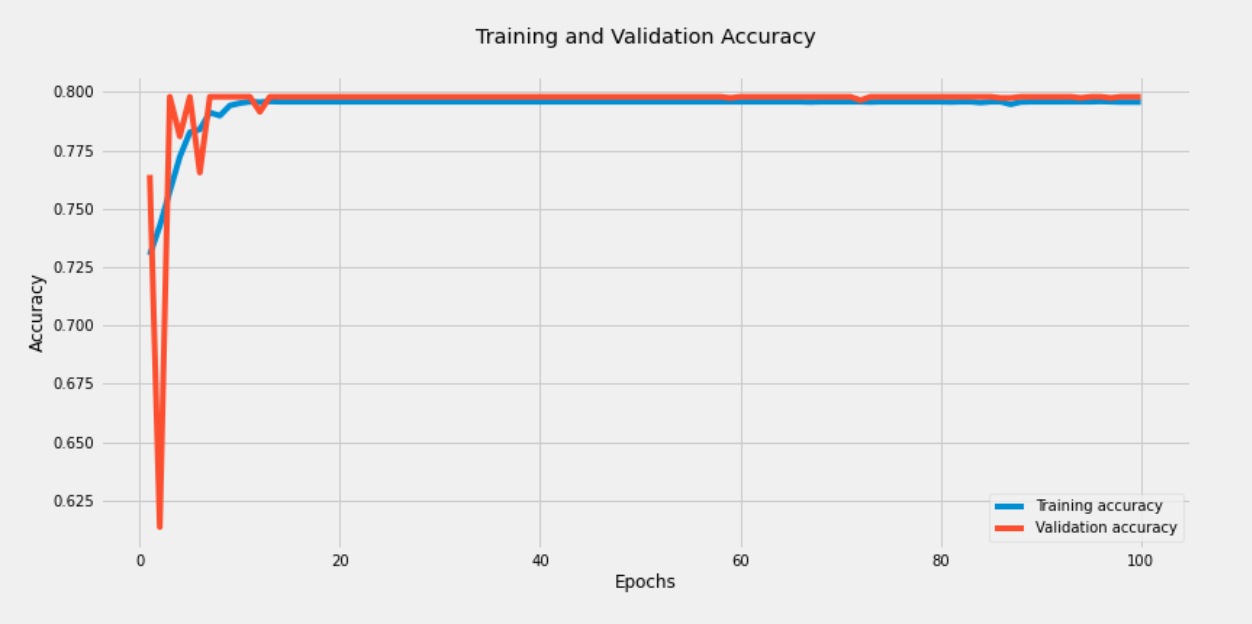
Wykres 1. Ilość klientów, którzy zrezygnowali z usług firmy.

Na **rysunku 1** jest widoczne, że 20.37% klientów zdecydowało się zrezygnować z usług firmy, a 79.63% zdecydowało się na pozostanie. Większość klientów firmy pochodzi z Francji, jednak najwięcej osób, które zdecydowały się zrezygnować z usług firmy pochodzi z Niemiec. Proporcjonalnie kobiety częściej rezygnują z usług firmy niż mężczyźni. Pomimo że większość klientów posiada 1 lub 2 produkty, to proporcjonalnie najwięcej rezygnacji odnotowano wśród osób posiadających 3 i 4 produkty, a mniej wśród klientów posiadających 1 produkt. Może to sugerować, że korzystanie z większej ilości produktów firmy zwiększa tendencję do rezygnacji. Optymalnym scenariuszem dla firmy jest, gdy klient posiada 2 produkty. Jak można się spodziewać, nieaktywni członkowie mają większą skłonność do odejścia, a ogólny odsetek nieaktywnych członków jest również bardzo wysoki **(Rysunek 1.)**

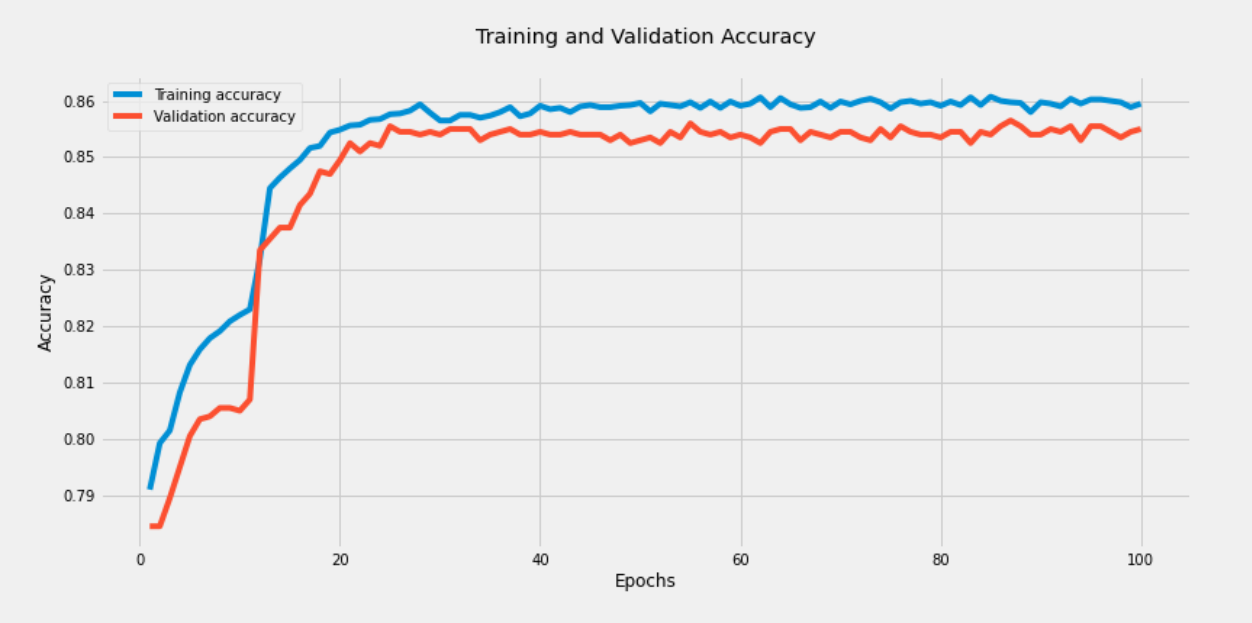
Rysunek 1. Rozkład klientów, którzy zrezygnowali z usług firmy na poszczególne zmienne.

3. Modele sieci neuronowych

Łącznie zostało stworzonych 6 modeli, których dokładność na zbiorze walidacyjnym wynosiła od 79.85% do 86.25%. Modele różniły się między sobą pod względem liczby neuronów w poszczególnych warstwach, ilości warstw ukrytych oraz zastosowanych technik przeciwdziałania przeuczeniu, takich jak dropout czy standaryzacja. W początkowych modelach zdecydowano się na niewielkie parametry, takie jak pojedyncza warstwa ukryta i niewielka liczba neuronów, aby uniknąć szybkiego przeuczenia. W późniejszych modelach została zwiększana ilość warstw ukrytych i modyfikowana ilość neuronów.

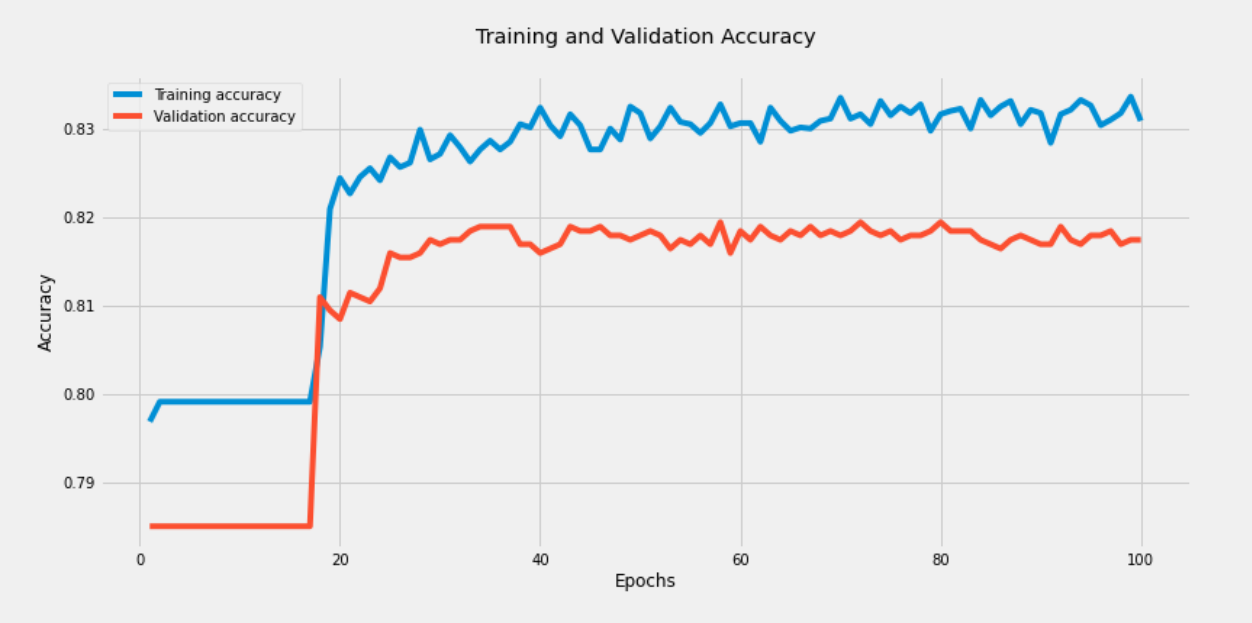


Rysunek 2. Accuracy pierwszego modelu.



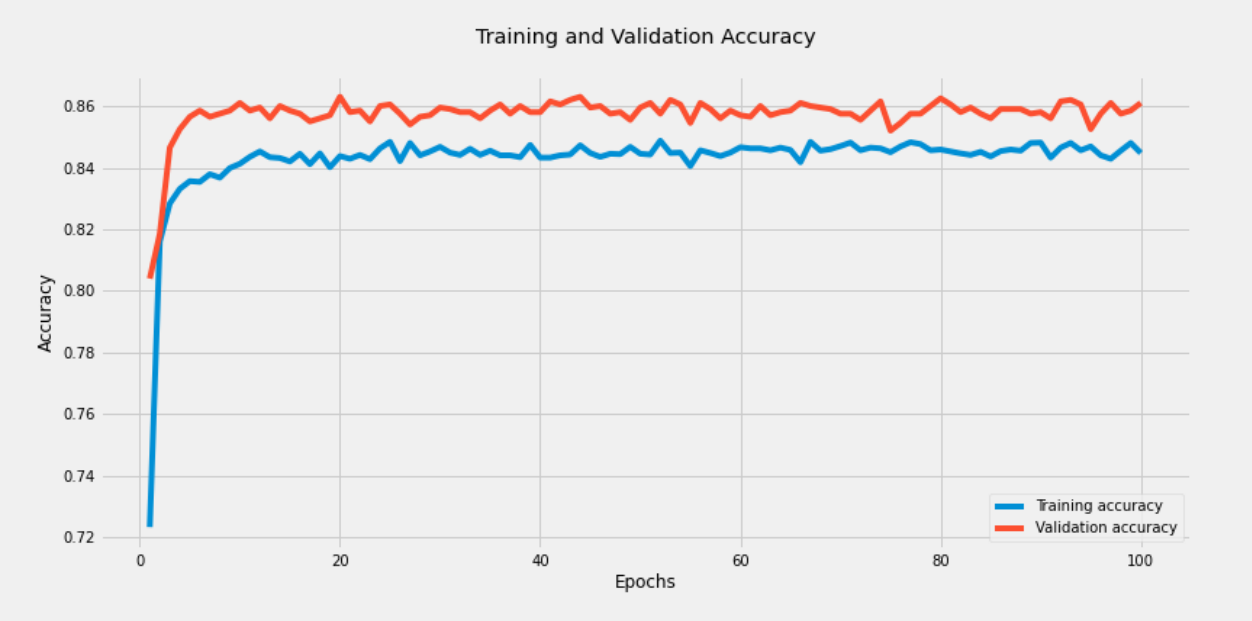
Rysunek 3. Accuracy drugiego modelu.

Na **rysunku 2** i **rysunku 3** pierwszy i drugi model są najprostszymi modelami posiadającymi jedną warstwę ukrytą o ilości 6 neuronów. W przypadku drugiego modelu została zastosowana tylko standaryzacja danych, co pozwoliło osiągnąć wynik 85.5% accuracy, w porównaniu do 79.85% w pierwszym modelu.



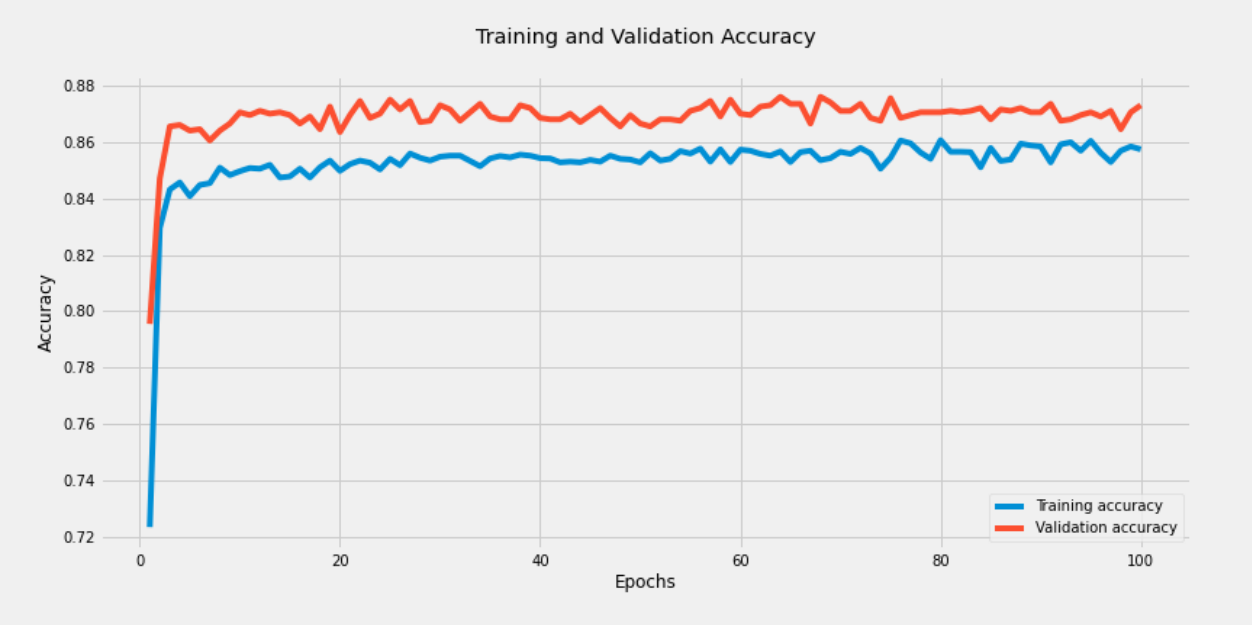
Rysunek 4. Accuracy trzeciego modelu.

W przypadku trzeciego modelu dodanie dodatkowej warstwy ukrytej oraz dropoutu spowodowało niewielki spadek accuracy.



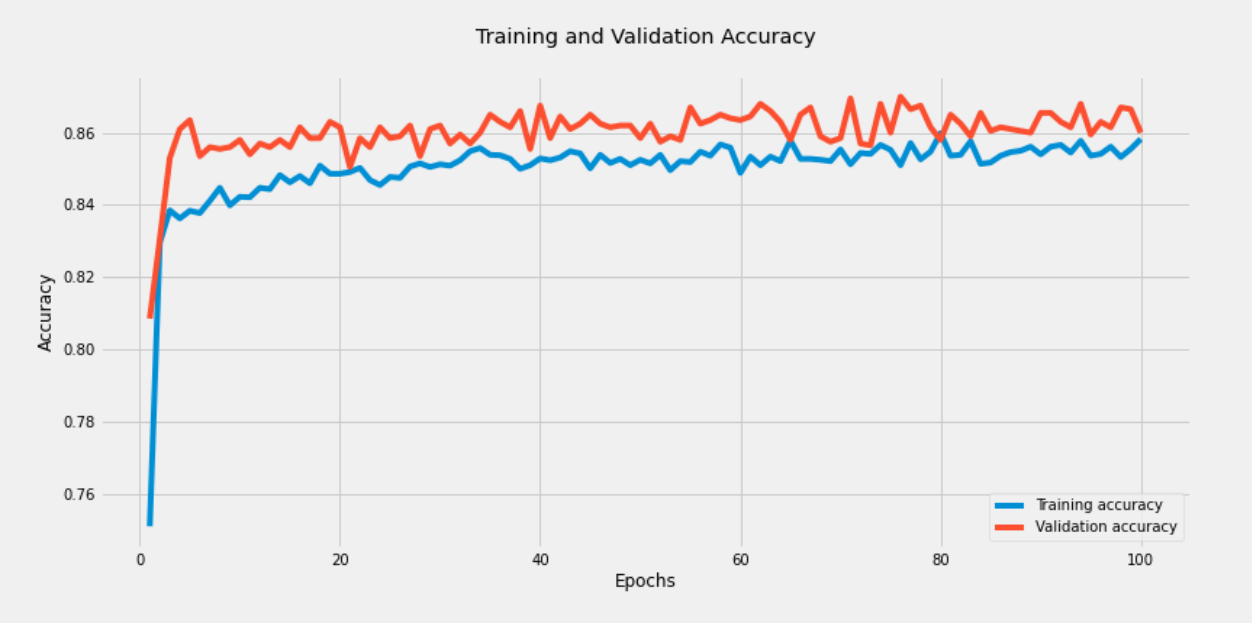
Rysunek 5. Accuracy czwartego modelu.

Czwarty model osiągnął accuracy dla zbioru walidacyjnego na poziomie 86.1%. Została w nim dodana nowa warstwa ukryta oraz dropout. Ilość neuronów pozostała w nim bez zmian.



Rysunek 6. Accuracy piątego modelu.

Piąty model osiągnął najlepsze accuracy spośród modeli na poziomie 87.3%. Została w nim znacznie zwiększona ilość neuronów oraz dropout rate zmieniony z 0.1 na 0.05, co wpłynęło w znaczny sposób accuracy. Model łącznie składa się z 2 warstw ukrytych po 12 neuronów.



Rysunek 7. Accuracy szóstego modelu.

W przypadku szóstego modelu zostało w nim dodane więcej warstw ukrytych i więcej neuronów. Accuracy wyniosło 86%.

Niezależnie od zmian wprowadzonych w ilości warstw ukrytych i ilości neuronów, w kolejnych modelach accuracy wynosiła od 85% do 87%.

4. Podsumowanie

Mimo, że piąty model osiągnął najlepsze accuracy na poziomie 87.3%, to nie stanowi to, że jest najlepszym dostępnym modelem. Niezależnie od zwiększania ilości warstw ukrytych i zmianie ilości neuronów, accuracy różnych modeli wahała się w przedziale od 85 do 87% i od różnych powtórzeń działanie jednego modelu różniło się od poprzedniego wyniku w okolicach 1-2%. Nie udało się osiągnąć wyższego accuracy, ale bardzo prawdopodobne jest, że accuracy na poziomie 87-88% jest maksymalnym accuracy dla tego zbioru danych. Z dostępnych danych nie

da się w lepszym stopniu przewidywać czy klient zrezygnuje z usług firmy i w przypadku konieczności na stworzenia lepszego modelu powinniśmy zgromadzić więcej danych o kliencie, aby stworzyć model o lepszym accuracy.