DATA MINING
PROJEKT

STATLOG (GERMAN CREDIT DATA) DATA SET

OPRACOWAŁA:
ALEKSANDRA GRZESZCZUK
NUMER ALBUMU: 255707

SPIS TREŚCI

1 CZĘŚĆ I - 30.04.2023 **2**
1.1 Analiza opisowa + wizualizacja danych 2
1.2 Klasyfikacja wraz z oceną dokładności 18
1.2.1 Liniowa analiza dyskryminacyjna (LDA) 20
1.2.2 Kwadratowa analiza dyskryminacja (QDA) 24
1.2.3 Metoda k - najbliższych sąsiadów (k -NN) 26
1.2.4 Metoda sieci neuronowych 27

1 CZĘŚĆ I - 30.04.2023

1.1 Analiza opisowa + wizualizacja danych

Kiedy bank otrzymuje wniosek o pożyczkę, na podstawie profilu wnioskodawcy musi podjąć decyzję, czy zatwierdzić pożyczkę, czy nie. Z tą decyzją wiążą się dwa rodzaje ryzyka:

- jeśli wnioskodawca ma dobre ryzyko kredytowe, czyli jest prawdopodobne, że spłaci pożyczkę, wówczas brak zatwierdzenia pożyczki danej osobie skutkuje utratą działalności dla banku
- jeśli wnioskodawca jest obarczony złym ryzykiem kredytowym, czyli nie jest prawdopodobne, aby spłacił pożyczkę, wówczas zatwierdzenie pożyczki dla tej osoby skutkuje stratą finansową dla banku

Aby zminimalizować stratę z punktu widzenia banku, potrzebna jest reguła decyzyjna dotycząca tego, komu udzielić zgody na pożyczkę, a komu nie. Profile demograficzne i społeczno - ekonomiczne wnioskodawcy są brane pod uwagę przez zarządzających pożyczkami przed podjęciem decyzji w sprawie jego wniosku o pożyczkę.

Niemieckie dane kredytowe (*German Credit Data* [1]) zawierają dane dotyczące 21 zmiennych i klasyfikację, czy wnioskodawca jest uznawany za dobre, czy złe ryzyko kredytowe dla 1000 osób ubiegających się o pożyczkę.

W zbiorze danych jest łącznie 21 atrybutów. Ich opisy i szczegóły zestawiono poniżej:

- *Status of existing checking account* - zmienna typu $factor$ - status istniejących rachunków bankowych klienta
- *Duration in month* - zmienna typu *numeric* - czas trwania kredytu w miesiącach
- *Credit history* - zmienna typu $factor$ - historia kredytowa
- *Purpose* - zmienna typu *character* - cel, na który brany jest kredyt
- *Credit amount* - zmienna typu *numeric* - ilość pieniędzy wzięta na kredyt
- *Savings account and bonds* - zmienna typu $factor$ - konto oszczędnościowe i obligacje
- *Present employment since* - zmienna typu $factor$ - okres obecnego zatrudnienia
- *Installment rate in percentage of disposable income* - zmienna typu *numeric* - stopa raty jako procent dochodu do dyspozycji
- *Personal status and sex* - zmienna typu $factor$ - status osobisty i płeć
- *Other debtors or guarantors* - zmienna typu $factor$ - inni dłużnicy lub poręczyciele
- *Present residence since* - zmienna typu *numeric* - okres obecnego miejsca zamieszkania
- *Property* - zmienna typu $factor$ - co kredytobiorca posiada na własność

[1] Dane dostępne pod adresem `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`

- *Age* - zmienna typu *numeric* - wiek w latach
- *Other installment plans* - zmienna typu $factor$ - inne plany ratalne
- *Housing* - zmienna typu $factor$ - nieruchomość, czy jest na własność, wynajmowana
- *Number of existing credits at this bank* - zmienna typu *numeric* - liczba istniejących kredytów w tym banku
- *Job* - zmienna typu $factor$ - czy jest to wykwalifikowany pracownik, niewykwalifikowany
- *Number of people being liable to provide maintenance for* - zmienna typu *numeric* - liczba osób zobowiązana do utrzymania kredytobiorcy
- *Telephone* - zmienna typu $factor$ - numer telefonu
- *Foreign worker* - zmienna typu $factor$ - czy jest to obcokrajowy pracownik
- *Response* - zmienna typu $factor$ - zdolność kredytowa - dobra lub zła

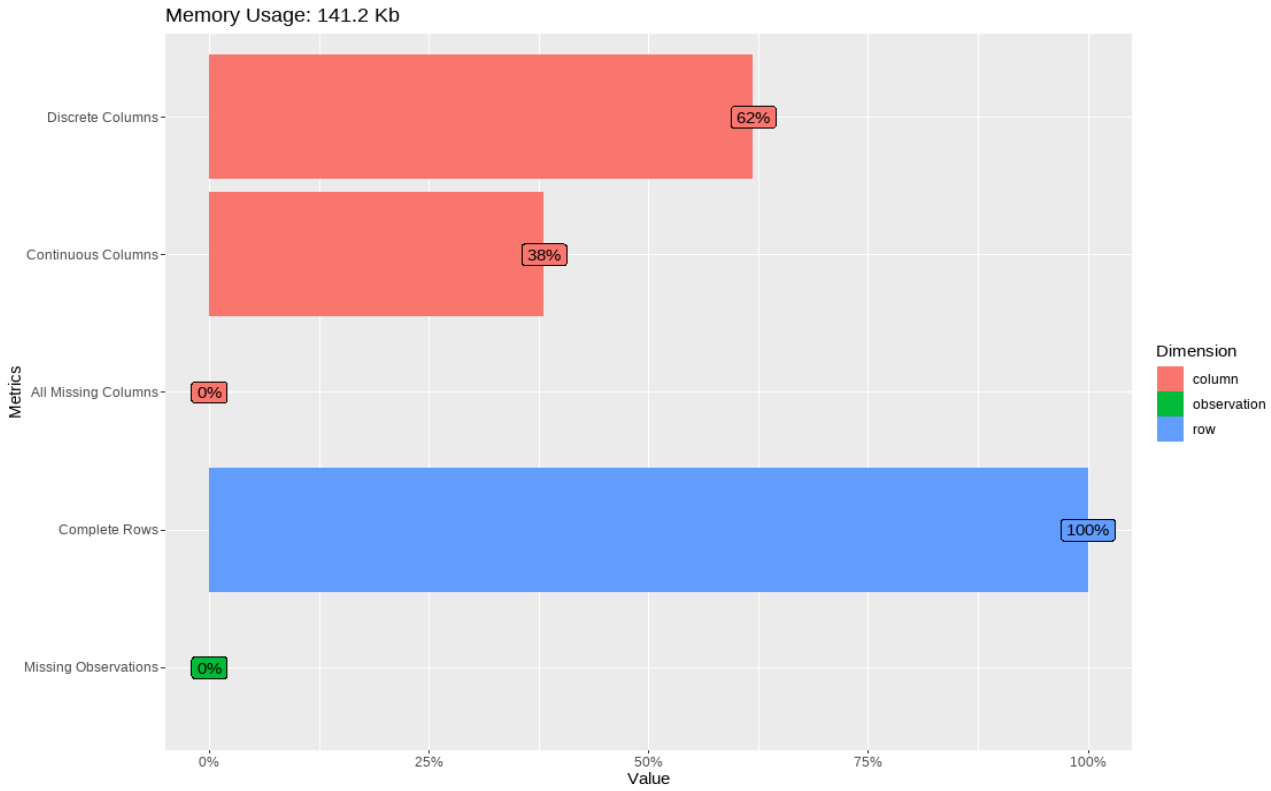
Biblioteka `DataExplorer` [2] zajmuje się zautomatyzowanym procesem eksploracji danych na potrzeby zadań analitycznych oraz modelowania predykcyjnego. Dzięki temu można między innymi skupić się na rozumieniu danych. Funkcje dostępne w tym pakiecie skanują oraz analizują każdą zmienną, po czym je wizualizują za pomocą podstawowych, typowych technik graficznych.

Na podstawie funkcji `introduce` oraz `plot_intro` z powyższej biblioteki dokonamy podsumowania podstawowych informacji naszych danych.

	Values
Rows	999.00
Columns	21.00
Discrete columns	13.00
Continuous columns	8.00
All missing columns	0.00
Total missing values	0.00
Complete rows	999.00
Total observations	20979.00

Tabela 1: Podstawowe informacje dla danych

[2] W `https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html` znajduje się przykładowe wprowadzenie do biblioteki `DataExplorer` wykonane przez Boxuan Cui, w roku 2020, na podstawie danych `nycflights13` zawierających informacje dotyczące lotów, które odleciały w 2013 roku z Nowego Jorku.



Rysunek 1: Wykres podstawowych informacji dla danych - opracowanie własne

Dobrą wiadomością jest to, że w zbiorze nie ma brakujących wartości. Struktura danych wygląda na spójną.

Funkcja `describe` z biblioteki `dkloork` [3] oblicza statystyki opisowe dla zmiennych ilościowych. Na jej podstawie możemy znaleźć między innymi średnią, odchylenie standardowe czy wartość kurtozy naszych danych.

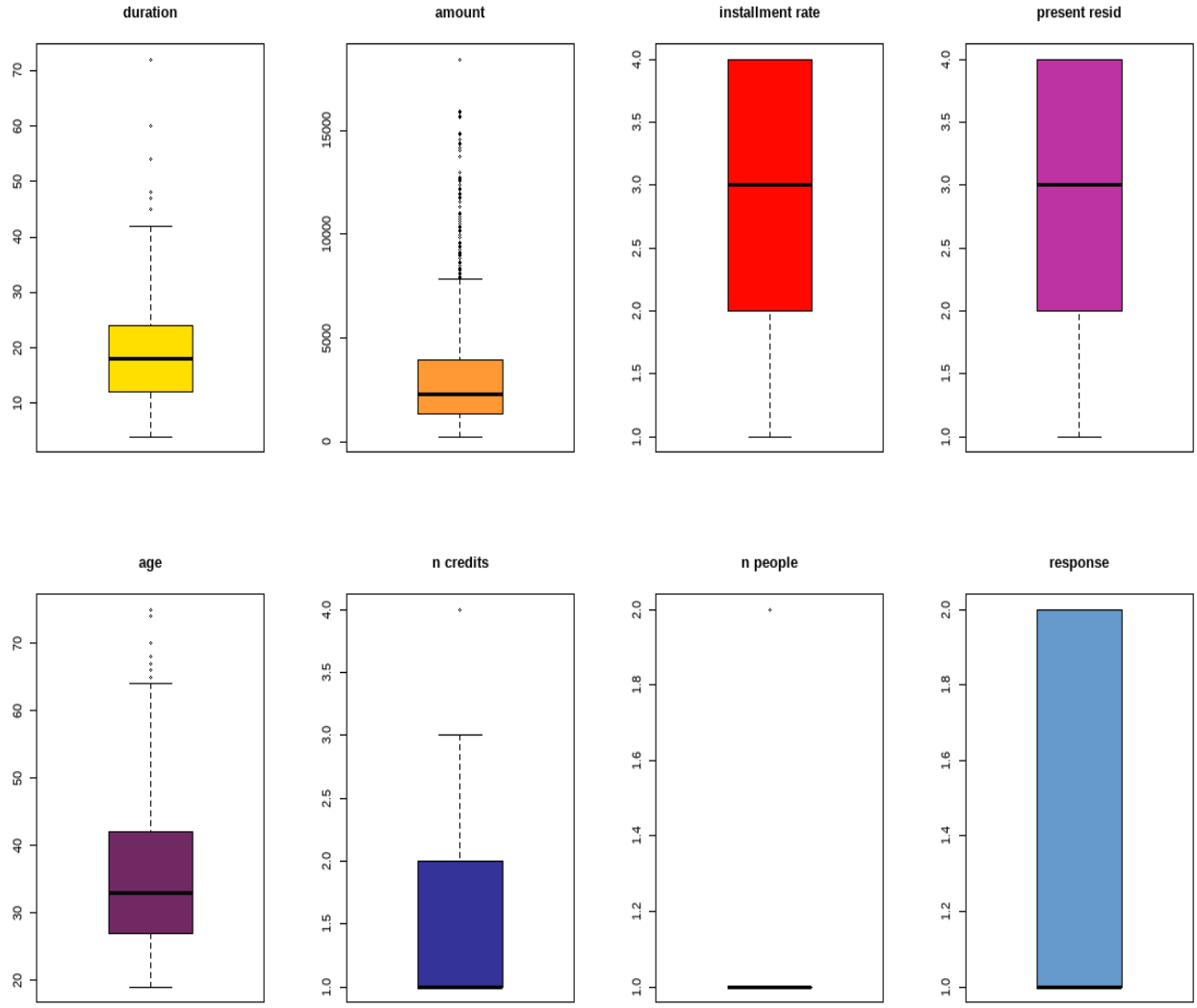
described_variables	n	na	mean	sd	se_mean	IQR	skewness	kurtosis	var
duration	999	0	20.92	12.06	0.38	12.00	1.09	0.92	145.3379
amount	999	0	3273.36	2823.37	89.33	2604.00	1.95	4.29	7971395
installment_rate	999	0	2.97	1.12	0.04	2.00	-0.53	-1.21	1.251719
present_resid	999	0	2.84	1.10	0.03	2.00	-0.27	-1.38	1.218076
age	999	0	35.51	11.34	0.36	15.00	1.02	0.60	128.5386
n_credits	999	0	1.41	0.58	0.02	1.00	1.28	1.61	0.3336663
n_people	999	0	1.16	0.36	0.01	0.00	1.91	1.64	0.1312134
response	999	0	1.30	0.46	0.01	1.00	0.87	-1.24	0.2103306

Tabela 2: Wybrane statystyki opisowe dla zmiennych ilościowych

Na podstawie powyższej tabelki widzimy, że średni okres kredytowania wynosi około 21 miesięcy, wartość kredytu 3200 zaś wiek - 36 lat. Skośność jest to miara symetrii bądź asymetrii rozkładu. Jeśli rozkład jest idealnie symetryczny, wartość skośności wynosi zero. Z kolei jej wartości ujemne wskazują na rozkład lewoskośny (wydłużone jest lewe ramię rozkładu), a dodatni na prawoskośny (wydłużone jest prawe ramię

[3] Więcej informacji na temat biblioteki `https://cran.r-project.org/web/packages/dlookr/index.html`.

rozkładu). Widzimy, że żadna z naszych zmiennych nie ma rozkładu idealnie symetrycznego. Zmienne *installment rate* oraz *present resid* posiadają rozkład lewostronnie skośny a pozostałe - prawostronnie skośny. Dalej w przypadku rozkładu normalnego wartość statystyki kurtozy wynosi zero. Kurtoza dodatnia oznacza, że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. Kurtoza ujemna wskazuje, że w danych istnieje mniej dodatnich wartości odstających niż w przypadku rozkładu normalnego. Widzimy zatem, że żadna z naszych zmiennych nie posiada rozkładu normalnego. Zmienne *installment rate*, *present resid* oraz *response* posiadają ujemny wskaźnik kurtozy, zatem mamy mniej dodatnich wartości odstających niż w rozkładzie normalnym. Pozostałe zmienne posiadają dodatni wskaźnik kurtozy, co jest równoważne, z tym że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. Korzystając w funkcji `boxplot` [4] wyznaczymy teraz wykresy pudełkowe zmiennych z naszych danych wypisanych w tabeli (2).



Rysunek 2: Wykres pudełkowe zmiennych zwartych w tabeli (2) - opracowanie własne

[4]Funkcja `boxplot` pochodzi z biblioteki `graphics`. Więcej informacji można znaleźć pod linkiem `https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/boxplot`

Przyglądając się powyższym wykresom, widzimy, że nasze wcześniejsze wnioski dotyczące skrajnych wartości są prawdziwe.

Analiza korelacji w statystyce polega na zbadaniu czy dwie zmienne są ze sobą istotnie statystycznie powiązane. Innymi słowy, sprawdza, czy jakiekolwiek dwie cechy, atrybuty lub własności współwystępują ze sobą. Korzystając z biblioteki `dlookr` oraz funkcji `correlate` obliczymy macierz korelacji. W tabeli umieszczamy tylko te zmienne, dla których wartość bezwzględna korelacji x jest większa niż 0.1 ($|x| > 0.1$).

variable 1	variable 2	coefficient correlation
amount	duration	0.62
response	duration	0.21
installment_rate	amount	-0.27
response	amount	0.15
age	present_resid	0.26
n_credits	age	0.15
n_people	age	0.12
n_people	n_credits	0.11

Tabela 3: Wyznaczone wartości współczynnika korelacji dla zmiennych

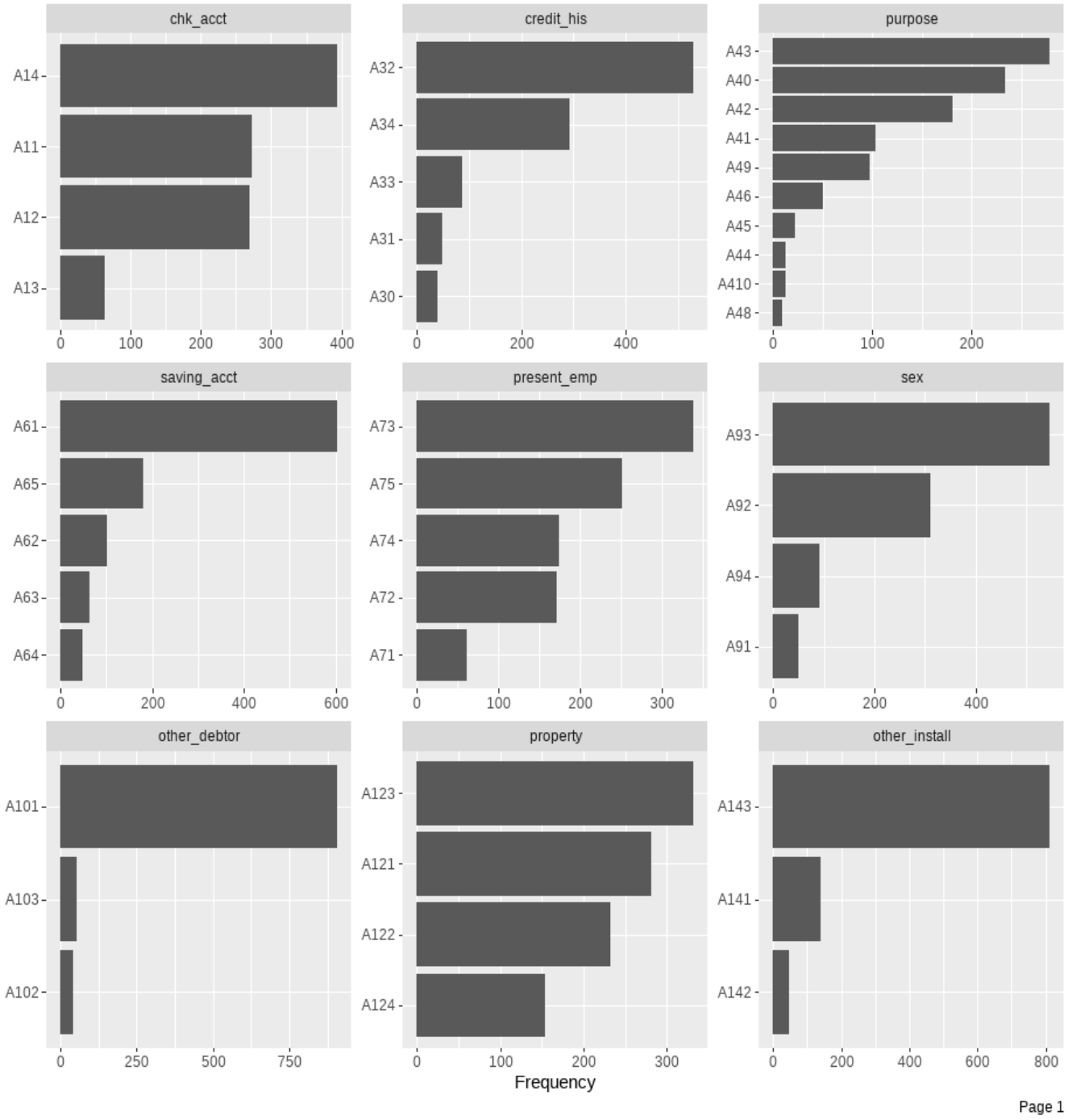
Metodą, którą użyliśmy w wyznaczaniu korelacji, była korelacja Pearsona. Jest to współczynnik określający poziom zależności liniowej między zmiennymi losowymi. Wartość współczynnika korelacji mieści się w przedziale domkniętym $[-1, 1]$. Im większa jest jego wartość bezwzględna, tym silniejsza jest zależność liniowa między zmiennymi. 0 - oznacza brak liniowej zależność, 1 – oznacza zależność dodatnią, a -1 – oznacza zależność ujemną między cechami. Przyglądając się tabeli (3) widzimy, że jednym zestawem zmiennych, przy których moglibyśmy stwierdzić, że istnieje trochę większa korelacja niż normalnie, jest zmienna *amount* oraz *duration*. Wydaje się to dość logiczne, ponieważ czas trwania kredytu w miesiącach faktycznie powinien zależeć od jego wielkości. To znaczy, im mniejszy kredyt, tym szybciej go spłacimy, zaś im większy, tym bardziej czas jego spłacania będzie się wydłużać. Im dłuższy okres kredytowania, tym miesięczna rata będzie niższa, bo zadłużenie jest rozkładane na większą liczbę płatności.

Zmienne jakościowe są to zmienne statystyczne, które wyrażają jakość lub cechę danego obiektu, lub osoby. Zmienne jakościowe zwykle nie odpowiadają liczbom. Funkcja `plot_bar` jest informacyjnym wykresem słupkowym, dzięki któremu możemy dowiedzieć się wielu rzeczy na temat właśnie danych jakościowych.

Niemieckie dane kredytowe posiadają bardzo wiele zmiennych atrybutów, dlatego na poniższych rysunkach (3) oraz (4) widzimy ich skrócone nazwy $A14$ czy $A34$. Wyjaśnione nazwy zmiennych można znaleźć pod linkiem `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`. Natomiast zmienne poddawane analizie będą objaśniane na bieżąco.

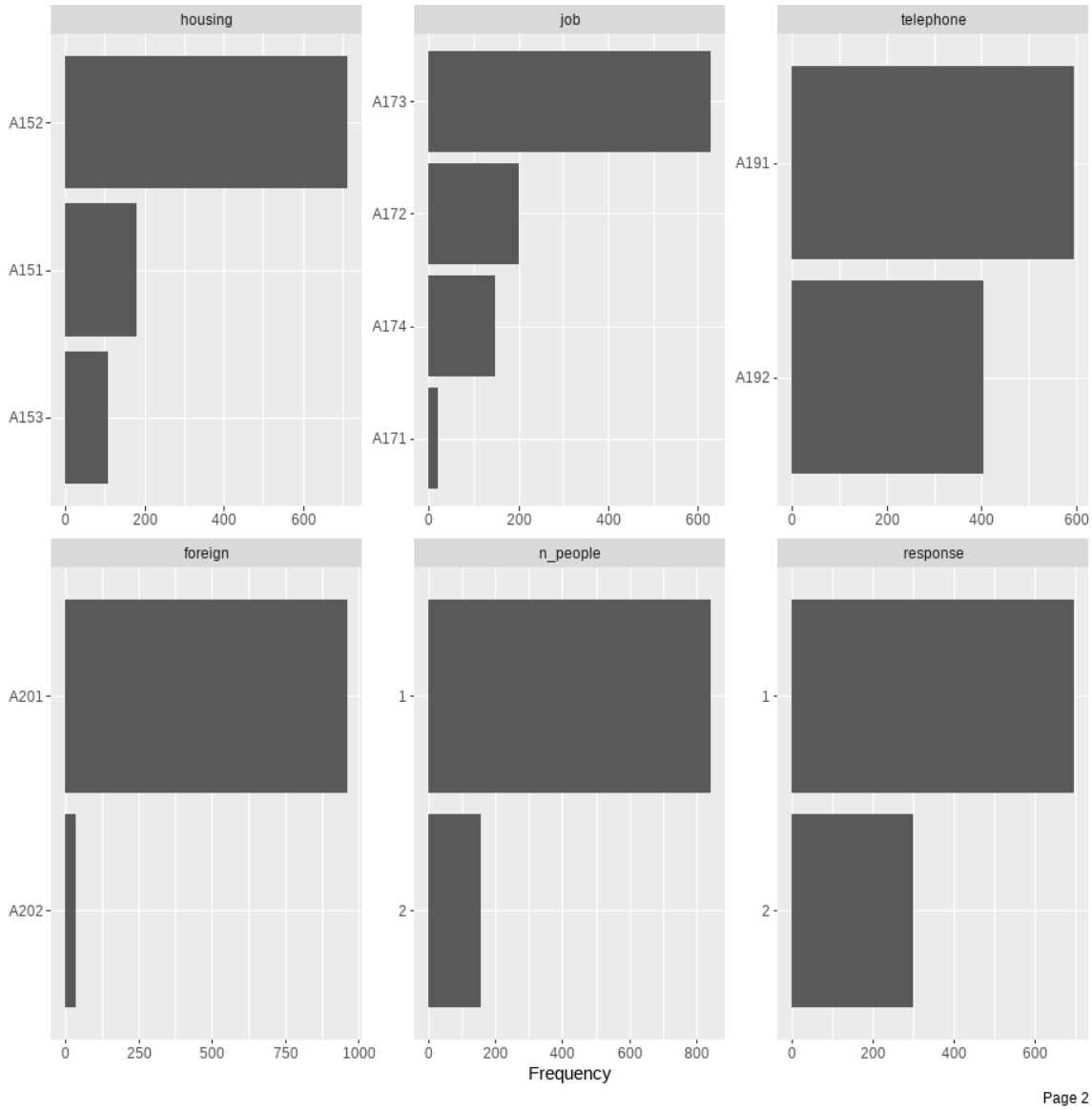
Przejdźmy więc do analizy ważniejszych zmiennych z rysunku (3). Wykres pierwszy *chk_acct* określa status istniejącego konta czekowego. Widzimy, że zdecydowanie dominuje tutaj zmienna $A14$ oznaczająca brak konta czekowego. Dalej w *credit_hist* dotyczącej historii kredytowania widzimy, że większość osób ubiegających się ponownie o kredyt dotychczasowo kredyty spłacała należycie. W późniejszym etapie analizy sprawdzimy dodatkowo, czy poprzednie kredyty wpływają na bycie dobrym bądź złym klientem. Celem, na który klienci najczęściej przeznaczaliby pieniądze z kredytu, jest radio bądź telewizja, zaś

najrzadziej - przekwalifikowanie. Może być spowodowane tym, że większość ludzi boi się zmian, nie chce ryzykować stracenia źródła dochodu w przypadku zmiany w pracy i poszerzenia swoich kwalifikacji. Kolejną zmienną jest posiadanie konta oszczędnościowego. Zadziwiające jest to, że największy odsetek klientów posiada mało pieniędzy na swoim koncie oszczędnościowym (atrybut $A61$ - < 100 Deutschmark) bądź wcale (atrybut $A65$).



Rysunek 3: Wykresy słupkowe zmiennych jakościowych - opracowanie własne

Według redaktora Mikołaja Paseckiego [5] średnia wartość oszczędności wśród osób odkładających pieniądze to 36 tysięcy złotych. Co ciekawe, blisko 70 % ankietowanych stwierdza, że w obecnych uwarunkowaniach gospodarczych odkładanie pieniędzy jest trudne. Większość nieoszczędzających respondentów twierdzi, że nie jest w stanie odłożyć pieniędzy przez zbyt wysokie wydatki bieżące. Z kolei 20% badanych uważa, iż odkładnie pieniędzy wiązałoby się ze zbyt dużymi wyrzeczeniami. Przejdźmy dalej do analizy naszych zmiennych - w przypadku zmiennej *present_emp* najczęstszą opcją jest zatrudnienie między 1 rokiem a 4 latami. Jest to dość typowe, ponieważ to średnio wtedy kończy się umowa na okres próbny, dostajemy stałe zatrudnienie i faktycznie możemy zaczynać ubiegać się o kredyt. Kolejną istotną kwestią w przypadku ubiegania się o kredyt może być fakt, czy posiadamy już nieruchomość na własność, czy też nie. Najczęściej klienci banku odpowiadali, że posiadają jedynie samochód, zaś dopiero drugą opcją było własne mieszkanie.



Rysunek 4: Wykresy słupkowe zmiennych jakościowych - opracowanie własne

[5]https://www.bankier.pl/smart/oszczednosci-polakow-w-dobie-inflacji-ile-oszczednosci-maja-polacy

Powyższy rysunek (4) jest kontynuacją wykresów słupkowych zmiennych jakościowych. W przypadku zmiennej *housing* widzimy znaczne dysproporcje między atrybutem $A152$ oznaczającym, że klienci posiadają swoje własne mieszkanie, a atrybutami $A151$, określającym wynajmowane mieszkanie i $A152$ oznaczającym mieszkanie za darmo. Zmienna *job* posiada również znaczące dysproporjce - najwięcej jednak naszych klientów jest z atrybutem $A173$ oznaczającym wykwalifikowanego pracownika bądź urzędnika. Zadziwiającą jest jednak dość kolejna kategoria, $foreign\ worker$, a właściwie jest wyniki. Pokazują one bowiem, że prawie cała nasza grupa analizowanych klientów pochodzi zza granicy.

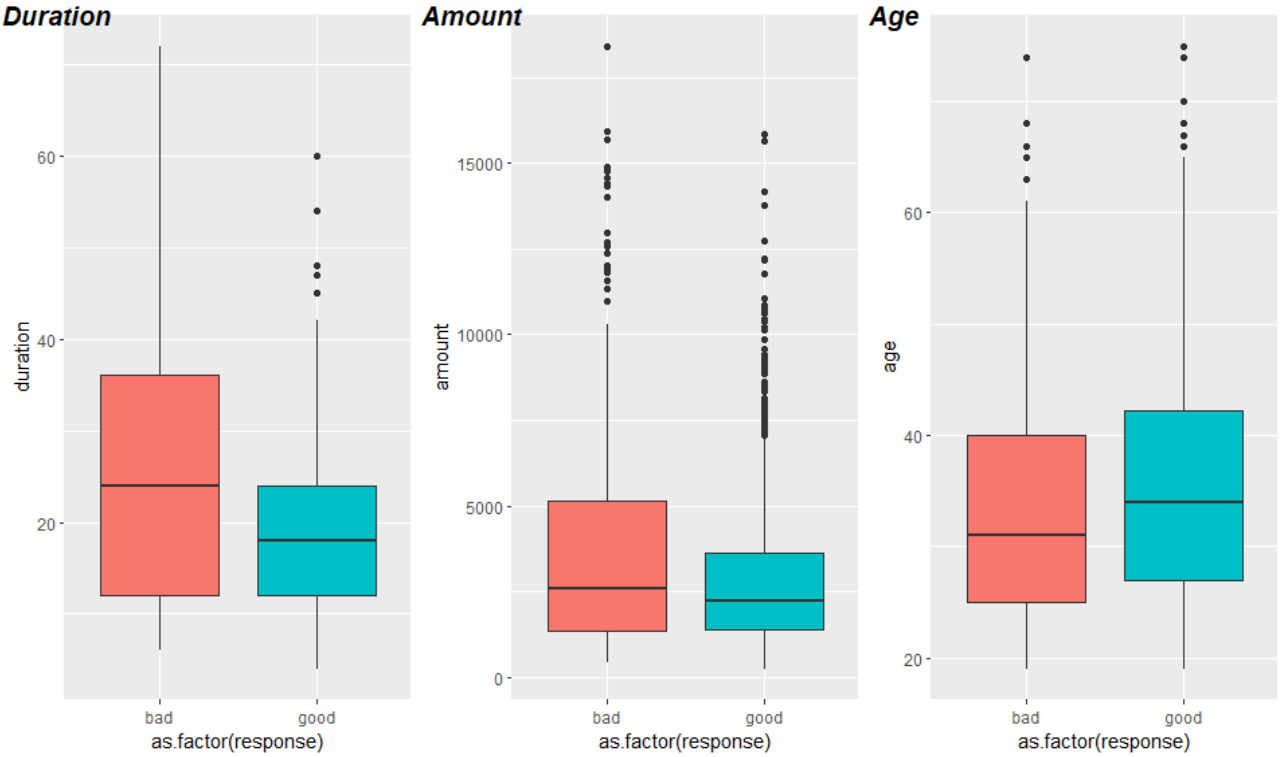
Niemal w każdym banku sprawdzenie kompletu dokumentacji do kredytu podzielone jest na cztery główne działy.

- Analiza osobista - polega na sprawdzeniu gospodarstwa domowego pod kątem liczby osób, obciążeń kredytowych, wieku czy sytuacji prawnej
- Analiza ekonomiczna - określa, czy dochody są stabilne oraz, czy dają możliwość prawidłowego regulowania przyszłego zobowiązania kredytowego
- Analiza prawna - bank kontroluje wszystkie dokumenty, sytuację prawną i inne kwestie mogące przeszkodzić w realizacji transakcji. Jeśli analityk kredytowy stwierdzi, że bank będzie narażony na ryzyko, decyzja odmowna dotycząca udzielenia kredytu będzie szybka i prosta
- Analiza nieruchomości - dotyczy zweryfikowania wartości rynkowej zabezpieczenia poprzez wycenę dostarczoną przez wnioskodawcę kredytu

Poniższy rysunek (5) zawiera trzy wykresy pudełkowe zmiennych *duration* (oznaczającą, ile miesięcy zajmie spłata kredytu), *amount* (wskazującą na kwotę kredytu) oraz *age*. Każdy z wykresów podzielony jest na dwie części względem zmiennej *response* oznaczającej zdolność kredytową, która może być *bad* (zła) lub *good* (dobra).

Zwróćmy uwagę na wykres *duration*. W obydwu przypadkach zdolności kredytowej, czarna kreska oznaczająca medianę wykresów, jest symetryczna. Oznacza to, że wykres cech w tym przypadku jest symetryczny. Ponadto, w przypadku złej zdolności kredytowej, mamy do czynienia z dłuższym wykresem, co oznacza, że dane są bardziej rozproszone, to znaczy, mogą przyjmować bardziej różniące się wartości. Co więcej, dostrzegamy bardzo długie wąsy, co sugeruje, że mamy dużo obserwacji skrajnych. W przypadku dobrej zdolności kredytowej tego samego wykresu zauważamy, że posiada on bardzo odległe wartości odstające. Może to odzwierciedlać rzeczywisty rozkład lub być rezultatem przypadku, ale może także świadczyć o błędnym pomiarze czy pomyłkach we wprowadzaniu informacji do bazy danych.

Przejdźmy do wykresu pudełkowego *amount* poniższego rysunku (5). Dokładnie widać, że mediana nie znajduje się na "środku" pudełka, zatem wykres nie jest symetryczny. Znacząco większa jest odległość górnego krańca pudełka od mediany, zatem mamy asymetrię prawostronną. Oznacza to, że obserwacje statystyczne skupiają się przy wartościach cechy mniejszych od średniej arytmetycznej. Co więcej, w przypadku i dobrej i złej zdolności kredytowej, spotykamy się z bardzo dużą ilością obserwacji skrajnych.



Rysunek 5: Wykres pudełkowy zmiennych *duration*, *amount* oraz *age* danych - opracowanie własne

Przyglądając się wykresowi *age* rysunku (5) dostrzegamy pewną niesymetryczność w obydwu przypadkach, ale w celu jej dokładnego określenia, korzystamy z funkcji `summary`. Tabela (4) zawiera podsumowanie zmiennej wiek w przypadku dobrej oraz złej zdolności kredytowej. Dostrzegamy, że wszystkie wartości (minimalne, maksymalne itd.) są ze sobą bardzo zbliżone. Ponownie w przypadku złej zdolności kredytowej spotykamy się z prawostronną asymetrią.

	age - good	age - bad
Min.	19.00	19.00
1st Qu.	32.00	30.50
Median	45.00	42.00
Mean	45.17	42.55
3rd Qu.	58.00	53.50
Max.	75.00	74.00

Tabela 4: Tabela podsumowująca zmienną wiek w przypadku dobrej oraz złej zdolności kredytowej danych

Kategoria wiekowa	liczba osób
18 - 30	411
30 - 50	476
50 - 80	113

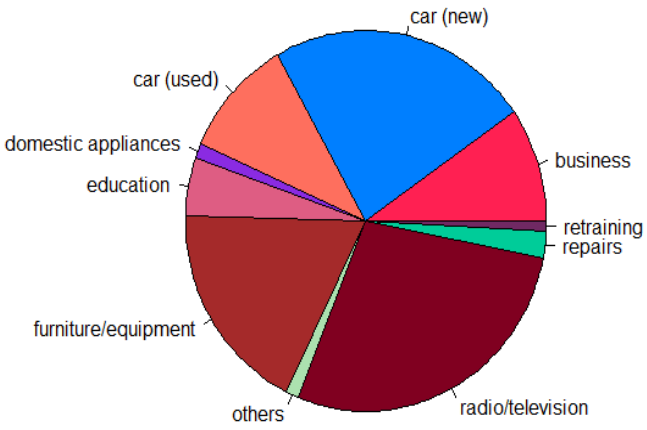
Tabela 5: Kategoryzacja zmiennej wiek

Powyższa tabela (5) zawiera kategoryzację zmiennej wiek na trzy grupy - osoby w wieku między 18 a 30, między 30 a 50 oraz powyżej 50 roku życia aż do 80. Zauważamy, że najwięcej kredytów biorą osoby w drugiej kategorii wiekowej. Wydaje się to dość standardowe, ponieważ jest to wiek, kiedy ludzie najczęściej chcą się ustatkować i biorą kredyty na mieszkania. Według Macieja Kazimierskiego [6], analityka finansowego, znaczna większość kredytobiorców w momencie zaciągnięcia zobowiązania była w wieku od 25 do 44 lat. Sprawdzimy, czy w przypadku zmiany kategoryzacji zmiennej wiek, również uzyskamy podobne wnioski.

Kategoria wiekowa	liczba osób
18 - 25	190
25 - 44	609
44 - 80	201

Tabela 6: Kategoryzacja zmiennej wiek

Widzimy faktycznie, że ze wszystkich 1000 danych prawie 61% jest osobami w wieku od 25 do 44 lat. Z oczywistych względów osoba, która dopiero co ukończyła 18 rok życia nie uzyska zbyt wielu punktów i jej ocena scoringowa może być niewystarczająca do uzyskania kredytu. W praktyce więc banki w ten sposób uniemożliwiają osobom poniżej 21 roku życia uzyskanie kredytu mieszkaniowego. Pokrywa się to z naszą tabelą (6), gdzie najmniej kredytów otrzymaliśmy w zadanej grupie wiekowej. Znacznie łatwiej mają osoby, które są w wieku 25 - 44 lat. Ta grupa kredytobiorców charakteryzuje się już względną stabilnością zawodową i rodzinną, co stanowi cechy bardzo pożądane przez banki. Te osoby mogą łatwiej uzyskać kredyt niż osoby dopiero rozpoczynające swoje dorosłe życie. Dlatego w tej kategorii wiekowej posiadamy najwięcej osób.



Rysunek 6: Wykres kołowy zmiennej *purpose* danych - opracowanie własne

[6] Źródło `https://direct.money.pl/artykuly/porady/kto-w-polsce-bierze-kredyt-na-mieszkanie,247,0,2353655`

purpose	number
business	97
car (new)	234
car (used)	103
domestic appliances	12
education	50
furniture/equipment	181
others	12
radio/television	280
repairs	22
retraining	9

Tabela 7: Tabela zawierająca podsumowanie zmiennej *purpose* danych

Powyższy rysunek (6) oraz tabela (7) określają nam, na co najczęściej ludzie chcieli wziąć kredyt w banku. Dostrzegamy, że najchętniej były to trzy rzeczy - radio/telewizor, nowy samochód bądź meble/wyposażenie, zapewne do nowego mieszkania.

purpose	number	response good	response bad
car (new)	234	145	89
education	50	28	22
furniture/equipment	181	123	58
radio/television	280	218	62

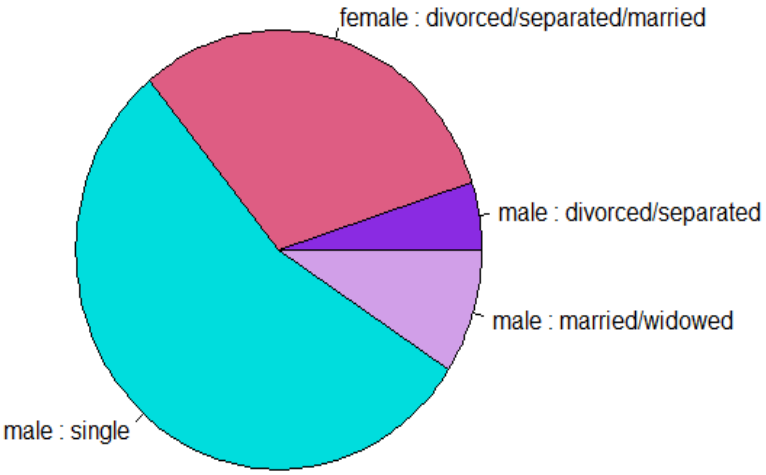
Tabela 8: Tabela zawierająca podsumowanie zmiennej *purpose* w kategorii dobrej bądź złej zdolności kredytowej danych

Tabela (8) zawiera podsumowanie zmiennej *purpose* w kategorii dobrej bądź złej zdolności kredytowej. Widzimy, że w każdym przypadku, więcej było klientów dobrych, czyli takich, którzy powinni spłacić zadłużenie. Możemy jednak dostrzec pewną zależność, że przy zmiennej celu *education* proporcje między dobrymi a złymi klientami nie różnią się znacząco.

Dalej przejdziemy do analizy statusu społecznego.

Personal status and sex	number
male : divorced/separated	50
female : divorced/separated/married	310
male : single	548
male : married/widowed	92

Tabela 9: Tabela zawierająca zmienną *Personal status and sex* danych

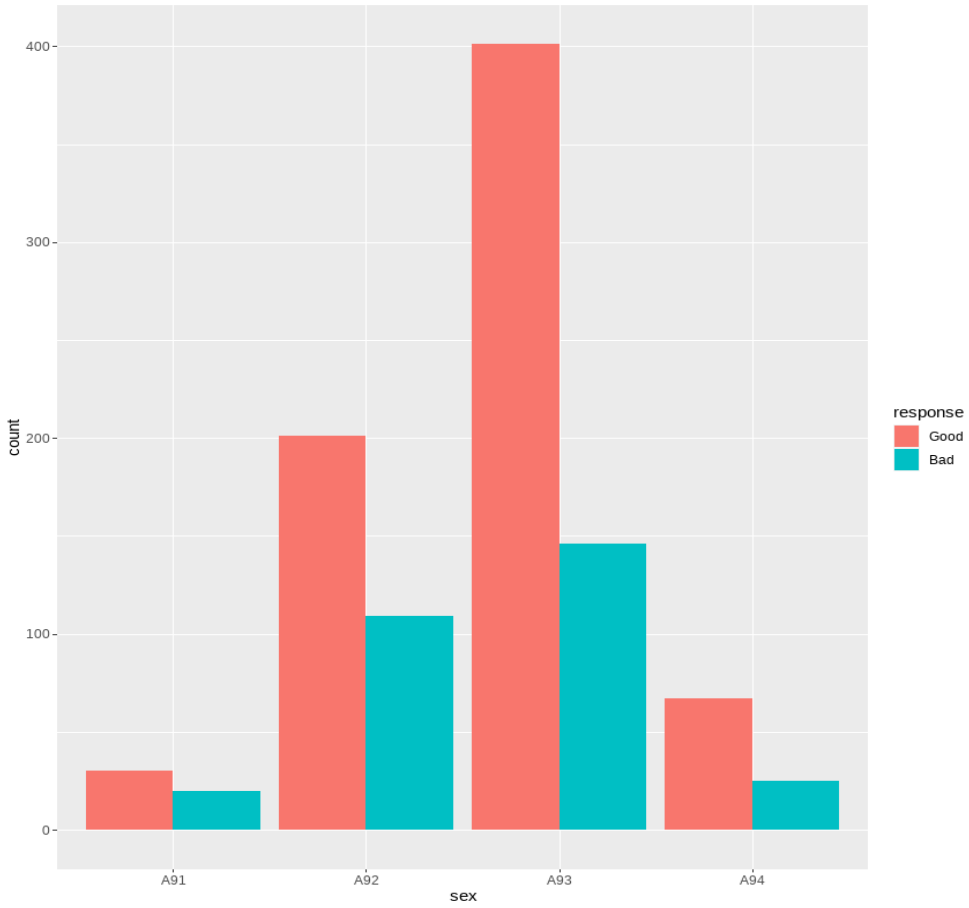


Rysunek 7: Wykres kołowy zmiennej *Personal status and sex* danych - opracowanie własne

Na podstawie powyższego rysunku (7) oraz tabeli (9) widzimy, że zdecydowaną większością kredytobiorców są samotni mężczyźni $(54,8\%)$. Równocześnie, najmniejszą liczbą kredytobiorców są również mężczyźni, ale rozwiedzeni bądź w separacji (5%). Sprawdzimy teraz, czy status społeczny oraz płeć miały wpływ na to, czy byliśmy dobrym, czy złym kredytobiorcą.

Rysunek (8) zawiera podsumowanie zmiennej płci oraz statusu społecznego w kategorii dobrego bądź złego klienta. Widzimy dokładnie, że pewien atrybut $A93$ zdecydowanie przewyższa pozostałe. Jest to mężczyzna o statusie singla. Jak widzimy, zdecydowana większość takich mężczyzn otrzymała pozytywną odpowiedź na starania o kredyt. Drugi atrybut, gdzie otrzymaliśmy najwięcej zapytań o kredyt jest $A92$ - kobieta : rozwiedziona/w separacji/zamężna. Według Seana LaPointe'a [7] mężczyźni są zdecydowanie bardziej skłonni do zaciągania pożyczek niż kobiety. Według niego wynika to z tego, że posiadają oni na ogół wyższe oraz bardziej pewne dochody niż kobiety. W rezultacie mężczyźni mogą czuć się bardziej komfortowo, biorąc pożyczkę, wiedząc, że będą w stanie ją spłacić. Ponadto badania wykazały, że mężczyźni mają zazwyczaj wyższą ocenę kredytową niż kobiety, co daje im większy dostęp do tańszych pożyczek.

[7] Artykuł w języku angielskim można znaleźć pod linkiem `https://www.fool.co.uk/2021/06/15/why-are-men-more-likely-to-take-out-loans-than-women/`



Rysunek 8: Wykres zmiennej *Personal status and sex* w kategorii dobrego bądź złego klienta - opracowanie własne

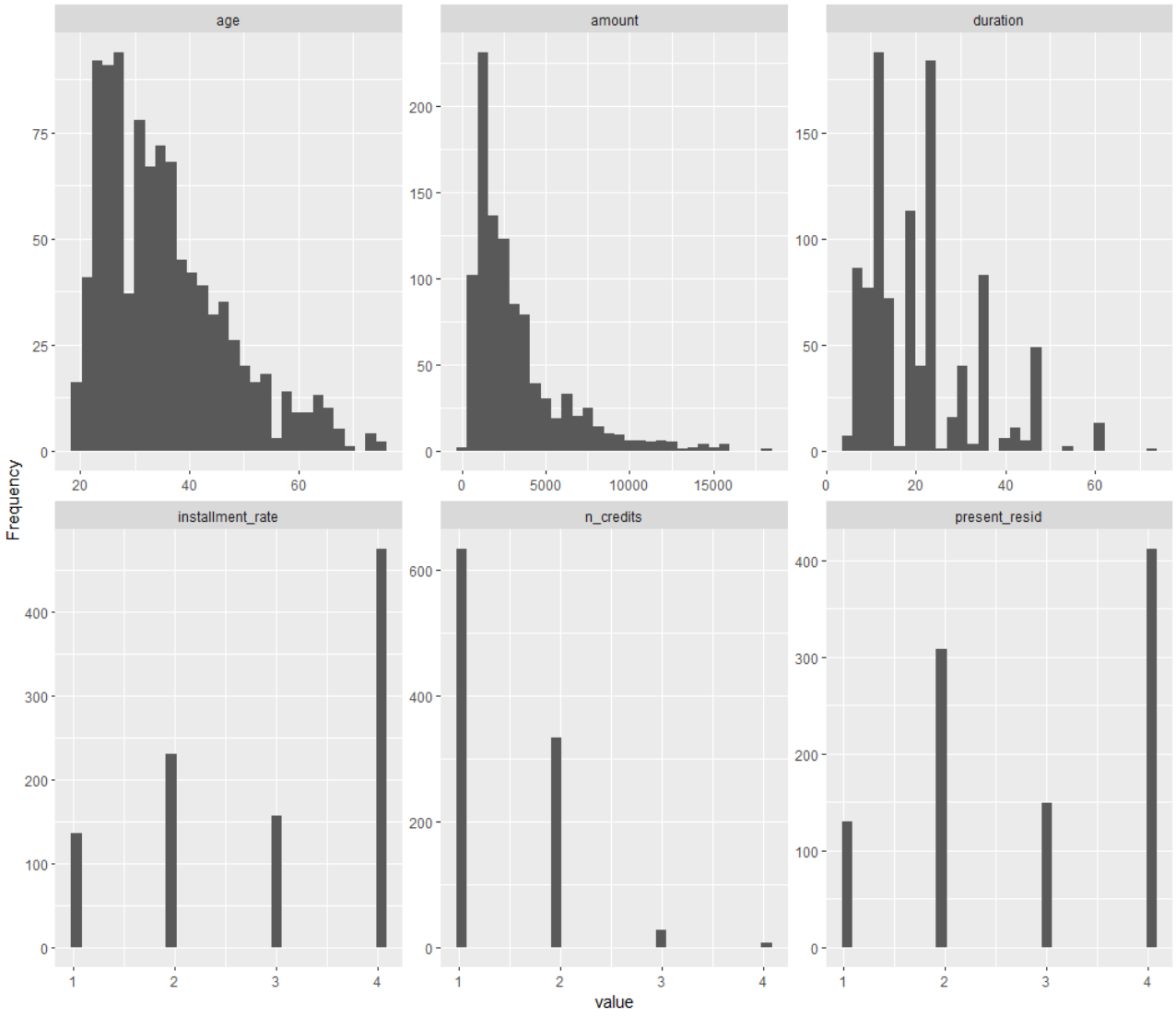
Sprawdzimy teraz, czy wcześniejsze kredyty, zadłużenia oraz to, czy były spłacane terminowo, wpływa na to, czy jesteśmy dobrym, czy złym klientem.

Historia kredytowa	Dobry klient	Zły klient
brak kredytów/wszystkie kredyty należycie spłacone	15	25
wszystkie kredyty w tym banku zostały należycie spłacone	21	28
dotychczasowe kredyty należycie spłacane	361	169
opóźnienie w spłacie w przeszłości	60	28
konto krytyczne/inne istniejące kredyty (nie w tym banku)	242	50

Tabela 10: Tabela przedstawiająca historię kredytową i porównanie dobrych i złych klientów

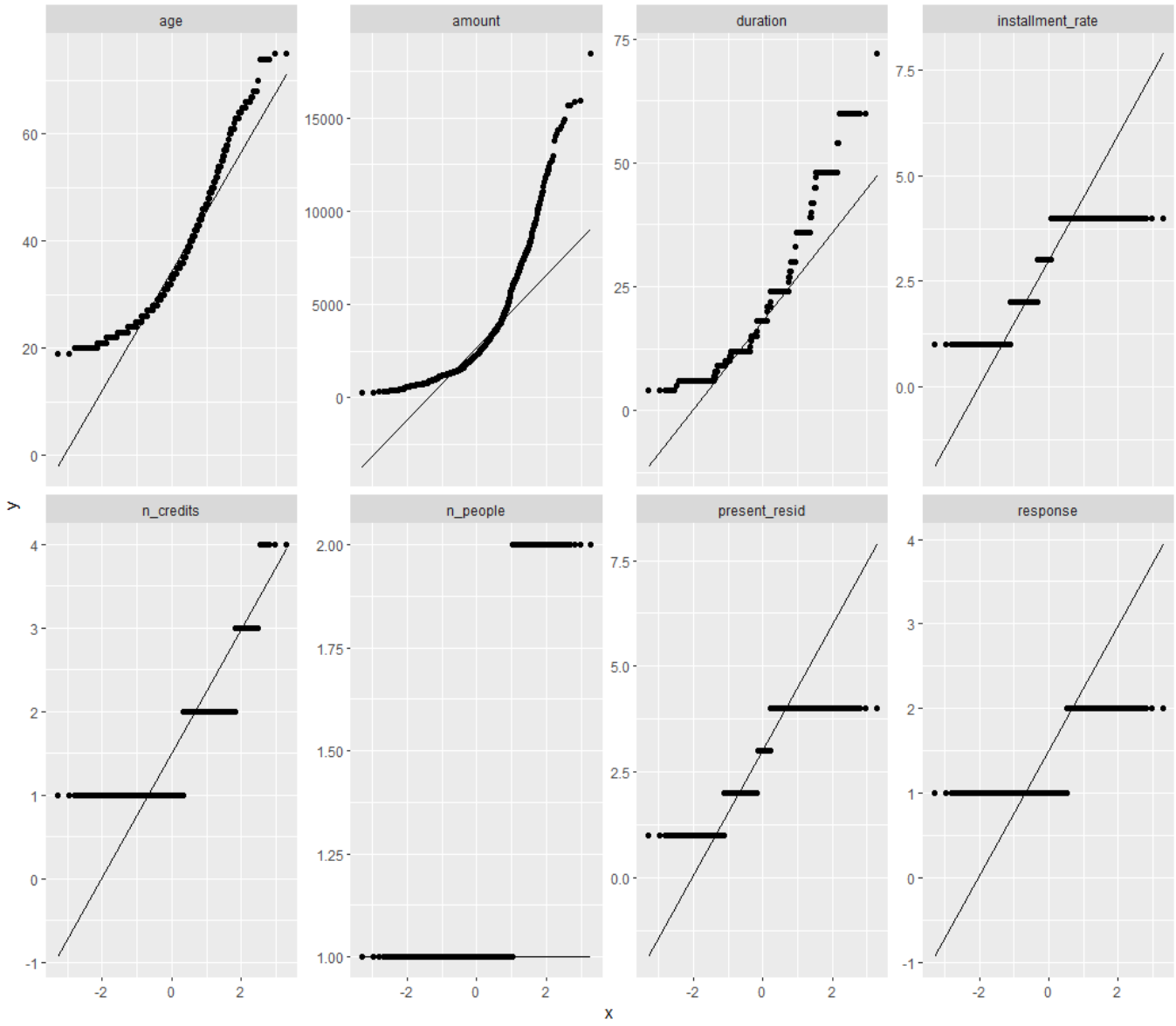
Tabela (10) przedstawia podsumowanie historii kredytowej oraz dobrych i złych klientów. Na jej podstawie widzimy, że fakt, czy mieliśmy opóźnienie w spłacie w przeszłości, czy posiadamy inne kredyty, nie ma wpływu na to, czy dostaniemy kredyt, czy nie. Zdecydowana większość kredytobiorców okazała się dobrym klientem i wniosek o kredyt został rozpatrzony pozytywnie.

Biblioteka `DataExplorer` ma bardzo wiele ciekawych funkcji. Poniżej prezentujemy histogramy oraz wykresy kwantylowe dla cech ilościowych naszych zmiennych. Zostały one stworzone kolejno na podstawie funkcji `plot_histogram` oraz `plot_qq`.



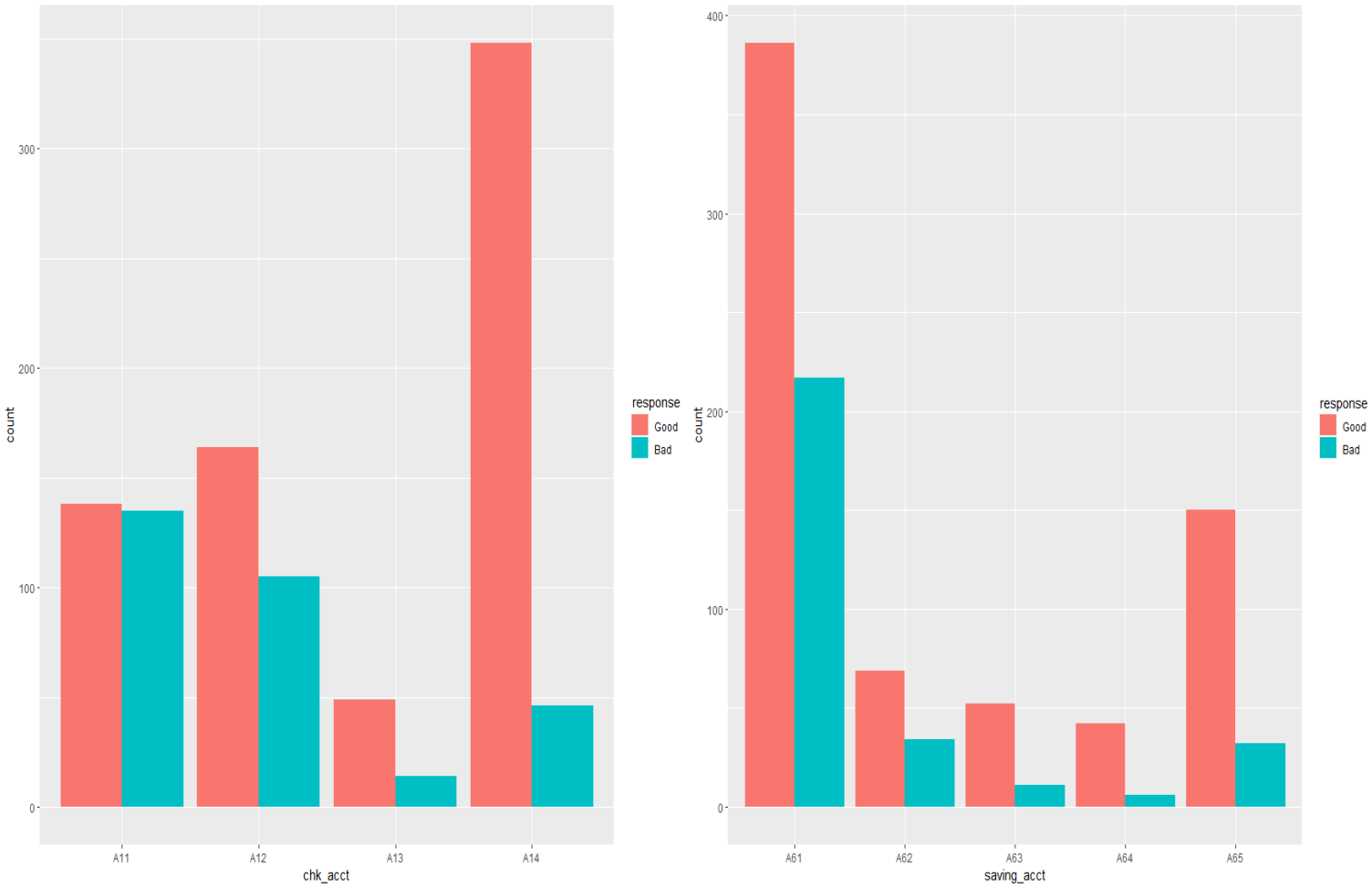
Rysunek 9: Histogramy zmiennych ilościowych - opracowanie własne

Histogram to zestawienie danych statystycznych w postaci wykresu powierzchniowego złożonego z przylegających do siebie prostokątów, których wysokość ilustruje liczebność występowania badanej cechy w populacji lub jej próbie, a podstawy (które spoczywają na osi odciętych) są rozpiętościami przedziałów klasowych. Analizując histogram *age* oraz *amount* powyższego rysunku (9) widzimy, że wskazują one na asymetrię prawostronną. Potwierdzają się ponownie nasze wcześniejsze wnioski. Na pozostałych wykresach właściwie nie widać żadnej konkretnej zależności, która mogłaby nas interesować w dalszych analizach.



Rysunek 10: Wykresy kwantylowe zmiennych ilościowych - opracowanie własne

Wykres kwantylowy to układ kropek, które w idealnej sytuacji powinny tworzyć prostą. To proste narzędzie świetnie nadaje się do wizualnej oceny tego, czy rozkład zmiennej jest podobny do dowolnego rozkładu teoretycznego. Przyglądając się powyższemu rysunkowi (10), widzimy, że żaden układ kropek z naszych zmiennych nie pokrywa się z wyznaczoną prostą linią.



Rysunek 11: Wykres zmiennych *chk_acct* oraz *saving_acct* w kategorii dobrego bądź złego klienta - opracowanie własne

Powyższy rysunek (11) podsumowuje zmienne *chk_acct* oraz *saving_acct* i sprawdza, w której grupie najwięcej było dobrych klientów, a w której najwięcej złych.
W przypadku statusu istniejącego już konta czekowego widzimy, że atrybut $A11$, oznaczający wartość $< 0DM$, ma praktycznie tyle samo dobrych klientów, co złych. Najwięcej dobrych klientów otrzymaliśmy w przypadku atrybutu $A14$ oznaczającego całkowity brak konta czekowego.
W przypadku posiadania konta oszczędnościowego nie występuje aż tak ogromna dysproporcja jak poprzednio, jednakże widzimy, że zdecydowana większość naszych dobrych klientów występuje przy atrybucie $A61$ oznaczającego posiadania konta oszczędnościowego $< 100DM$. Ponadto można zauważyć, że w przypadku $A65$ określającego całkowity brak konta oszczędnościowego zdecydowana większość klientów została uznawana za dobrych.

Przygotujemy nasze dane do dalszej analizy. Wiemy już, że nie posiadamy żadnych wartości brakujących bądź inaczej oznaczonych. Nasza struktura danych wydaje się spójna. Zauważyliśmy już również, że żadne ze zmiennych nie są bardzo mocno ze sobą skorelowane. Dodatkowo należy jeszcze usunąć z naszego zbioru cechy, o zbyt małej zmienności, czyli wariancji. Na podstawie tabeli (2) i analizie jej wariancji, do dalszej analizy usuniemy zmienne $installment\_rate$, $present\_resid$, $n\_credits$, n_people. Zmienna $response$ również cechuje się małą zmiennością, jednakże jest to nasza zmienna wynikowa, której nie możemy usunąć.

1.2 Klasyfikacja wraz z oceną dokładności

Przed przystąpieniem do klasyfikacji, przypomnijmy, że zmienną wynikową jest zmienna *response* o wartości 1 gdy mamy do czynienia z dobrym klientem oraz wartości 2 w przypadku złego klienta. Nasz zbiór danych wymaga użycia macierzy następujących kosztów

	1	2
1	0	1
2	5	0

Tabela 11: Macierz kosztów oryginalna

Warto zaznaczyć, że w naszych danych zmienimy trochę wartości. To znaczy, 1 będzie oznaczać cały czas dobrego klienta, jednak 2 zamieniany na 0, co będzie oznaczało złego klienta. W takim wypadku macierz kosztów będzie wyglądała następująco

	0	1
0	0	5
1	1	0

Tabela 12: Macierz kosztów po zamianie

Wiersze powyższej macierzy przedstawiają rzeczywistą klasyfikację, zaś kolumny - przewidywaną (dokładnie zostanie wyjaśnione to później). Widzimy, że zdecydowanie gorzej jest sklasyfikować klienta złego jako dobrego, niż sklasyfikować klienta dobrego jako złego. Wydaje się to dość logiczne. W przypadku, gdy klienta dobrego klasyfikujemy jako złego, nie udzielając mu tym samym pożyczki, tracimy ewentualną prowizję. W momencie, gdy udzielimy kredytu złemu klientowi, oceniając go jako dobrego, jako bank jesteśmy narażeni, że zadłużenie nie zostanie spłacone. Wiąże się to z większymi kosztami, większą stratą dla banku.

W przypadku klasyfikacji interesuje nas problem prognozowania (predykcji) etykiety G na podstawie wektora cech $X = (X_1, \ldots, X_p)'$. Zadanie klasyfikacji polega na skonstruowaniu reguły decyzyjnej $d(x)$, która będzie dla dowolnej obserwacji $x \in X$ przypisywała przynależność do jednej z klas ze zbioru G.

Przy doborze algorytmów predykcyjnych zastosowaliśmy kilka kryteriów. Pierwszym z nich było to, aby algorytm był odpowiedni do analizy problemu klasyfikacji binarnej, z uwagi na charakter naszego zbioru danych. Kolejnym kryterium było uwzględnienie popularności i powszechności stosowania algorytmu w praktyce oraz dostępności implementacji w popularnych narzędziach do analizy danych. Wzięliśmy także pod uwagę czas potrzebny na przetworzenie danych przez algorytm, a także jego zdolność do obsługi zbiorów danych o dużej liczbie cech i obserwacji. Ostatecznie wybraliśmy do analizy cztery algorytmy: regresję logistyczną, kwadratowa anlizę dyskryminacyjna, metodę k - najbliższych sąsiadów oraz sieć neuronową.

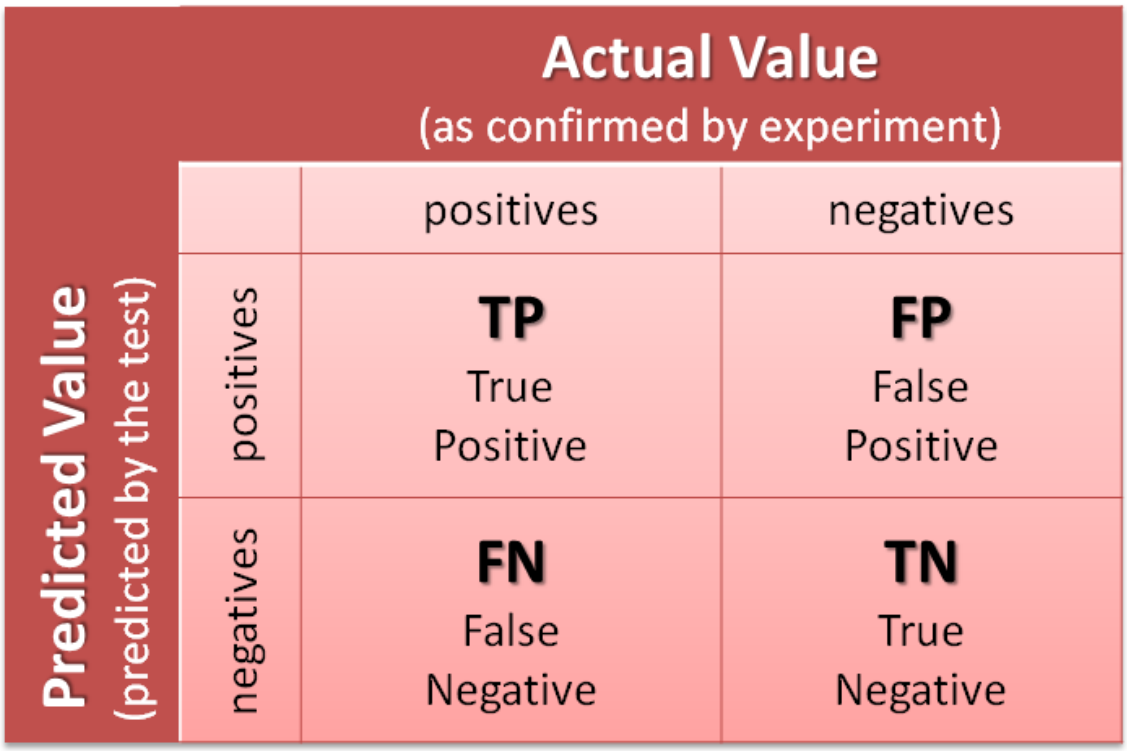
Możemy się domyślać, że różne modele będą dawały różne rezultaty. Dlatego wybierzemy kilka modeli, które następnie będziemy analizować i oceniać ich dokładność.

- (M1) Model w oparciu o wszystkie zmienne
- (M2) Model w oparciu o Backward AIC $\rightarrow$ $response \sim chk\_acct + duration + credit\_his + purpose + saving\_acct + present\_emp + other\_debtor + age + other\_install + housing + foreign$

- (M3) Model w oparciu o Backward BIC $\rightarrow$ $response \sim chk\_acct + duration + credit\_his$
- (M4) Model w oparciu o bibliotekę `klaR` [8] oraz metodę $lda \rightarrow response \sim duration$
- (M5) Model w oparciu o bibliotekę `klaR` oraz metodę $qda \rightarrow response \sim duration + amount$

Przy ocenie klasyfikacji pomogą nam macierze błędu a także niektóre wskaźniki - dokładność, czułość, fałszywie pozytywna wartość, swoistość, fałszywie negatywna wartość oraz współczynnik kolreacji Matthewsa.

Macierz błędu powstaje z przecięcia klasy prognozowanej i klasy faktycznie zaobserwowanej, mamy zatem 4 przypadki (2 dla zgodności i 2 dla niezgodności prognozy ze stanem faktycznym). Używane są do wydajności modelu klasyfikacji, gdzie N oznacza liczbę klas docelowych. Macierz porównuje rzeczywiste wartości docelowe z przewidywanymi przed model. Daje nam to całościowy obraz tego, jak dobrze działa nasz model klasyfikacji i jakie rodzaje błędów popełnia. W przypadku problemu klasyfikacji binarnej otrzymujemy macierz 2×2, która wygląda następująco:



Rysunek 12: Przykładowa, teoretyczna macierz błędów

Kolumny macierzy reprezentują rzeczywiste wartości zmiennej docelowej, zaś wiersze - przewidywane wartości. Wyjaśnimy teraz kolejno pojęcia TP, FP, TN, FN na przykładzie testów medycznych.

- TP ludzie chorzy poprawnie zdiagnozowani jako chorzy
- FP ludzie zdrowi błędnie zdiagnozowani jako chorzy

[8] Biblioteka posiada wiele różnych funkcji służących do klasyfikacji oraz wizualizacji danych. Więcej informacji można znaleźć pod linkiem `https://www.rdocumentation.org/packages/klaR/versions/1.7-2`

- TN ludzie zdrowi poprawnie zdiagnozowani jako ludzie zdrowi
- FN ludzie chorzy błędnie zdiagnozowani jako zdrowi

Z macierzy pomyłek można wyliczyć wiele wskaźników dla klasyfikatora binarnego, takich jak:

- Dokładność (*Accuracy*) pozwala nam ocenić jakość klasyfikacji testu. Daje nam informacje na temat tego, jaka część testów, ze wszystkich zaklasyfikowanych, została oceniona poprawnie. Im wyższa wartość dokładności, tym lepiej. $ACC = 1$ oznacza idealnie dopasowanie i brak pomyłki ani razu.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- Prawdziwie pozytywna wartość (*True Positive Rate*) zwana inaczej czułością (*sensivity*). Mówi nam o tym, jaki jest udział prawidłowo prognozowanych przypadków pozytywnych wśród wszystkich przypadków pozytywnych. Wartość ta powinna być jak najbliższa 1.

$$TPR = \frac{TP}{TP + FN}$$

- Fałszywie pozytywna wartość (*False Positive Rate*) określa jaki jest udział fałszywie pozytywnych przypadków wśród wszystkich negatywnych przypadków. Wartość jego powinna być jak najbliższa 0.

$$FPR = \frac{FP}{TN + FP}$$

- Prawdziwie negatywna wartość (*True Negative Rate*) inaczej swoistość (*specifity*) mierzy, jak dużo ze wszystkich negatywnych przypadków zostało rzeczywiście zaklasyfikowanych do tej kategorii.

$$TNR = \frac{TN}{TN + FP}$$

- Współczynnik korelacji Matthews (*ang. Matthews Correlation Coefficient*) przyjmujący wartości od -1 do 1. Dla $MCC = 1$ nasz model bardzo dobrze, wręcz idealnie klasyfikuje wszystko do prawidłowej kategorii, zaś dla $MCC = -1$ otrzymujemy informację, że wszystko zostało zaliczone do niepoprawnej kategorii.

$$MCC = \frac{TN \cdot TP - FP \cdot FN}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}}$$

1.2.1 Liniowa analiza dyskryminacyjna (LDA)

Reguła klasyfikacyjna dla LDA polega na tym, że klasyfikujemy obserwację do klasy, do której ma najbliżej w przestrzeni zmiennych wyznaczonych przez wektory dyskryminacyjne. Innymi słowy, dla nowej obserwacji obliczamy jej odległość od każdego z wektorów dyskryminacyjnych, a następnie przypisujemy ją do klasy, której wektor jest najbliższy.

Poniżej przedstawimy w krokach, jak będzie wyglądać nasza liniowa analiza dyskryminacyjna.

1. Przygotowanie danych

- Podział danych na zbiór treningowy i testowy
- Standaryzacja zmiennych (aby każda zmienna miała średnią równą 0 i wariancję równą 1)

2. Przeprowadzenie analizy dyskryminacyjnej
   - Obliczenie macierzy kowariancji
   - Obliczenie wektora średnich dla każdej klasy
   - Obliczenie współczynników dyskryminacyjnych
3. Predykcja
   - Predykcja na zbiorze testowym
   - Klasyfikacja nowych obserwacji na podstawie obliczonych współczynników dyskryminacyjnych

Model M1

	0	1
0	50	22
1	60	201

Model M2

	0	1
0	49	26
1	61	197

Model M3

	0	1
0	43	27
1	67	196

Tabela 13: Ocena dokładności klasyfikacji modeli M1, M2 oraz M3

Model M4

	0	1
0	11	12
1	99	211

Model M5

	0	1
0	12	15
1	98	208

Tabela 14: Ocena dokładności klasyfikacji modeli M4, M5

Model	Wartość błędu
M1	0.2462462
M2	0.2612613
M3	0.2822823
M4	0.3333333
M5	0.3393393

Tabela 15: Błąd klasyfikacji na zbiorze testowym we wszystkich modelach

Przyglądając się tabelom (18) oraz (19) widzimy, że właściwie żaden z modeli nie dał nam efektu "wow" w klasyfikacji. Bardzo dobrze klasyfikują one dobrych klientów jako faktycznie dobrych, jednak tych złych jako prawdziwie złych - znacznie gorzej.

Analizując tabelę (15) widzimy, że najmniejszy błąd wyznacza nam model M1.

Pamiętając o macierzy kosztów (12), sprawdzimy, który z modeli wyznacza nam najmniej błędów.

- Model M1 $\rightarrow 22 * 5 + 60 * 1 = 170$

- Model M2 $\rightarrow 26 * 5 + 61 * 1 = 191$
- Model M3 $\rightarrow 27 * 5 + 67 * 1 = 202$
- Model M4 $\rightarrow 12 * 5 + 99 * 1 = 159$
- Model M5 $\rightarrow 15 * 5 + 98 * 1 = 173$

Możemy stwierdzić zatem, że najmniej błędów na zbiorze treningowym wyznacza nam model M4. Moglibyśmy zatem uznać, że jest to najlepszy model. Jednak zanim dokonamy takiego stwierdzenia, musimy sprawdzić, jak będzie działał nasz model na danych ze zbioru testowego.

Korzystając z funkcji `confusionMatrix` dostępnej w pakiecie `caret` [9] wyznaczymy podstawowe wskaźniki, każdego z naszych modeli, na zbiorze testowym. Sprawdzimy tym samym, który z naszych modeli najlepiej się sprawdził.

Warto dodać jeszcze, że współczynnik Kappa Cohena określa stopień zgodności dwukrotnych pomiarów tej samej zmiennej w różnych warunkach. Jego wartość zawiera się w przedziale od -1 do 1. Wartość 1 oznacza pełną zgodność, wartość 0 oznacza zgodność na poziomie takim samym jaki powstałby dla losowego rozłożenia danych w tabeli kontyngencji. Poziom pomiędzy 0 a -1 jest w praktyce niewykorzystywany.

Model	Accuracy	Kappa	Sensitivity	Sensitivity	PPV	NPV
M1	0.7357	0.3353	0.4000	0.9013	0.6667	0.7528
M2	0.7267	0.3282	0.4273	0.8744	0.6267	0.7558
M3	0.7087	0.2578	0.3364	0.8924	0.6066	0.7316
M4	0.6637	0.0465	0.09091	0.94619	0.45455	0.67846
M5	0.6577	0.0182	0.06364	0.95067	0.38889	0.67302

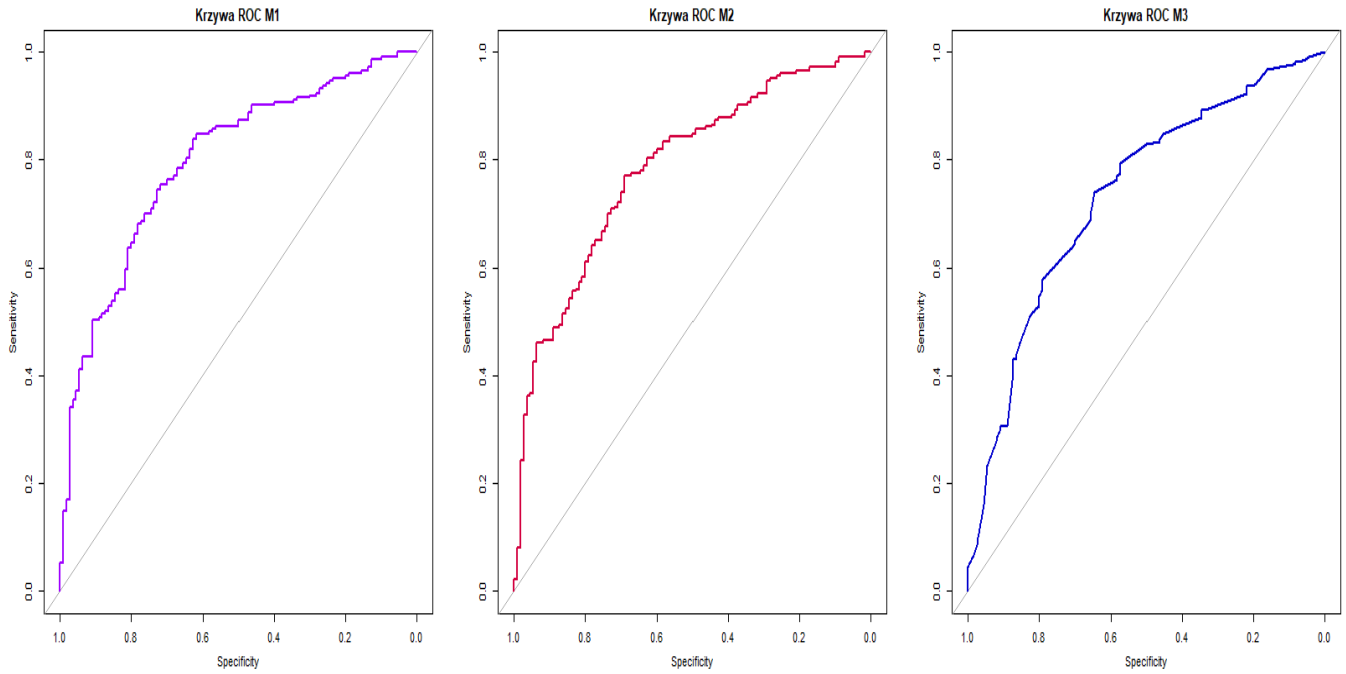
Tabela 16: Błąd klasyfikacji na zbiorze testowym we wszystkich modelach

Powyższa tabela (17) pokazuje nam analizę danych testowych na podstawie modeli, jakie wybraliśmy do zbioru uczącego. Widzimy, że najlepsza dokładność występuje w przypadku modelu M1. Równocześnie, dla tego samego modelu mamy najwyższy współczynnik Kappa. Pozostałe, bardzo ważne współczynniki, w większości również spełniają ustalone, potrzebne założenia. Możemy zatem uznać, że najlepszym modelem jest model M1.

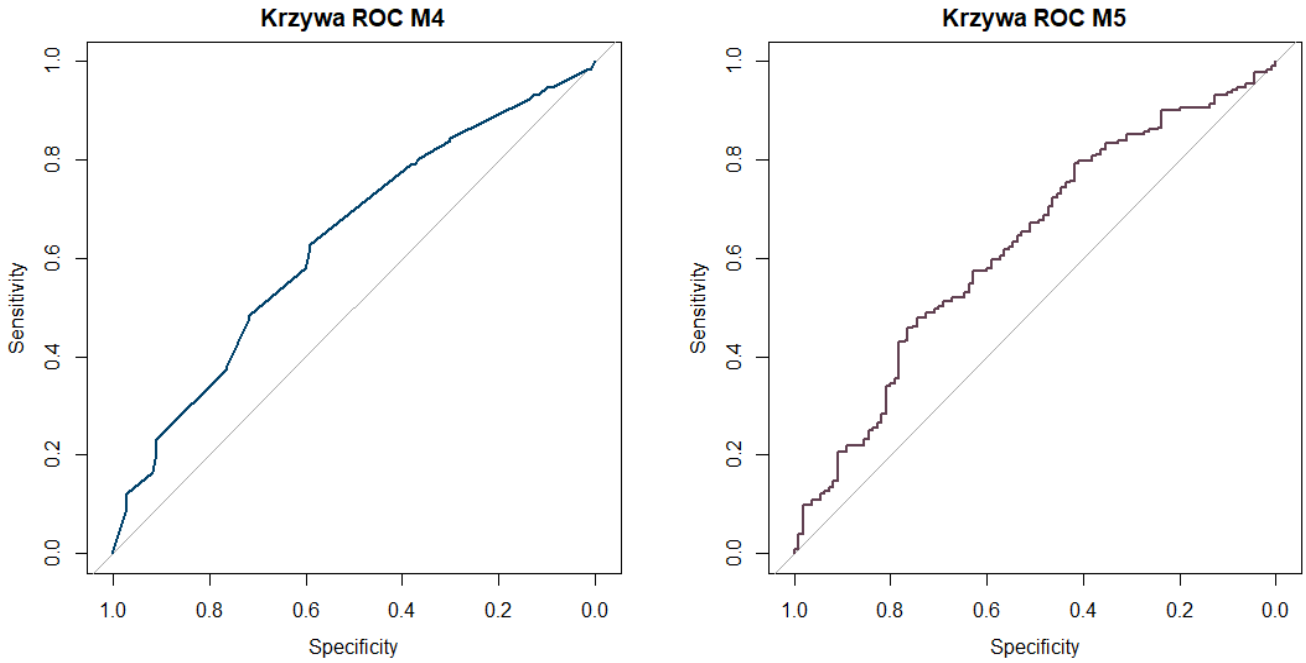
Poniżej prezentujemy krzywe ROC każdego z wybranych modeli. Krzywa ROC (ang. Receiver Operating Characteristic) to wykres przedstawiający wydajność klasyfikatora binarnego w zależności od wartości progu decyzyjnego. Krzywa ROC przedstawia stosunek liczby prawdziwie pozytywnych przypadków (True Positive Rate - TPR) do liczby fałszywie pozytywnych przypadków (False Positive Rate - FPR) w zależności od różnych wartości progowych. TPR to stosunek liczby prawdziwie sklasyfikowanych pozytywnych przypadków do liczby wszystkich pozytywnych przypadków w populacji. FPR to stosunek liczby fałszywie sklasyfikowanych pozytywnych przypadków do liczby wszystkich negatywnych przypadków w populacji. Im wyższa jest wartość TPR, tym lepsza jest zdolność klasyfikatora do wykrywania pozytywnych przypadków,

[9] Informacje na temat funkcji dostępne pod adresem `https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix`, zaś informacje na temat pakietu `https://cran.r-project.org/web/packages/caret/index.html`

a im niższa wartość FPR, tym mniejsza jest liczba negatywnych przypadków sklasyfikowanych błędnie jako pozytywne. Krzywa ROC jest przydatna w ocenie jakości modeli klasyfikacji binarnej, ponieważ umożliwia porównanie wydajności różnych modeli lub różnych wartości progowych w jednym wykresie. Im większa powierzchnia pod krzywą ROC (AUC - Area Under Curve), tym lepsza jakość klasyfikacji.



Rysunek 13: Krzywe ROC model M1, M2, M3 - opracowanie własne



Rysunek 14: Krzywe ROC model M4, M5 - opracowanie własne

Model	AUC
M1	0.7949
M2	0.7911
M3	0.7579
M4	0.6334
M5	0.6268

Tabela 17: Wartość krzywej ROC oraz AUC wszystkich modeli

Ponownie, największe pole AUC ma model M1. Czyli możemy stwierdzić, że najlepszym modelem do tej pory jest właśnie ten model. Zawiera on wszystkie zmienne z naszego założenia.

Będąc jednak pracownikami banku, nie uznalibyśmy 73% dokładności jako coś zadowalającego. Śmiało uznalibyśmy, że jest to zbyt niedokładny model, żeby się nim posługiwać i mógłby przynieść duże straty dla banku. Dlatego w dalszej części sprawozdania postaramy się znaleźć lepszy model, z lepszą jakością dopasowania.

1.2.2 Kwadratowa analiza dyskryminacja (QDA)

Kwadratowa reguła klasyfikacyjna (QDA) polega na tym, że dzielimy zbiór danych na klasy i dla każdej klasy estymujemy funkcję gęstości prawdopodobieństwa. Następnie dla nowej obserwacji obliczamy wartości funkcji gęstości dla każdej klasy, a następnie przypisujemy ją do klasy, dla której wartość funkcji gęstości była największa.

W przypadku klasyfikacji za pomocą kwadratowej reguły klasyfikacyjnej korzystać będziemy z funkcji `qda` ze znanego pakietu `MASS`. Przeprowadzimy klasyfikację opartą na estymacji funkcji gęstości dla każdej z klas. Dane przygotowujemy tak samo, jak w poprzednim punkcie.
W przypadku kwadratowej analizy dyskryminacyjnej również będziemy analizować te same modele, co wcześniej, to znaczy

- (M1) Model w oparciu o wszystkie zmienne
- (M2) Model w oparciu o Backward AIC $\rightarrow response \sim chk\_acct + duration + credit\_his + purpose + saving\_acct + present\_emp + other\_debtor + age + other\_install + housing + foreign$
- (M3) Model w oparciu o Backward BIC $\rightarrow response \sim chk\_acct + duration + credit\_his$
- (M4) Model w oparciu o bibliotekę `klaR` [10] oraz metodę $lda \rightarrow response \sim duration$
- (M5) Model w oparciu o bibliotekę `klaR` oraz metodę $qda \rightarrow response \sim duration + amount$

Model M1

	0	1
0	57	46
1	40	190

Model M2

	0	1
0	61	50
1	36	186

Model M3

	0	1
0	45	21
1	52	215

Tabela 18: Ocena dokładności klasyfikacji modeli M1, M2 oraz M3

[10] Biblioteka posiada wiele różnych funkcji służących do klasyfikacji oraz wizualizacji danych. Więcej informacji można znaleźć pod linkiem `https://www.rdocumentation.org/packages/klaR/versions/1.7-2`

Model M4

	0	1
0	15	16
1	82	226

Model M5

	0	1
0	27	21
1	70	215

Tabela 19: Ocena dokładności klasyfikacji modeli M4, M5

Podobnie jak wcześniej, żaden z modeli nie uzyskał bardzo wysokiej skuteczności, jednakże można dostrzec poprawę wszystkich modeli. Sprawdźmy, który z modeli daje nam najmniej błędów (uwzględniając macierz kosztów (12)).

- Model M1 $\rightarrow 1 * 40 + 5 * 46 = 270$
- Model M2 $\rightarrow 1 * 36 + 50 * 5 = 286$
- Model M3 $\rightarrow 1 * 52 + 21 * 5 = 157$
- Model M4 $\rightarrow 1 * 82 + 5 * 16 = 162$
- Model M5 $\rightarrow 1 * 70 + 5 * 21 = 175$

Na podstawie powyższego, możemy stwierdzić, że kandydatem na najlepszy model, jest M3. Analizując dodatkowo poniższą tabelę (20) widzimy, że faktycznie model ten posiada najniższą wartość błędu klasyfikacji na zbiorze treningowym.

Model	Wartość błędu
M1	0.2582583
M2	0.2582583
M3	0.2192192
M4	0.2762763
M5	0.2732733

Tabela 20: Błąd klasyfikacji na zbiorze testowym we wszystkich modelach

Podsumowując, kwadratowa reguła klasyfikacyjna (QDA) polega na estymacji funkcji gęstości dla każdej z klas, a następnie na przypisaniu nowych obserwacji do klasy, dla której wartość funkcji gęstości była największa.

Nie ma jednoznacznej odpowiedzi na pytanie, która metoda - LDA czy QDA - jest lepsza w przypadku naszych danych, ponieważ wybór metody zależy od charakterystyki danych oraz od celów analizy. LDA jest mniej skomplikowana niż QDA i zwykle wymaga mniejszej liczby parametrów do oszacowania, co oznacza, że jest bardziej stabilna, gdy liczba obserwacji jest mała w stosunku do liczby zmiennych. Jednakże LDA zakłada, że macierze kowariancji dla poszczególnych klas są równe, co może być nieprawdziwe w przypadku, gdy macierze te są różne. Z drugiej strony, QDA może być bardziej elastyczna niż LDA, ponieważ nie zakłada, że macierze kowariancji dla poszczególnych klas są równe. Dzięki temu QDA może lepiej radzić sobie z różnorodnością klas. Jednak QDA może być bardziej podatna na przeuczenie, zwłaszcza gdy liczba zmiennych jest duża w stosunku do liczby obserwacji. W przypadku danych *German Credit Data*, zarówno LDA, jak i QDA, osiągają podobną dokładność klasyfikacji, co sugeruje, że obie metody są w stanie dobrze poradzić sobie z tym zbiorem danych.

1.2.3 Metoda k - najbliższych sąsiadów (k -NN)

Metoda k-NN polega na przyporządkowaniu etykiet klasy nowej obserwacji na podstawie etykiet k najbliższych sąsiadów w zbiorze treningowym. K to parametr, który należy ustawić przed analizą.

Ostateczna dokładność klasyfikacji zależy od wybranej wartości k oraz od tego, czy zastosowane zmienne są odpowiednie dla danego problemu klasyfikacyjnego. Jednakże warto pamiętać, że metoda k-NN może być mniej skuteczna niż LDA i QDA w przypadku danych o wyższych wymiarach lub gdy występują skomplikowane zależności między zmiennymi.

	0	1
0	27	41
1	72	193

Tabela 21: Macierz klasyfikacji dla k = 5

Widzimy, że w przypadku losowo wybranego parametru k, nasz wynik jest całkowicie zły. Ponadto, obliczyliśmy błąd klasyfikacji, który wynosi 0.3393393. Losowe wybieranie parametru nie jest sensowne, dlatego, korzystając z krzyżowej walidacji wybierzemy najlepszy parametr k dla naszych danych.

W przypadku metody k-najbliższych sąsiadów, jeden z parametrów, który musi być ustalony, to wartość k, czyli liczba najbliższych sąsiadów branych pod uwagę w procesie klasyfikacji. Aby wybrać odpowiednią wartość k, można zastosować walidację krzyżową.

W przypadku walidacji krzyżowej dla k-najbliższych sąsiadów, dzielimy zbiór danych na k części (np. 5 lub 10). Następnie w każdej iteracji wybieramy jedną z tych części jako zbiór testowy, a pozostałe części jako zbiór treningowy. Dla każdej wartości k, przeprowadzamy proces walidacji krzyżowej i obliczamy średnią skuteczność klasyfikacji dla wszystkich iteracji. Wartość k, dla której osiągana jest najwyższa skuteczność, wybieramy jako ostateczną wartość parametru k.

Poniższy rysunek (15) pokazuje, w jaki sposób możemy wyznaczyć najlepszy parametr k.

	0	1
0	10	4
1	89	230

Tabela 22: Macierz klasyfikacji dla k = 23

Widzimy, że dzięki wyznaczeniu najlepszego parametru k dostaliśmy bardzo mało błędów kosztujących bank najwięcej. Ponadto błąd klasyfikacji wynosi 0.2792793, co jest zdecydowanie mniej niż poprzedni przypadek.

```r
set.seed(255707)
kfold <- trainControl(method = "cv", number = 10)
knn_model <- train(response ~ ., data = learning.set, method = "knn",
  tuneLength = 10, trControl = kfold)
print(knn_model)
```

```
## k-Nearest Neighbors
##
## 666 samples
##  16 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 599, 599, 599, 600, 599, 600, ...
## Resampling results across tuning parameters:
##
##   k   RMSE       Rsquared    MAE
##    5  0.4937193  0.01742214  0.4149578
##    7  0.4873324  0.01576819  0.4200976
##    9  0.4817702  0.01444665  0.4221303
##   11  0.4761993  0.01358016  0.4193403
##   13  0.4723061  0.01779507  0.4192551
##   15  0.4730883  0.01532098  0.4214795
##   17  0.4694970  0.02300635  0.4203745
##   19  0.4672833  0.02901475  0.4189340
##   21  0.4676565  0.02754502  0.4199524
##   23  0.4671440  0.02821604  0.4206652
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 23.
```

Rysunek 15: Wybór najlepszego parametru k

Ostatecznie, skuteczność metody k-NN zależy od wielu czynników, takich jak jakość danych, liczba wymiarów cech, wartość parametru k, oraz stopień niezrównoważenia klas. W niektórych przypadkach może być to skuteczna metoda klasyfikacji, ale w innych lepiej sprawdzą się bardziej zaawansowane algorytmy.

Podsumowując, najlepszą metodą okazała się kwadratowa analiza dyskryminacyjna w przypadku modelu M3 wyznaczonego w oparciu o Backward BIC $\rightarrow$ $response \sim chk\_acct + duration + credit\_his$.

1.2.4 Metoda sieci neuronowych

Istnieje oczywiście wiele metod klasyfikacji, które mogą być zastosowane do naszego zbioru danych, przykładowo

- Drzewa decyzyjne - to metoda klasyfikacji polegająca na budowaniu drzewa, którego węzłami są testy na wartościach cech, a liśćmi - etykiety klas. Drzewa decyzyjne mogą być wykorzystane do klasyfikacji, jak i do regresji.

- Las losowy - to metoda klasyfikacji polegająca na budowaniu wielu drzew decyzyjnych i wykorzystywaniu ich do klasyfikacji nowych przykładów. W każdym drzewie losowo wybiera się podzbiór danych oraz podzbiór cech.

- Metoda wektorów nośnych (SVM) - to metoda klasyfikacji polegająca na znalezieniu hiperpłaszczyzny, która najlepiej separuje klasy. SVM może być stosowane do problemów klasyfikacji binarnej, jak i wieloklasowej.

- Sieci neuronowe - to metoda klasyfikacji, która naśladuje działanie ludzkiego mózgu. Sieci neuronowe składają się z neuronów, które przetwarzają sygnały wejściowe i przekazują je do kolejnych warstw neuronów, aż do uzyskania wyniku.

- Klasyfikator Bayesa - to metoda klasyfikacji, która opiera się na założeniu, że prawdopodobieństwo przynależności do klasy jest zależne od wartości cech. Klasyfikator Bayesa może być stosowany do problemów binarnej klasyfikacji, jak i wieloklasowej.

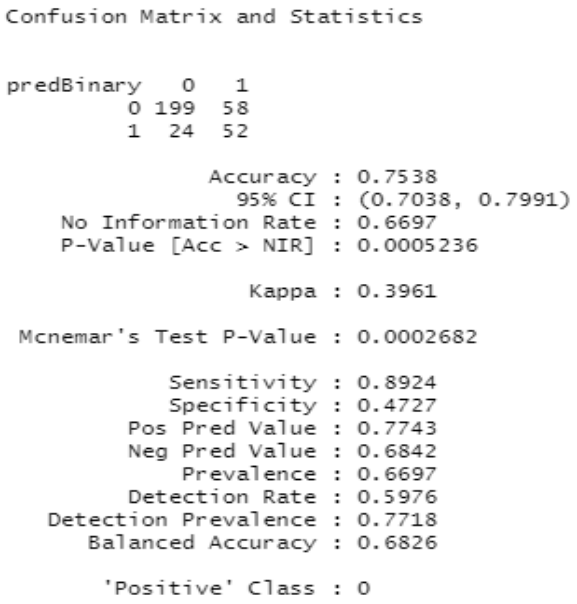
Oczywiście, do klasyfikacji naszych danych na podstawie sieci neuronowych można użyć wielu różnych podejść. Jednym z popularnych podejść jest sieć neuronowa typu MLP (Multilayer Perceptron), którą właśnie wypróbujemy.

Sieć MLP (Multilayer Perceptron) jest jednym z popularnych modeli sieci neuronowych stosowanych do klasyfikacji i predykcji. Składa się z jednej lub więcej warstw ukrytych, a każda warstwa zawiera wiele neuronów. Każdy neuron otrzymuje sygnały wejściowe, wykonuje obliczenia na podstawie tych sygnałów i przekazuje wynik do kolejnej warstwy lub do wyjścia sieci.
Wejścia sieci są zwykle normalizowane, a wagi neuronów są inicjalizowane losowo. Sieć MLP uczy się na podstawie zestawu danych treningowych poprzez dostosowywanie wag w celu minimalizacji błędu predykcji. Do tego celu stosowane są różne algorytmy optymalizacyjne, np. algorytm propagacji wstecznej.
Sieć MLP jest stosunkowo prostą i łatwą do nauczenia siecią neuronową, która może być skuteczna w klasyfikacji i predykcji, szczególnie gdy dane są dobrze znormalizowane i mają duże rozmiary. Jednakże, sieci MLP mogą być podatne na problemy związane z overfittingiem i wymagają odpowiedniego doboru liczby warstw ukrytych oraz liczby neuronów w każdej warstwie, co może być czasochłonne i wymagać doświadczenia.

Poniższy rysunek (16) przedstawia nasze wartości odwrotnie. To znaczy, wartość 0 w tym przypadku oznacza dobrego klienta, zaś wartość 1 - złego klienta.
Na podstawie macierzy błędu z rysunku (16) można wywnioskować, że model ma tendencję do przewidywania pozytywnego wyniku (1) z większą czułością niż specyficznością, co oznacza, że model ma mniejszą skłonność do przewidywania negatywnego wyniku (0).

Dokładność modelu wynosi 0.7538, co oznacza, że model dobrze klasyfikuje wyniki w 75% przypadków.

Wartość kappa wynosi 0.3961, co oznacza, że poziom zgodności między przewidywaniami a rzeczywistymi wynikami jest umiarkowany.



Rysunek 16: Model sieci neuronowych

Można zauważyć, że wartość negatywnego predykcyjnego wyniku (Neg Pred Value) wynosi 0.6842, co sugeruje, że model ma tendencję do błędnego przewidywania negatywnych wyników. Natomiast wartość pozytywnego predykcyjnego wyniku (Pos Pred Value) wynosi 0.7743, co sugeruje, że model ma tendencję do poprawnego przewidywania pozytywnych wyników.

Podsumowując, przeprowadzone zostały cztery różne modele predykcyjne: regresja logistyczna, kwadratowa analiza dyskryminacyjna (QDA), metoda k - najbliższych sąsiadów oraz wielowarstwowy perceptron (MLP). Do oceny jakości modeli wykorzystaliśmy metrykę Accuracy, która określa procent poprawnych predykcji. Najwyższe Accuracy osiągnął model regresji logistycznej. Kolejną metryką była precyzja, która określa procent przypadków pozytywnych, które zostały prawidłowo sklasyfikowane. W przypadku pozytywnych przypadków, najlepszy wynik osiągnął model MLP. Metryka Recall, określa procent pozytywnych przypadków, które zostały poprawnie wykryte. W przypadku pozytywnych przypadków, najwyższy wynik uzyskał ponownie model MLP z recall wynoszącym 0.66. Warto zauważyć, że wartości metryk dla klasy pozytywnej są w ogólnym przypadku niższe niż dla klasy negatywnej, co sugeruje, że modele mają trudności z poprawnym rozpoznawaniem przypadków pozytywnych.

Literatura

[1] Adam Zagdański *Eksploracyjna analiza danych (EDA) - wprowadzenie*, https://eportal.pwr.edu.pl/pluginfile.php/559567/mod_resource/content/0/EDA_wprowadzenie.pdf