Pozyskiwanie Wiedzy
Projekt - Część I

Justyna Domańska i Julia Lenczewska

29 stycznia 2020

Spis treści

1 Cel raportu

Bank otrzymując wszelkie dokumenty dotyczące klientów chcących wziąć kredyt musi na podstawie przedstawionych informacji podjąć decyzję odnośnie tego czy powinien przyznać kredyt czy też nie. Wiąże się to z dwoma rodzajami ryzyka. Możliwe jest przyznanie "dobremu" klientowi statusu "złego" lub "złemu" "dobrego" (za dobrego klienta uważamy takiego, który będzie w stanie spłacić kredyt, za złego - który prawdopodobnie go nie spłaci). W pierwszej sytuacji bank straci możliwość uzyskania przychodzu związanego m.in. z odsetkami. W drugim natomiast naraża się na wysokie straty w związku z możliwością konieczności przeprowadzenia postępowania windykacyjnego. Oczywistym wydaje się, że dla banku lepiej jest stracić jednego dobrego klienta, niż popełnić błąd przyznając kredyt osobie, która nie będzie w stanie go spłacać.

Celem raportu jest uzyskanie modelu, służącego do klasyfikacji klientów jako dobrych i złych.

2 Opis danych

Zbiór danych *German Credit* został zebrany i udostępniony przez profesora dr Hansa Hofmanna w 1994 roku. Pochodzi ze strony http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data). Są to dane finansowe zawierające informacje o klientach ubiegających się o kredyt. Dane składają się z 21 kolumn oraz 1000 obserwacji (21. kolumna mówi o tym czy klient otrzymał kredyt czy nie). W omawianym zbiorze nie występują wartości brakujące. Wśród zmiennych wpływających na ocenę zdolności kredytowej klienta mamy 13 zmiennych jakościowych i 7 ilościowych. Typ ostatniej zmiennej (creditability) jest domyślnie ustawiony jako *integer*, jednak my zmieniamy go na *factor*. Dodatkowo wartości, które przyjmuje ta zmienna, tj. 1 i 2, zmieniamy na odpowiednio 0 i 1. Wartość 0 oznacza, że klient jest "dobry", zaś 1 - "zły".

L.p.	Nazwa zmiennej	Opis zmiennej	Typ
1	Status of existing checking account	Stan konta bankowego	factor
2	Duration in month	Czas trwania (w miesiącach)	integer
3	Credit history	Historia kredytowa	factor
4	Purpose	Cel	factor
5	Credit amount	Kwota kredytu	integer
6	Savings account	Oszczędności	factor
7	Present employment since	Obecne zatrudnienie od	factor
8	Installment rate in percentage of disposable income	Udział obecnych rat w zarobkach	integer
9	Personal status and sex	Status i płeć	factor
10	Other debtors	Inni dłużnicy/poręczyciele	factor
11	Present residence since	Obecne miejsce zamieszkania od	integer
12	Property	Własność	factor
13	Age in years	Wiek (w latach)	integer
14	Other installment plans	Inne plany ratalne	factor
15	Housing	Dom	factor
16	Number of existing credits at this bank	Liczba kredytów w tym banku	integer
17	Job	Praca	factor
18	Number of people being liable to provide maintenance for	Liczba osób na utrzymaniu	integer
19	Telephone	Telefon	factor
20	foreign worker	Pracownik zagraniczny	factor
21	creditability	Zdolność kredytowa	factor

Tabela 1: Lista zmiennych (kolumn) dla danych German Credit

Tabela 1 przedstawia informacje dotyczące zmiennych w zbiorze German Credit takie jak nazwa zmiennej, jej opis - a dokładniej tłumaczenie na język polski oraz typ zmiennej.

Nazwa zmiennej	Kategorie	Opis
Status of existing checking account	A11	... < 0 DM
	A12	0 <= ... < 200 DM
	A13	... >= 200 DM /zapewniona pensja przez co najmniej rok
	A14	brak rachunku bankowego
Credit history	A30	brak kredytów/wszystkie kredyty spłacone prawidłowo
	A31	wszystkie kredyty w tym banku spłacone prawidłowo
	A32	istniejące kredyty spłacane do tej pory prawidłowo
	A33	opóźnienie w spłatach w przeszłości
	A34	kredyt zagrożony/istnieją inne kredyty (nie w tym banku)
Purpose	A40	nowy samochód
	A41	używany samochód
	A42	meble, wyposażenie
	A43	radio/telewizja
	A44	sprzęt gospodarstwa domowego
	A45	remonty
	A46	edukacja
	A47	wakacje
	A48	przekwalifikowanie
	A49	biznes
	A410	inne
Savings account	A61	... < 100 DM
	A62	100 <= ... < 500 DM
	A63	500 <= ... < 1000 DM
	A64	.. >= 1000 DM
	A65	nieznane/ brak konta oszczędnościowego
Present employment since	A71	bezrobotny
	A72	... < 1 rok
	A73	1 <= ... < 4 lat
	A74	4 <= ... < 7 lat
	A75	.. >= 7 lat
Personal status and sex	A91	mężczyzna: rozwiedziony/w separacji
	A92	kobieta: rozwiedziona/w separacji/zamężna
	A93	mężczyzna: singiel
	A94	mężczyzna: żonaty/wdowiec
	A95	kobieta: singielka
Other debtors	A101	brak
	A102	współkredytobiorca
	A103	poręczyciel
Property	A121	nieruchomość
	A122	jeśli nie A121: kasa mieszkaniowa/ubezpieczenie na życie
	A123	jeśli nie A121/A122: samochód lub inne, poza kontem oszcz.
	A124	nieznane/ brak własności
Other installment plans	A141	bank
	A142	sklepy

	A143	brak
Housing	A151	wynajmowane
	A152	własne
	A153	za darmo
Job	A171	bezrobotny/ niewykwalifikowany - zamiejscowy
	A172	niewykwalifikowany - rezydent
	A173	pracownik wykwalifikowany / urzędnik
	A174	kierownictwo/ samozatrudniony/pracownik wysoko wykwal. / f
Telephone	A191	brak
	A192	tak, zarejestrowany pod nazwiskiem klienta
foreign worker	A201	tak
	A202	nie

Tabela 2: Opis kategorii zmiennych typu factor

Tabela 2 informuje nas o tym, jakie poziomy przyjmują zmienne kategoryczne, a także co dane poziomy oznaczają.

3 Statystyki opisowe

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd
Duration in month	4.00	12.00	18.00	20.90	24.00	72.00	12.06
Credit amount	250.00	1365.50	2319.50	3271.26	3972.25	18424.00	2822.74
Installment rate in percentage of disposable income	1.00	2.00	3.00	2.97	4.00	4.00	1.12
Present residence since	1.00	2.00	3.00	2.85	4.00	4.00	1.10
Age in years	19.00	27.00	33.00	35.55	42.00	75.00	11.38
Number of existing credits at this bank	1.00	1.00	1.00	1.41	2.00	4.00	0.58
Number of people being liable to provide maintenance for	1.00	1.00	1.00	1.16	1.00	2.00	0.36

Tabela 3: Statystyki opisowe dla zmiennych ilościowych

Tabela 3 przedstawia statystyki opisowe dla zmiennych, których typ ustawiony jest jako ”integer”. Obliczone statystyki to minimum, pierwszy kwartyl, mediana, średnia, trzeci kwartyl, maksimum oraz odchylenie standardowe. Widzimy, że najkrótszy czas trwania kredytu wynosi 4 miesiące, a najdłuższy 6 lat. Kwoty kredytu są bardzo zróżnicowane. Średnia to ok. 3271 marek niemieckich. Wysokość najwyższego kredytu, dużo większa niż średnia, może świadczyć o tym, że mamy do czynienia z obserwacjami odstającymi. Wiek ubiegających się o kredyt waha się od 19 do 75 roku życia.

Status of existing checking account	Credit history	Purpose	Savings account	Present employment since	Personal status and sex	Other debtors
A11:274	A30: 40	A43 :280	A61:603	A71: 62	A91: 50	A101:907
A12:269	A31: 49	A40 :234	A62:103	A72:172	A92:310	A102: 41
A13: 63	A32:530	A42 :181	A63: 63	A73:339	A93:548	A103: 52
A14:394	A33: 88	A41 :103	A64: 48	A74:174	A94: 92	
	A34:293	A49 : 97	A65:183	A75:253		
		A46 : 50				
		(Other): 55				

Tabela 4: Statystyki opisowe dla zmiennych jakościowych

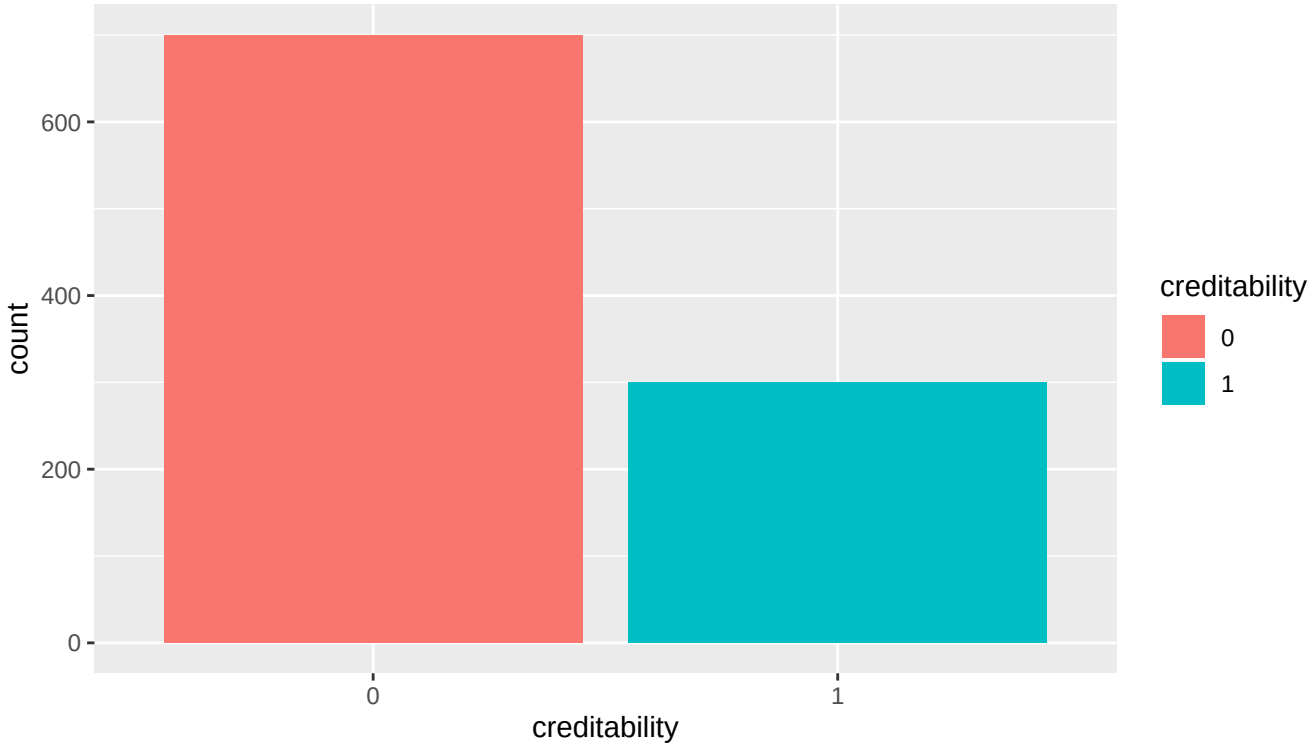
Property	Other installment plans	Housing	Job	Telephone	foreign worker	creditability
A121:282	A141:139	A151:179	A171: 22	A191:596	A201:963	0:700
A122:232	A142: 47	A152:713	A172:200	A192:404	A202: 37	1:300
A123:332	A143:814	A153:108	A173:630			
A124:154			A174:148			

Tabela 5: Statystyki opisowe dla zmiennych jakościowych

W tabelach 4-5 widzimy liczność wszystkich poziomów zmiennych jakościowych. Najważniejszą informacją jest dla nas liczność grup w podziale ze względu na zmienną creditability. Liczba klientów "dobrych" (poziom 0) wynosi 700, zaś "złych" (poziom 1) - 300. Zatem nasze dane możemy uznać za niezbalansowane.

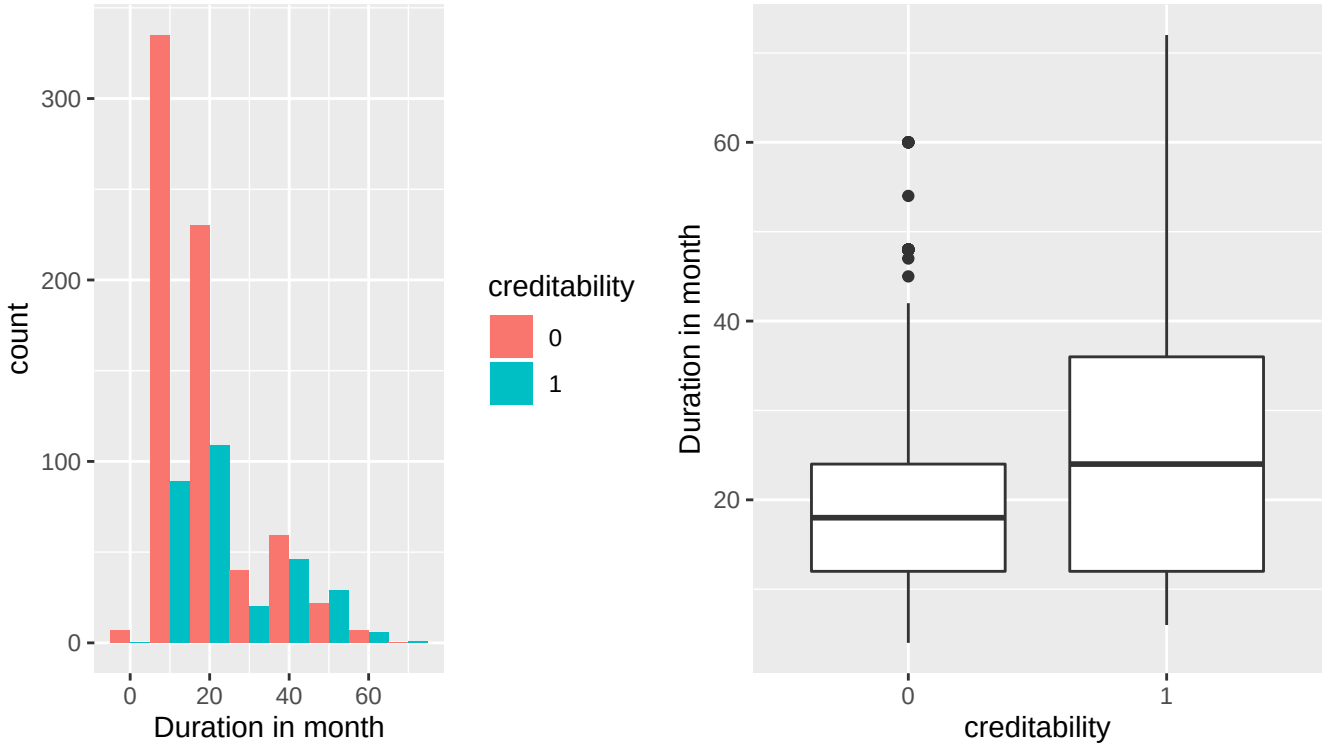
4 Wykresy

W tym rozdziale zaprezentujemy wykresy mogące pomóc w zrozumieniu danych. Dla zmiennych ciągłych zaprezentujemy histogramy, dodatkowo dla tych z nich które przyjmują zróżnicowane wartości, umieścimy box-ploty. Dla zmiennych kategorycznych narysujemy wykresy słupkowe pokazujące liczność danej zmiennej z podziałem na creditability.



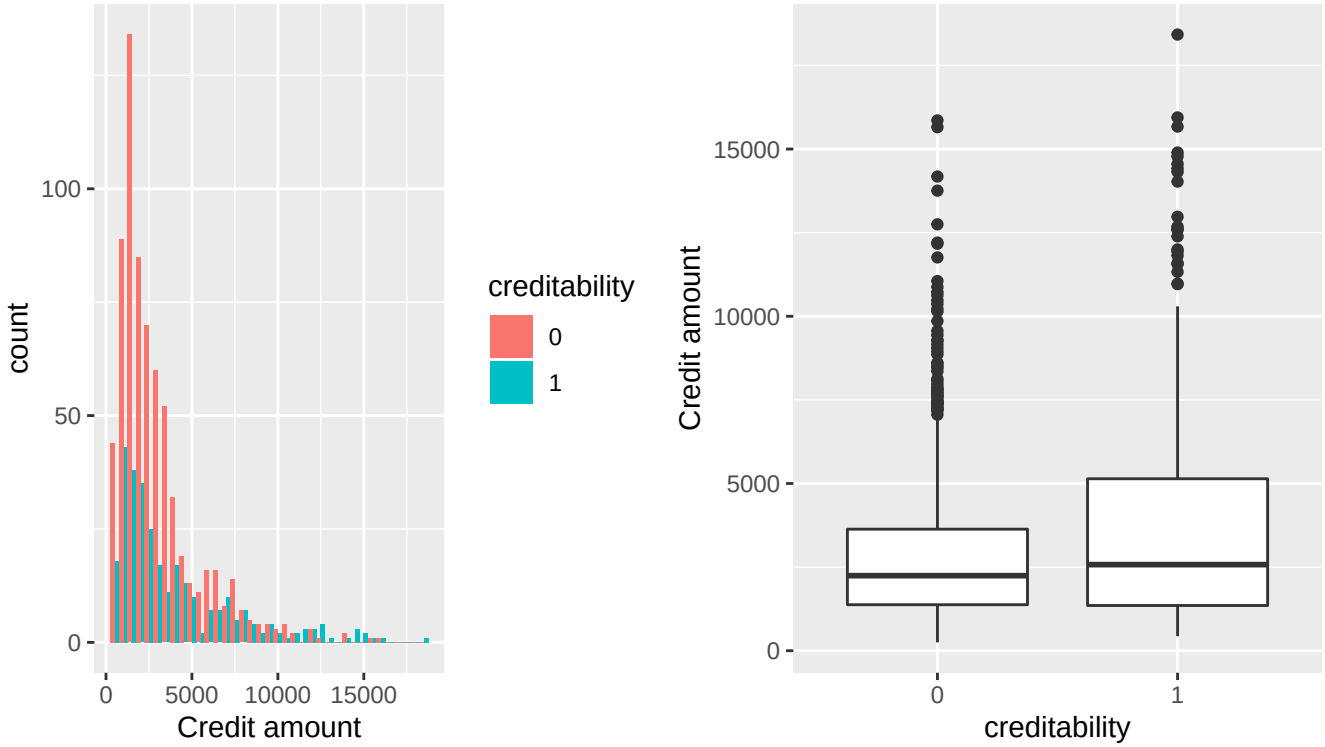
Rysunek 1: Histogram dla zmiennej creditability

Rysunek 1 potwierdza rozkład zmiennej creditability, mamy 300 obserwacji należących do klasy pozytywnej oraz 700 do negatywnej.



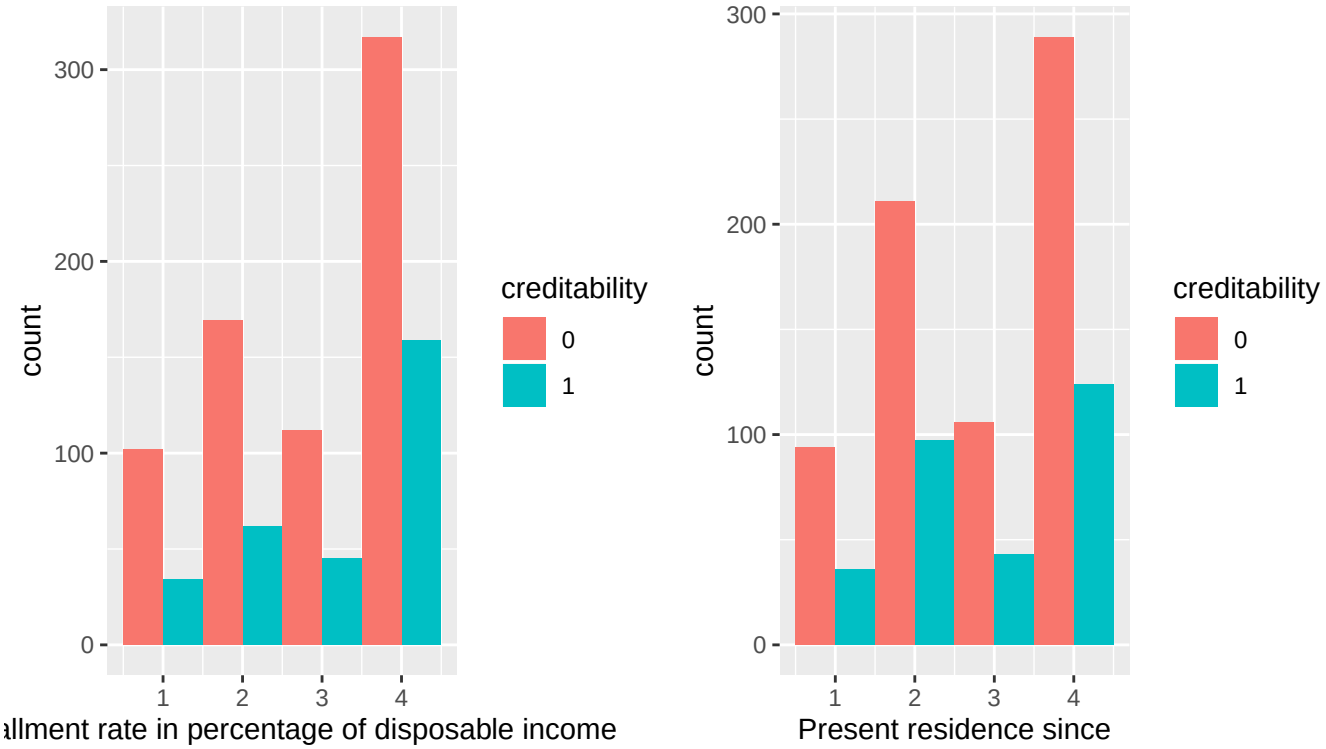
Rysunek 2: Histogram i box-plot dla zmiennej Duration in month

Na Rysunku 2 zaprezentowany jest box-plot oraz histogram dla zmiennej Duration in month. Możemy zauważyć, że mediana czasu trwania jest wyższa w przypadku osób które kredytu nie dostały.

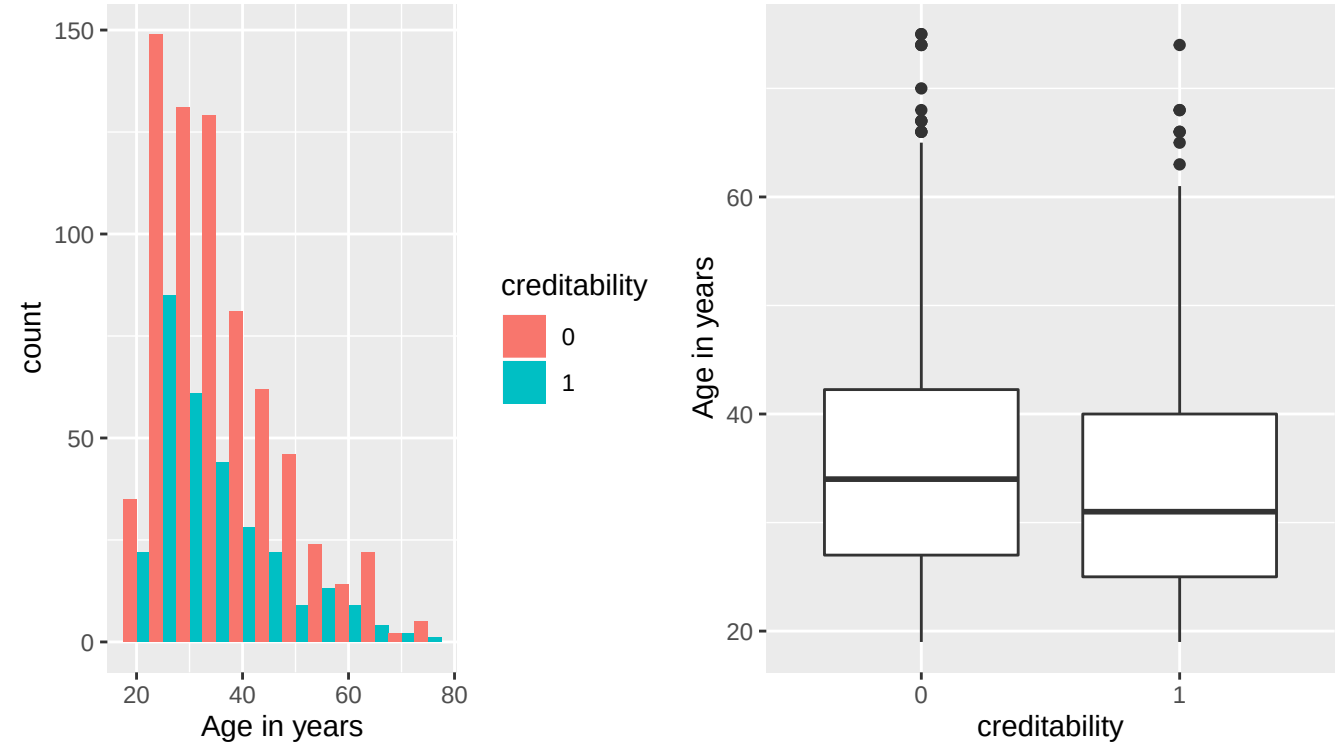


Rysunek 3: Histogram i box-plot dla zmiennej Credit amount

Na zawartym w Rysunku 3 box-plocie widzimy, że to czy otrzymano kredyt czy nie, zależy od jego kwoty. Przy wyższych kwotach kredytu liczba osób, które nie dostały kredytu zaczyna przeważać nad liczbą tych, które go otrzymały.



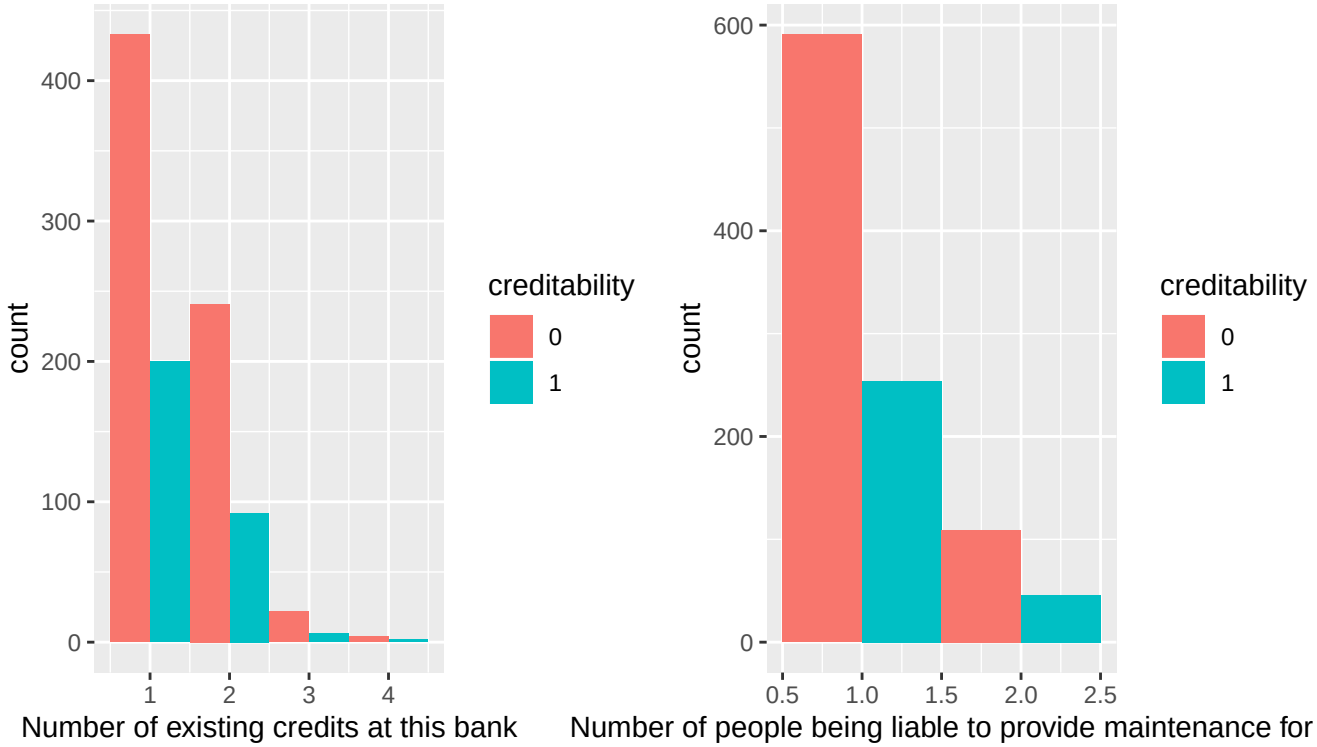
Rysunek 4: Histogramy dla zmiennych kolejno Installment rate in percentage of disposable income i Present residence since



Rysunek 5: Histogram i box-plot dla zmiennej Age in years

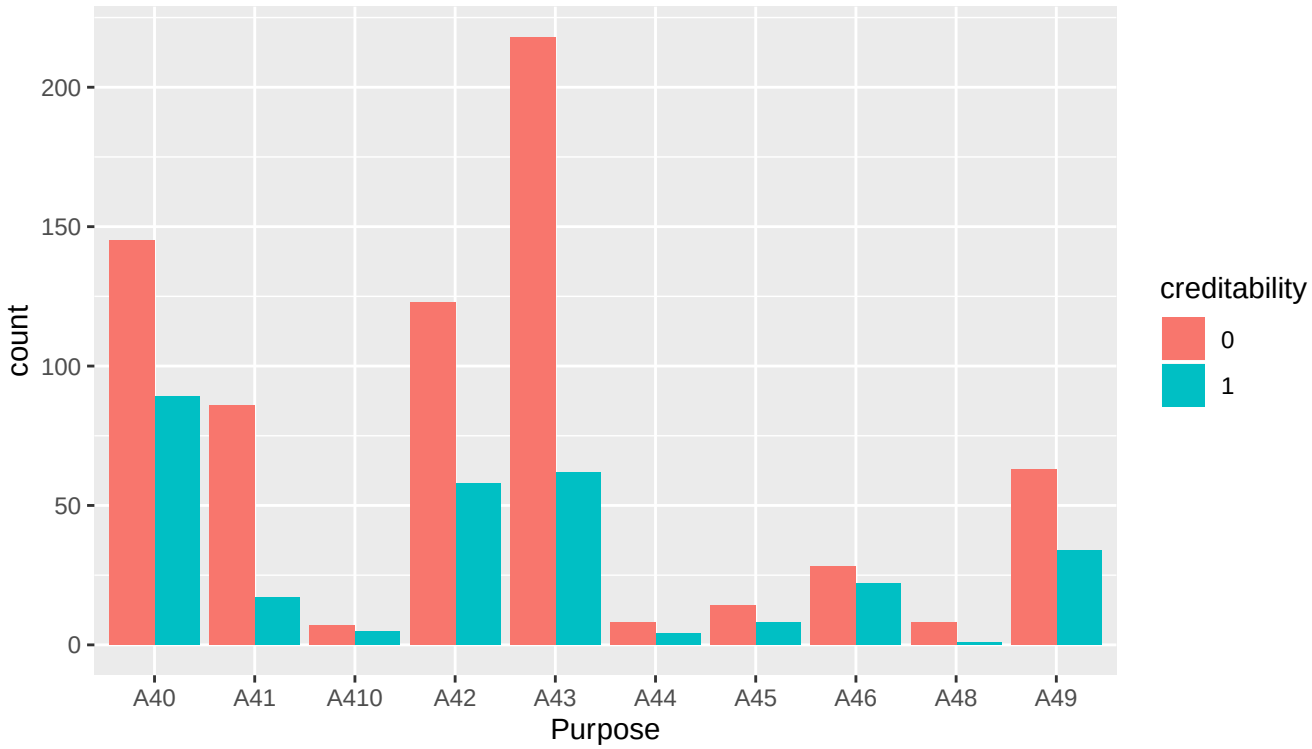
Na box-plocie z Rysunku 5 widzimy, że mediana wieku osób, które nie otrzymały kredytu jest niższa niż

osób, które go dostały. Najwięcej wnioskujących o kredyt osób jest między 25 a 35 rokiem życia.

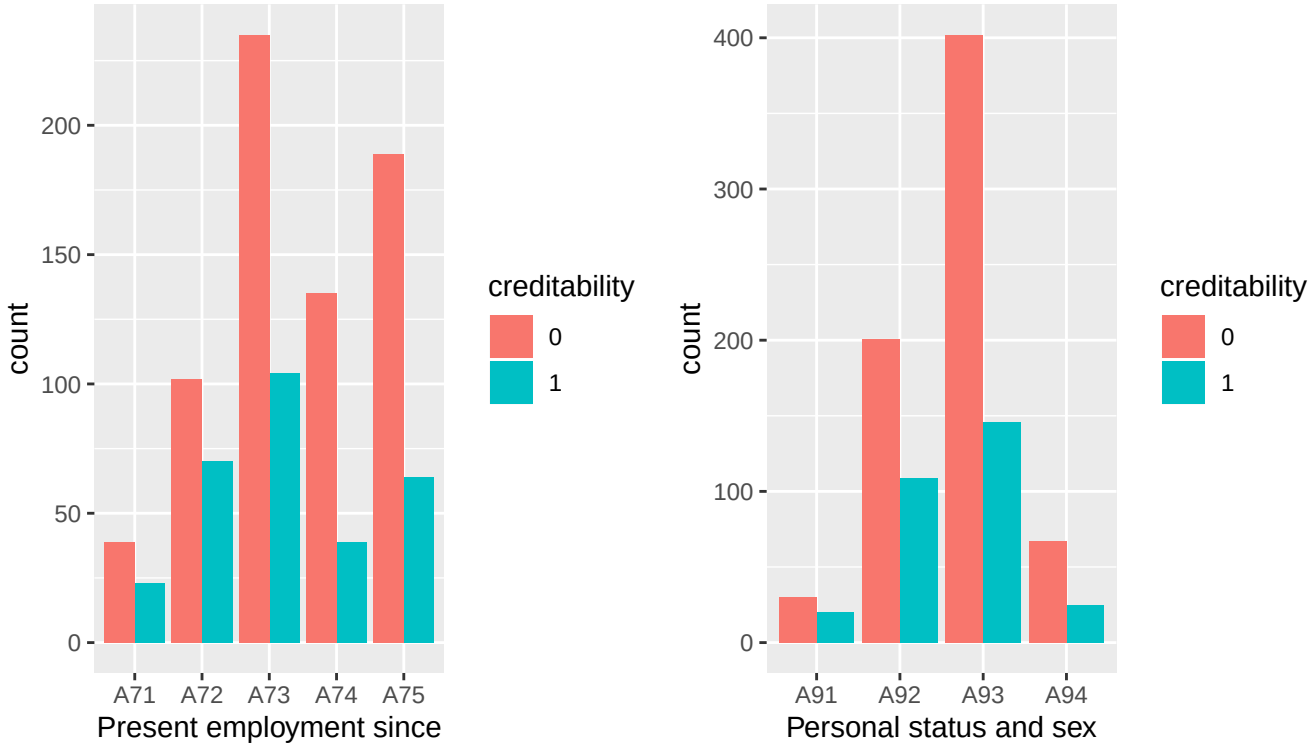


Rysunek 6: Histogramy dla zmiennych Number of existing credits at this bank i Number of people being liable to provide maintenance for

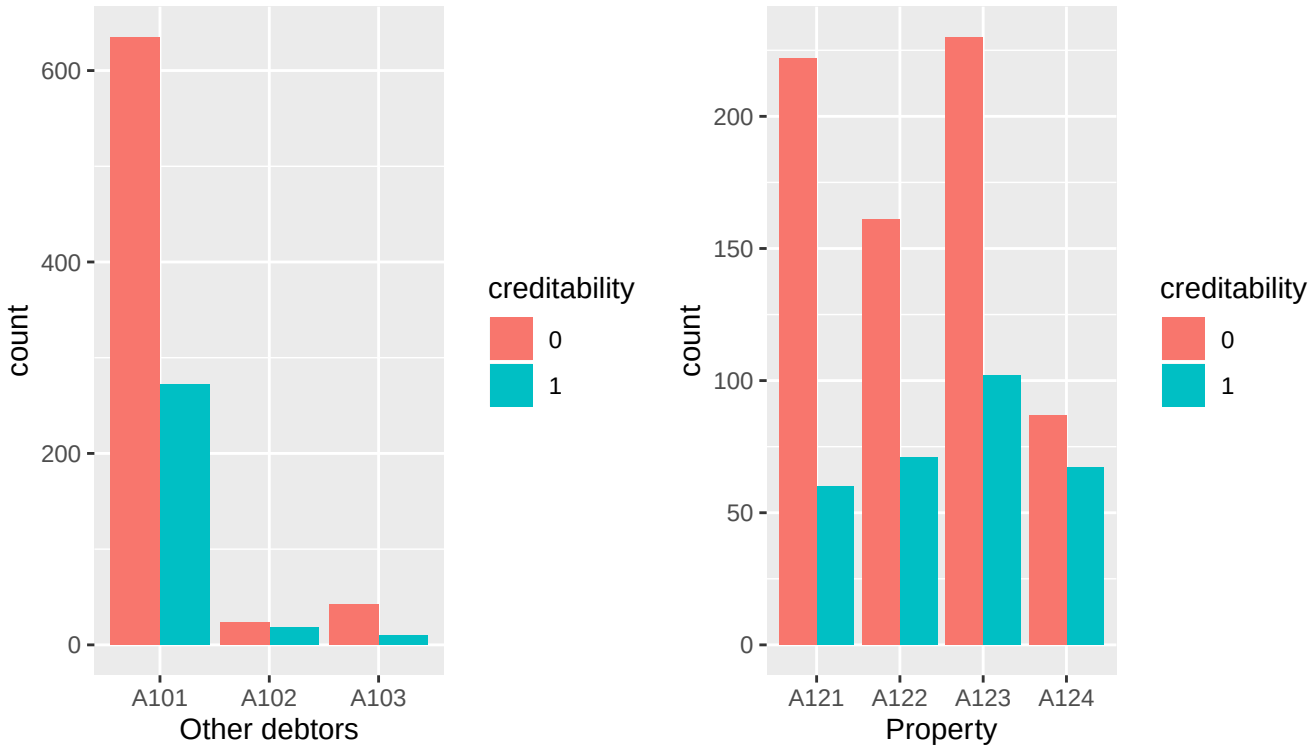
Rysunek 7: Wykresy słupkowe dla zmiennych Status of existing checking account i Credit history



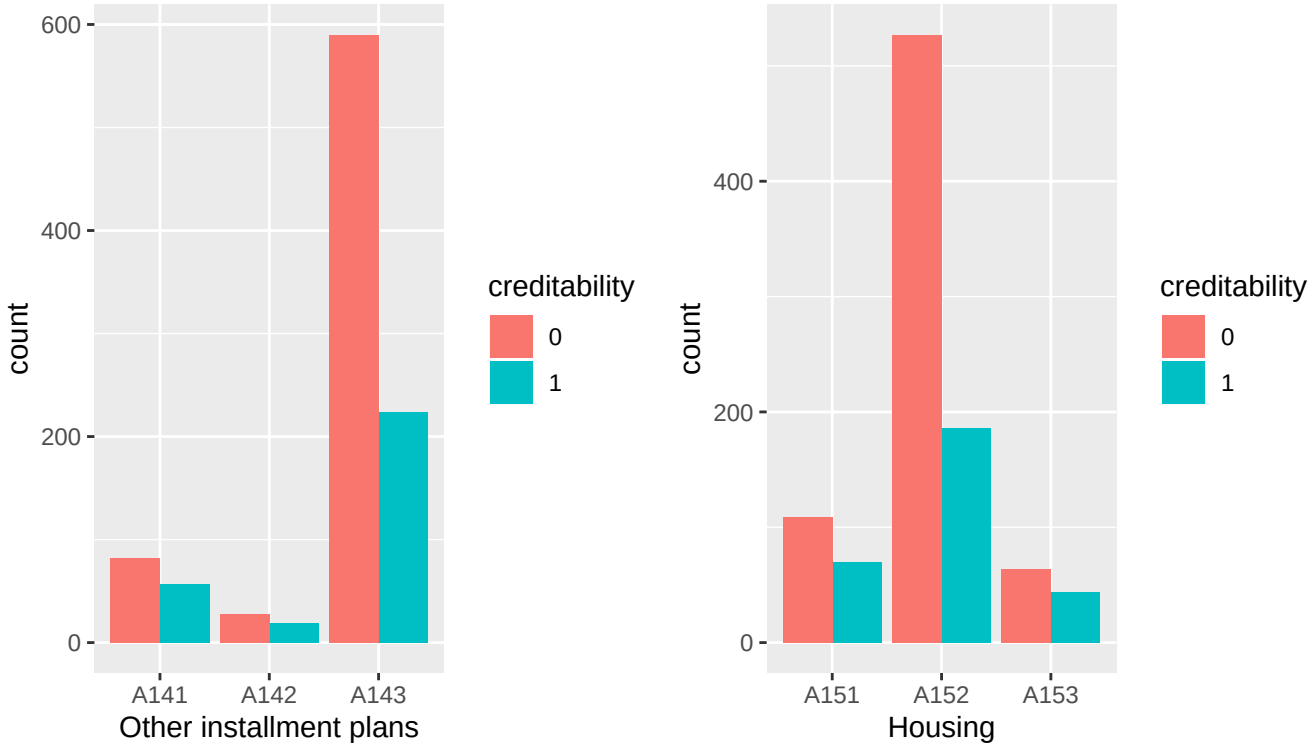
Rysunek 8: Wykres słupkowy dla zmiennej Purpose



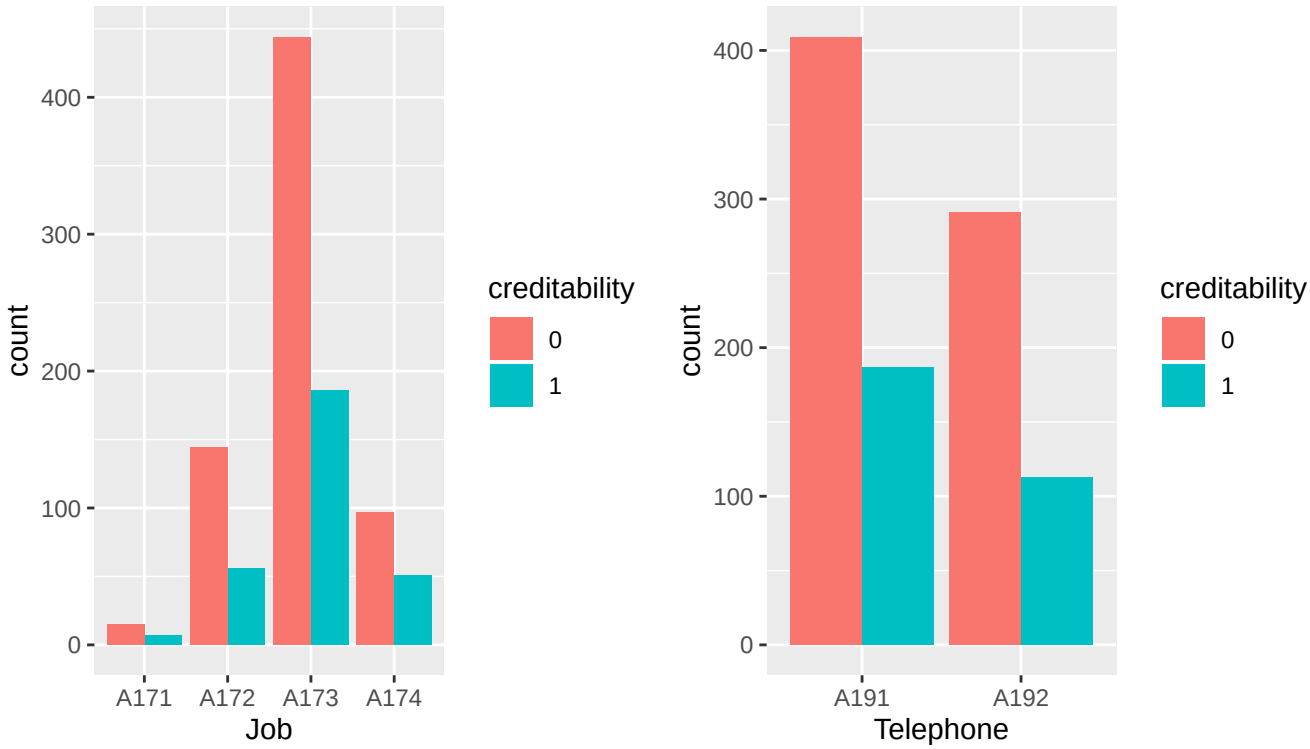
Rysunek 9: Wykresy słupkowe dla zmiennych Present employment since i Personal status and sex



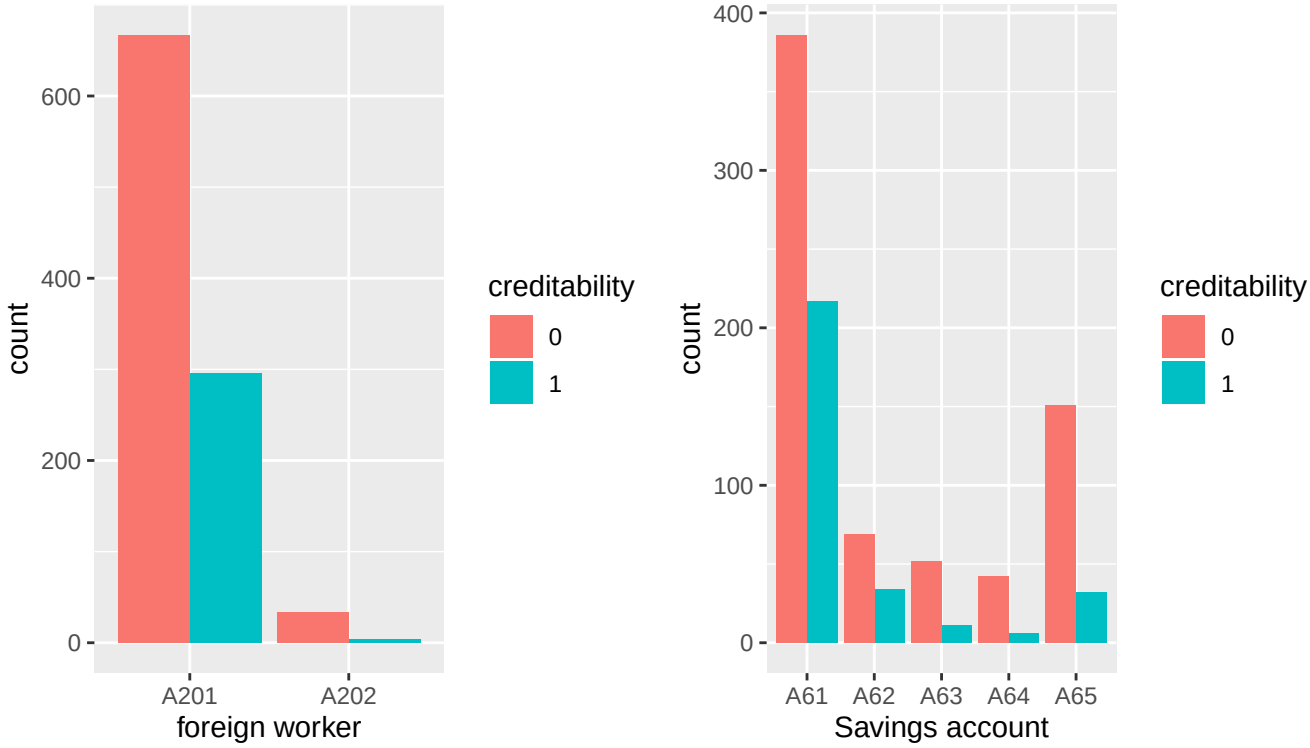
Rysunek 10: Wykresy słupkowe dla zmiennych Other debtors i Property



Rysunek 11: Wykresy słupkowe dla zmiennych Other installment plans i Housing

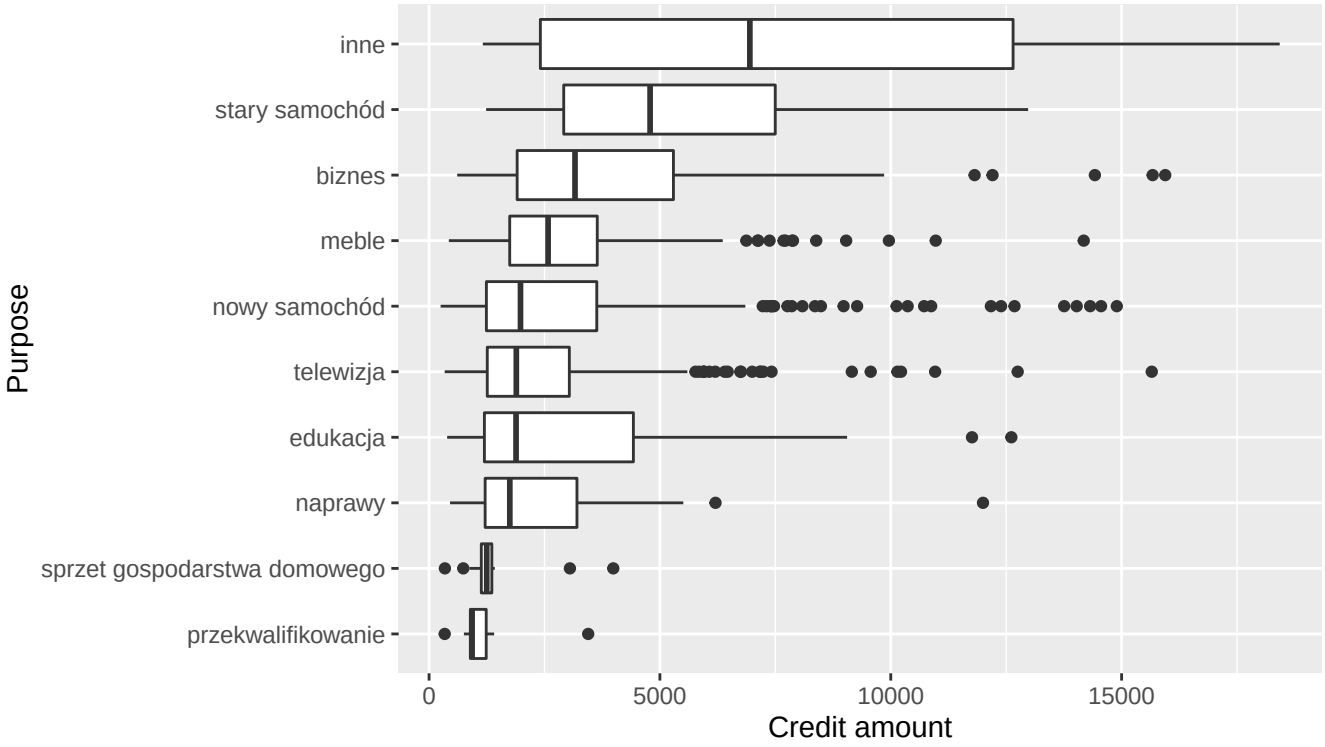


Rysunek 12: Wykresy słupkowe dla zmiennych Job i Telephone



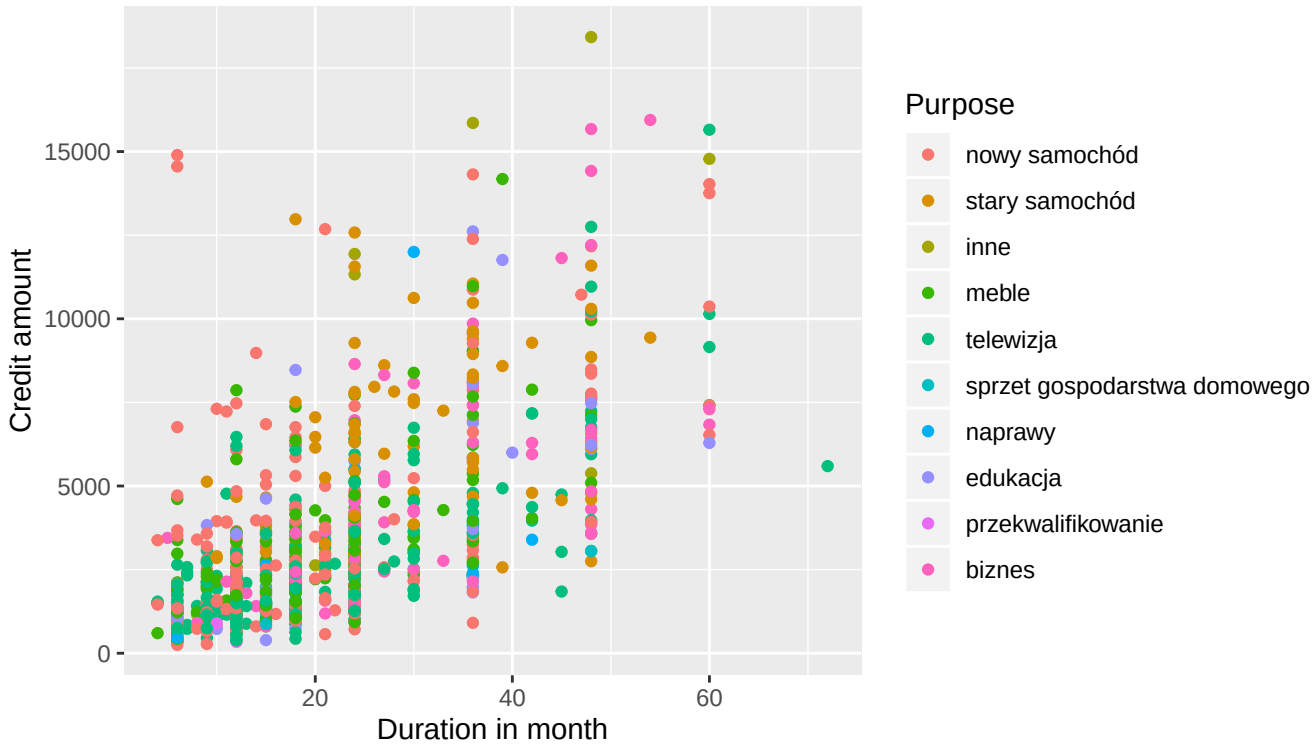
Rysunek 13: Wykresy słupkowe dla zmiennych foreign worker i Savings account

Rysunki 7-13 przedstawiają wykresy słupkowe dla zmiennych kategorycznych. Możemy z nich odczytać jak na danym poziomie rozkładają się obserwacje z klasy pozytywnej oraz negatywnej.



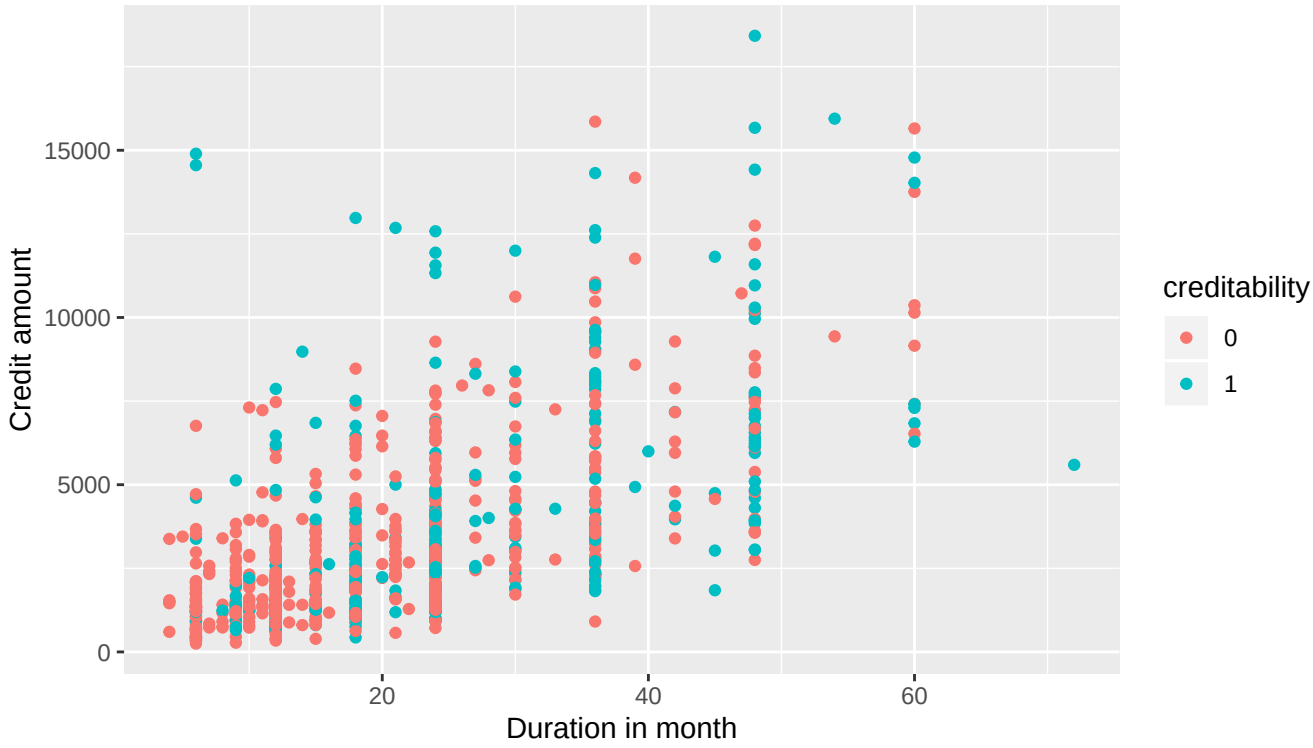
Rysunek 14: Rozkład kwoty kredytu w zależności od celu

Ciekawym pytaniem, na które możemy uzyskać odpowiedź analizując dane, jest np. pytanie ”Na co brane są najwyższe kredyty?”. Rysunek 14 wskazuje nam odpowiedź. Najwyższą medianę mamy w przypadku, gdy cel określony jest jako inny. Nie jest to zaskoczeniem, ponieważ w celu kredytu nie mamy zawartego m.in. zakupu mieszkania. Co zaskakujące, większą medianę kwoty kredytu mamy w przypadku starego samochodu, a nie jakbyśmy się spodziewali nowego. Można się również zastanawiać nad tym czy większe kredyty spłaca się dłużej. Na Rysunku 15 mozemy dostrzec pewną zależność między zmiennymi Credit amount oraz Duration in month. Widzimy także kilka obserwacji odstających od pozostałych. Dwóch klientów chciało otrzymać bardzo wysoki kredyt rozłożony na krótki okres czasu. Sprawdzimy, czy te osoby otrzymały kredyt czy nie.



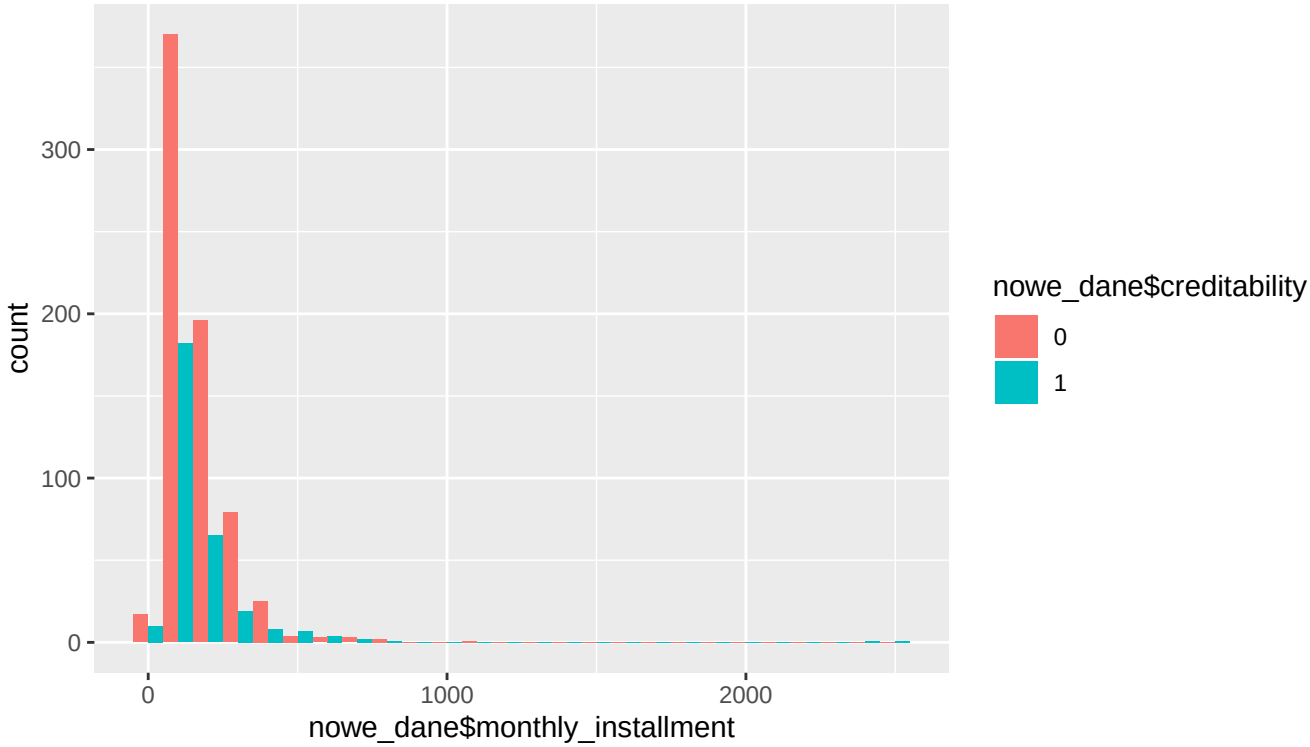
Rysunek 15: Zależność między kwotą kredytu a trwaniem kredytu wyrażonym w miesiącach, z podziałem na cel

Z Rysunku 16 możemy odczytać że osoby te nie otrzymały kredytu, zatem obserwacje te nie powinny wpływać na tworzone modele.



Rysunek 16: Zależność między kwotą kredytu a trwaniem kredytu wyrażonym w miesiącach, z podziałem na zdolność kredytową

Możemy także zastanowić się, czy na otrzymanie kredytu ma wpływ wysokość miesięcznej raty. Tworzymy w tym celu nową zmienną monthly_installment.



Rysunek 17: Histogram zmiennej monthly_installment

Rysunek 17 pokazuje, że osoby, które wnioskowały o kredyty z dużą miesieczną ratą, nie otrzymały kredytu. Można także zauważyć, że od pewnej wysokości raty, liczba osób które otrzymały kredyt i go nie otrzymały wyrównuje się.

	Duration in month	Credit amount	Installment rate in percentage of disposable income	Present residence since	Age in years	Number of existing credits at this bank	Number of people being liable to provide maintenance for
Duration in month	1.00	0.62	0.07	0.03	-0.04	-0.01	-0.02
Credit amount	0.62	1.00	-0.27	0.03	0.03	0.02	0.02
Installment rate in percentage of disposable income	0.07	-0.27	1.00	0.05	0.06	0.02	-0.07
Present residence since	0.03	0.03	0.05	1.00	0.27	0.09	0.04
Age in years	-0.04	0.03	0.06	0.27	1.00	0.15	0.12
Number of existing credits at this bank	-0.01	0.02	0.02	0.09	0.15	1.00	0.11
Number of people being liable to provide maintenance for	-0.02	0.02	-0.07	0.04	0.12	0.11	1.00

Tabela 6: Korelacje dla zmiennych ciągłych

W Tabeli 6 widzimy wartości korelacji dla zmiennych ciągłych. Większość zmiennych nie jest skorelowana. Wartość powyżej 0.5 mamy jedynie w przypadku Credit amount oraz Duration in month.

5 Modele

Do danych dopasujemy następujące modele: regresji logistycznej, regresji logistycznej ze stepem, regresji logistycznej z kategoryzacją, regresji logistycznej z wartościami WOE, liniowej analizy dyskryminacyjnej, kwadratowej analizy dyskryminacyjnej, drzew decyzyjnych i k najbliższych sąsiadów. W celu uzyskania wiarygodniejszych wyników, dane podzielimy na dwa zbiory: treningowy i testowy. Najpierw sprawdzimy działanie modeli na zbiorze treningowym, przeprowadzając w tym celu 5-krotną walidację krzyżową. Następnie, najlepsze modele zbudujemy dla całego zbioru treningowego i porównamy wyniki na zbiorze testowym. Ponieważ analizowane dane mają nierówny rozkład klas, zastosujemy tzw. stratified sampling. W każdej iteracji walidacji wyznaczymy wskaźniki modelu takie jak: dokładność(ACC), czułość(TPR), specyficzność(TNR), F1 oraz średni koszt złej klasyfikacji(MMC). Następnie podamy średnie tych wskaźników w danym modelu. Ponadto, zmodyfikujemy te modele, aby stały się "cost-sensitive" za pomocą zmiany thresholdu. Wartość 0.5 powyżej której do tej pory klasyfikowaliśmy do klasy 1 zamienimy na $p^* = \frac{C(1,0)}{C(1,0)+C(0,1)}$, gdzie $C(1,0)$ oznacza koszt zaklasyfikowania dobrego klienta jako złego, a $C(0,1)$ złego jako dobrego. W przypadku naszych danych macierz kosztów wygląda następująco:

	actual good	actual bad
predicted good	0.00	5.00
predicted bad	1.00	0.00

Tabela 7: Macierz kosztów dla danych German Credit

Zakładając, że dysponujemy macierzą pomyłek:

	actual good	actual bad
predicted good	TP	FN
predicted bad	FP	TN

Tabela 8: Macierz pomyłek

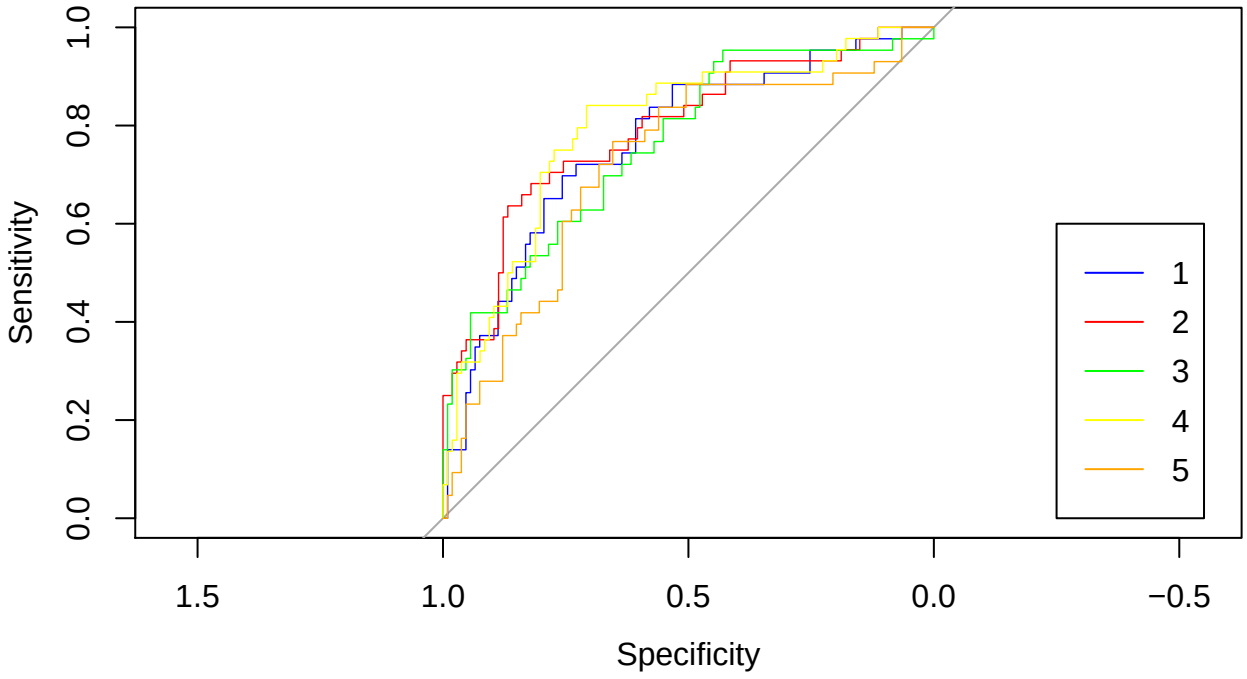
pozostałe wskaźniki możemy obliczyć korzystając ze wzorów:

- $ACC = \frac{TP+TN}{TN+TP+FP+FN}$
- $TPR = \frac{TP}{TP+FN}$,
- $TNR = \frac{TN}{TN+FP}$,
- $F_1 = \frac{2 \cdot TP}{2 \cdot TP+FP+FN}$,
- $MMC = \frac{FP \cdot C(0,1)+FN \cdot C(1,0)}{TP+TN+FP+FN}$.

Zaznaczmy, że w przypadku cost-sensitive learning uzyskanie wysokiej dokładności nie jest najważniejsze. Bardziej pożądane jest uzyskanie wysokiego F_1, a w przypadku tak rozłożonych kosztów jak u nas także wysokiej czułości. Przypomnijmy, że im F_1 jest bliższe 1 tym model jest lepszy. Ważne jest również uzyskanie jak najmniejszych kosztów, czyli minimalizacja wskaźnika MMC.

5.1 Model regresji logistycznej

Pierwszy model, który został zbudowany oparty jest na regresji logistycznej. Do jego utworzenia wykorzystujemy wszystkie zmienne objaśniające zawarte w zbiorze danych. Następnie wyliczamy wskaźniki dla każdej z iteracji walidacji.



Rysunek 18: Krzywe ROC uzyskane dla każdej walidacji (Regresja logistyczna)

Rysunek 18 prezentuje krzywe ROC uzyskane w każdej z iteracji dla modelu, w którym punkt odcięcia ustalony jest jako 0.5.

	ACC	TPR	TNR	F1	MMC
1	0.75	0.51	0.85	0.23	0.81
2	0.80	0.61	0.88	0.20	0.65
3	0.73	0.53	0.81	0.28	0.80
4	0.76	0.43	0.90	0.17	0.91
5	0.72	0.33	0.88	0.18	1.05
średnia	0.75	0.48	0.86	0.21	0.84

Tabela 9: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej

	ACC	TPR	TNR	F1	MMC
1	0.65	0.81	0.58	0.55	0.57
2	0.63	0.82	0.56	0.58	0.58
3	0.61	0.81	0.52	0.60	0.61
4	0.67	0.84	0.59	0.55	0.52
5	0.63	0.79	0.57	0.56	0.61
średnia	0.64	0.82	0.56	0.57	0.58

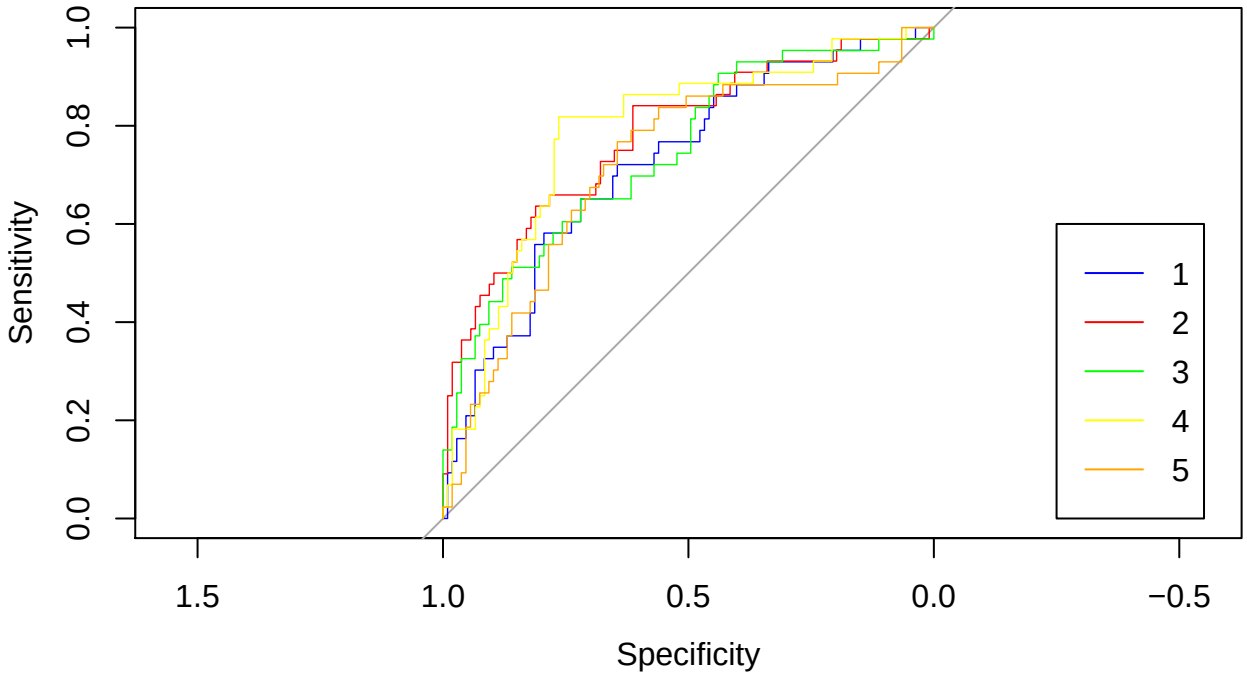
Tabela 10: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z p=1/6

Tabele 9-10 przedstawiają wartości wskaźników takich jak dokładność, czułość, specyficzność, współczynnik F_1 oraz MMC. Pierwsza z tabel dotyczy modelu, w którym nie uwzględniono różnych kosztów popełnianych

błędów, tzn. punkt odcięcia wynosi 0.5. Dokładność uzyskana przy tak ustalonym p jest stosunkowo wysoka. Wynosi średnio 0.75. Jednak jak już wcześniej wspomnieliśmy, ważniejszym wskaźnikiem jest m.in. czułość, która wynosi średnio 0.48. Otrzymaliśmy również dość niską wartość F_1. Średnia wynosi 0.21 (dla modelu losowego wynosi 0). Obserwujemy także wysokie koszty złej klasyfikacji. Sytuacja zmiena się jeśli zmienimy p z 0.5 na 1/6. Wówczas, jak widzimy w drugiej z tabel, dokładność nieco spada, ale za to znacząco rośnie czułość oraz wartość F_1. Zauważamy także znaczną redukcję kosztów z 0.84 do 0.58.

5.2 Regresja logistyczna ze stepem

Kolejnym z rozpatrywanych modeli jest model regresji logistycznej, w której dobór zmiennych następuje z użyciem metody krokowej (funkcja *step*). Wybierane są te zmienne dla których wartość AIC jest najmniejsza i na nich budowany jest model.



Rysunek 19: Krzywe ROC uzyskane dla każdej walidacji (Regresja logistyczna ze stepem)

Rysunek 19 prezentuje krzywe ROC uzyskane w każdej z iteracji dla modelu, w którym punkt odcięcia ustalony jest jako 0.5.

	ACC	TPR	TNR	F1	MMC
1	0.71	0.37	0.84	0.22	1.01
2	0.78	0.48	0.91	0.16	0.83
3	0.73	0.51	0.82	0.26	0.83
4	0.75	0.43	0.89	0.18	0.91
5	0.72	0.35	0.87	0.19	1.03
średnia	0.74	0.43	0.87	0.20	0.92

Tabela 11: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej ze stepem

	ACC	TPR	TNR	F1	MMC
1	0.61	0.77	0.54	0.57	0.66
2	0.64	0.84	0.56	0.58	0.55
3	0.57	0.84	0.46	0.66	0.62
4	0.69	0.86	0.61	0.53	0.47
5	0.63	0.79	0.57	0.56	0.61
średnia	0.63	0.82	0.55	0.58	0.58

Tabela 12: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej ze stepem z p=1/6

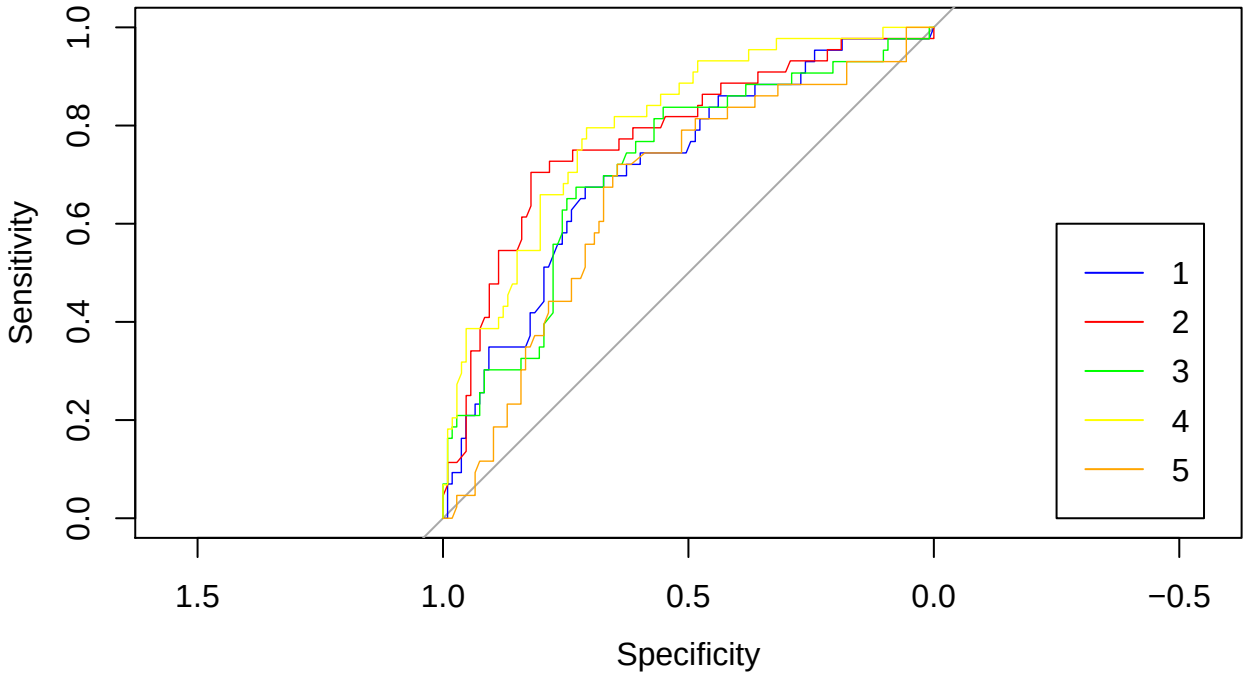
Tabele 11-12 zawierają wartości wskaźników uzyskane w 5-krotnej walidacji. Podobnie jak w przypadku modelu regresji logistycznej nieuwzględniającym metody krokowej bardziej satysfakcjonujące wyniki otrzymaliśmy w przypadku, gdy threshold został zmieniony na 1/6. Otrzymana wówczas średnia czułość jest wyższa niż dla p=0.5.

5.3 Model regresji logistycznej z kategoryzacją

Do stworzenia kolejnego modelu po raz kolejny wykorzystamy regresję logistyczną z metodą krokową. Tym razem do wybrania zmiennych posłużymy się tzw. Information Value, zdefiniowane następującym wzorem:

$$IV = \sum_i \ln \frac{p_G(x_i)}{p_B(x_i)}(p_G(x_i) - p_B(x_i)),$$

gdzie p_B i p_G są odpowiednio gęstościami rozkładów $X|B$ i $X|G$, przy czym $X|B$ i $X|G$ oznaczają kolejno zmienne losowe w grupie złych i dobrych klientów. W przypadku, gdy dla pewnej zmiennej dyskretnej X, $IV \leqslant 0.02$ to predyktor X należy pominąć. Gdy $IV \in (0.02, 0.1]$ to X ma słabą moc predykcyjną. Gdy $IV \in (0.1, 0.3]$ to X ma średnią moc predykcyjną. Jeżeli $IV > 0.5$ to powinniśmy skontrolować predyktor X. Zgodnie z powyższym usuwamy zmienne, dla których otrzymujemy $IV \leqslant 0.02$.



Rysunek 20: Krzywe ROC uzyskane dla każdej walidacji (Regresja logistyczna z kategoryzacją)

	ACC	TPR	TNR	F1	MMC
1	0.71	0.35	0.86	0.20	1.03
2	0.78	0.52	0.89	0.19	0.78
3	0.70	0.30	0.86	0.19	1.10
4	0.74	0.39	0.89	0.18	0.98
5	0.69	0.35	0.83	0.23	1.05
średnia	0.73	0.38	0.86	0.20	0.99

Tabela 13: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z kategoryzacją

	ACC	TPR	TNR	F1	MMC
1	0.60	0.74	0.54	0.57	0.69
2	0.59	0.82	0.50	0.62	0.62
3	0.62	0.84	0.53	0.60	0.57
4	0.65	0.86	0.56	0.59	0.51
5	0.59	0.77	0.51	0.60	0.68
średnia	0.61	0.81	0.53	0.59	0.61

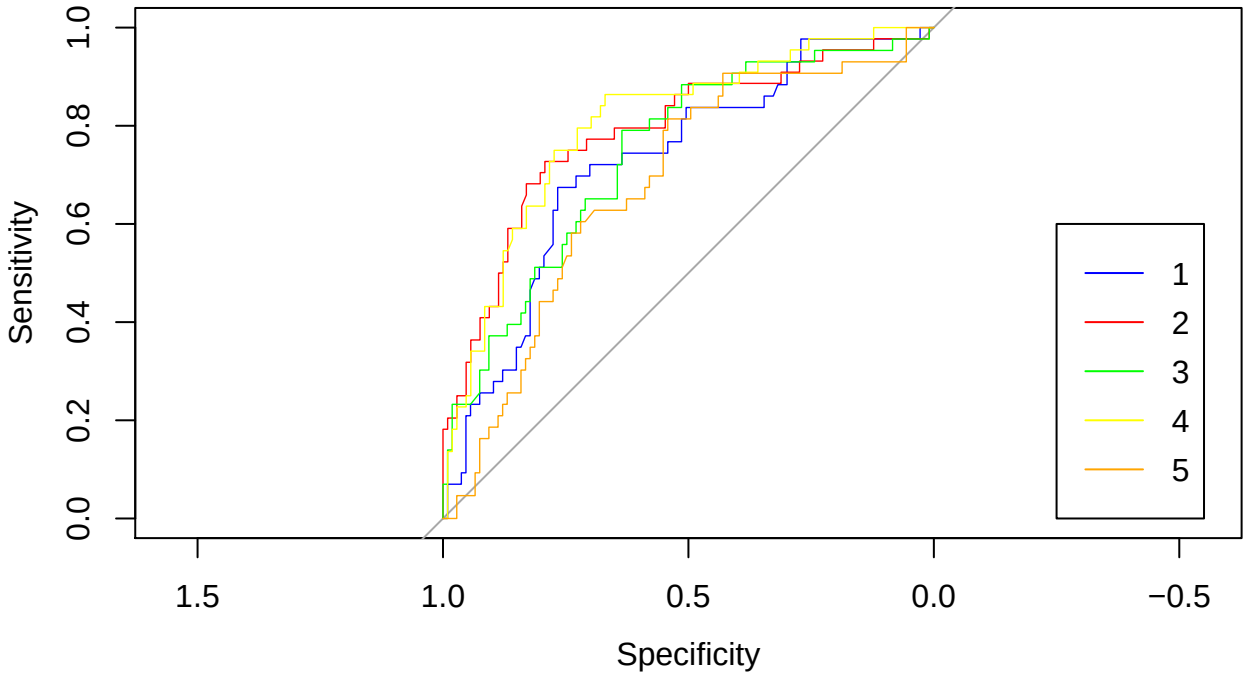
Tabela 14: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z kategoryzacją z p=1/6

5.4 Model regresji logistycznej z wartościami WoE

Kolejny model, który zbudujemy również jest oparty na regresji logistycznej z metodą krokową. Jednak tym razem po zastosowaniu kryterium IV, wartości zmiennych zastąpimy odpowiadającymi im wartościami WoE,

zdefiniowanej jako

$$WoE(x_i) = \ln \frac{p_G(x_i)}{p_B(x_i)}.$$



Rysunek 21: Krzywe ROC uzyskane dla każdej walidacji (Regresja logistyczna z wartościami WOE)

	ACC	TPR	TNR	F1	MMC
1	0.71	0.28	0.88	0.17	1.12
2	0.77	0.50	0.89	0.18	0.81
3	0.72	0.40	0.85	0.22	0.97
4	0.77	0.43	0.92	0.14	0.89
5	0.69	0.26	0.86	0.18	1.17
średnia	0.73	0.37	0.88	0.18	0.99

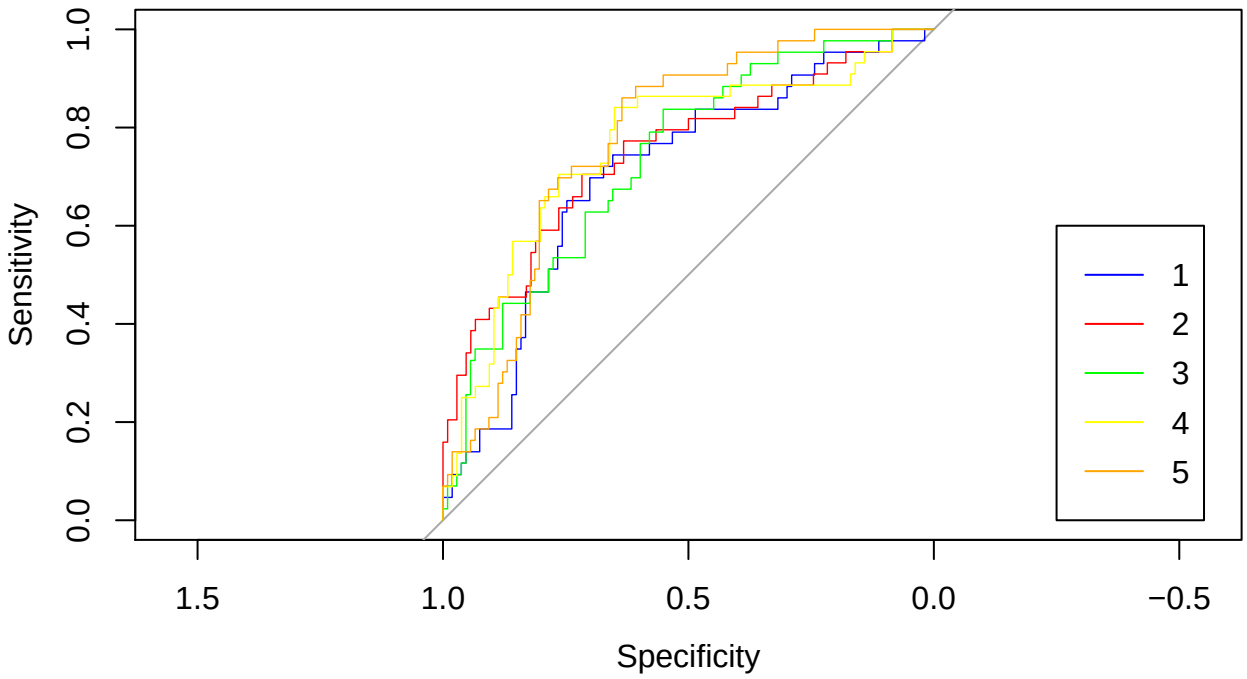
Tabela 15: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z WOE

	ACC	TPR	TNR	F1	MMC
1	0.62	0.74	0.57	0.54	0.67
2	0.61	0.89	0.50	0.64	0.52
3	0.61	0.86	0.51	0.62	0.55
4	0.65	0.86	0.57	0.58	0.51
5	0.61	0.81	0.53	0.59	0.60
średnia	0.62	0.83	0.54	0.60	0.57

Tabela 16: Tabela wskaźników dla 5-fold cross-validation w modelu regresji logistycznej z WOE z p=1/6

5.5 Liniowa analiza dyskryminacyjna

Następny model zbudujemy w oparciu o liniową analizą dyskryminacyjną. Ideą tej metody jest wyznaczenie hiperpłaszczyzny „najlepiej" separującej klasy. Granica decyzyjna między klasami 0 i 1 jest liniową funkcją wektora cech niezależnych $\mathbf{x}$. LDA możemy zastosować wyłącznie do zmiennych ilościowych, więc pominiemy inne zmienne przy tworzeniu modelu.



Rysunek 22: Krzywe ROC uzyskane dla każdej walidacji (LDA)

	ACC	TPR	TNR	F1	MMC
1	0.71	0.35	0.85	0.21	1.04
2	0.76	0.41	0.91	0.15	0.93
3	0.70	0.49	0.79	0.30	0.89
4	0.73	0.30	0.91	0.14	1.10
5	0.71	0.26	0.89	0.16	1.15
średnia	0.72	0.36	0.87	0.19	1.02

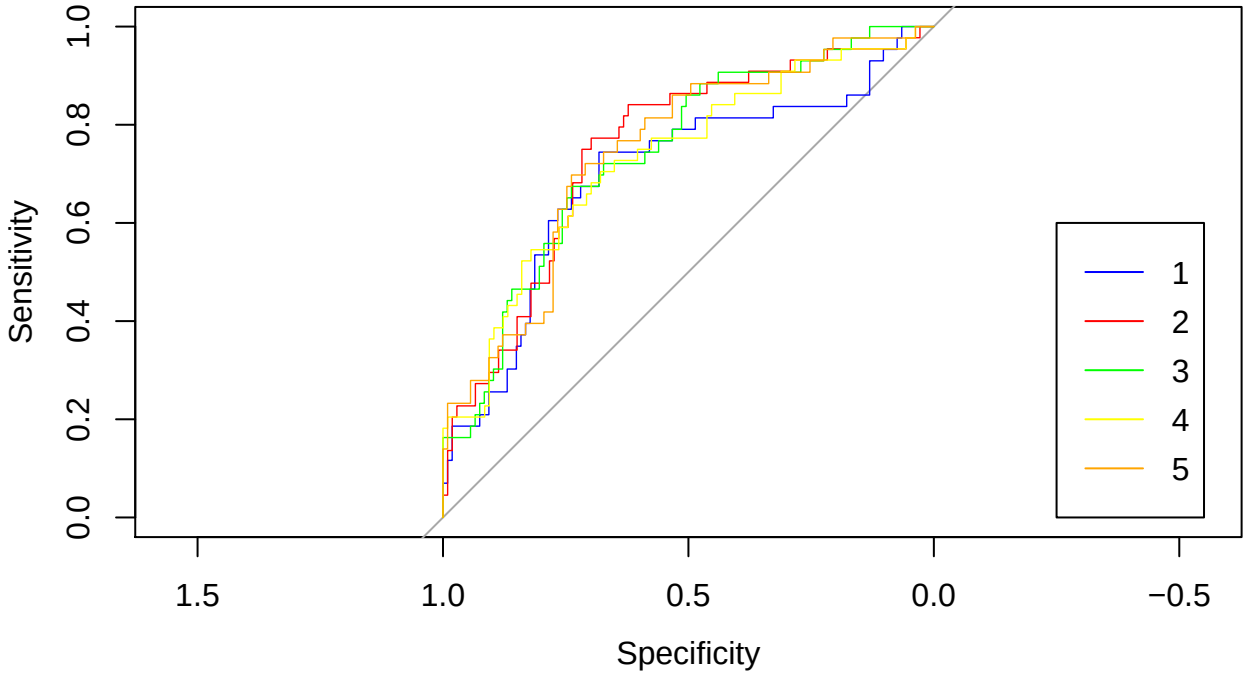
Tabela 17: Tabela wskaźników dla 5-fold cross-validation w LDA

	ACC	TPR	TNR	F1	MMC
1	0.58	0.81	0.49	0.63	0.63
2	0.58	0.82	0.48	0.64	0.63
3	0.56	0.88	0.43	0.69	0.57
4	0.64	0.86	0.55	0.59	0.52
5	0.60	0.91	0.48	0.66	0.51
średnia	0.59	0.86	0.48	0.64	0.57

Tabela 18: Tabela wskaźników dla 5-fold cross-validation w LDA z p=1/6

5.6 Kwadratowa analiza dyskryminacyjna

Kolejny model zbudujemy w oparciu o kwadratową analizą dyskryminacyjną, która jest modyfikacją wykorzystywanej wcześniej liniowej analizy dyskryminacyjnej. Ideą tej metody jest również wyznaczenie hiperpłaszczyzny „najlepiej" separującej klasy. Jednak granica decyzyjna jest w tym przypadku wielomianem drugiego rzędu wektora cech niezależnych $\mathbf{x}$. Podobnie jak LDA, QDA możemy zastosować wyłącznie do zmiennych ilościowych, więc pominiemy inne zmienne przy tworzeniu modelu.



Rysunek 23: Krzywe ROC uzyskane dla każdej walidacji (QDA)

	ACC	TPR	TNR	F1	MMC
1	0.71	0.63	0.74	0.37	0.72
2	0.72	0.68	0.74	0.38	0.65
3	0.69	0.67	0.69	0.42	0.69
4	0.69	0.64	0.71	0.40	0.74
5	0.72	0.65	0.75	0.36	0.68
średnia	0.70	0.65	0.72	0.39	0.70

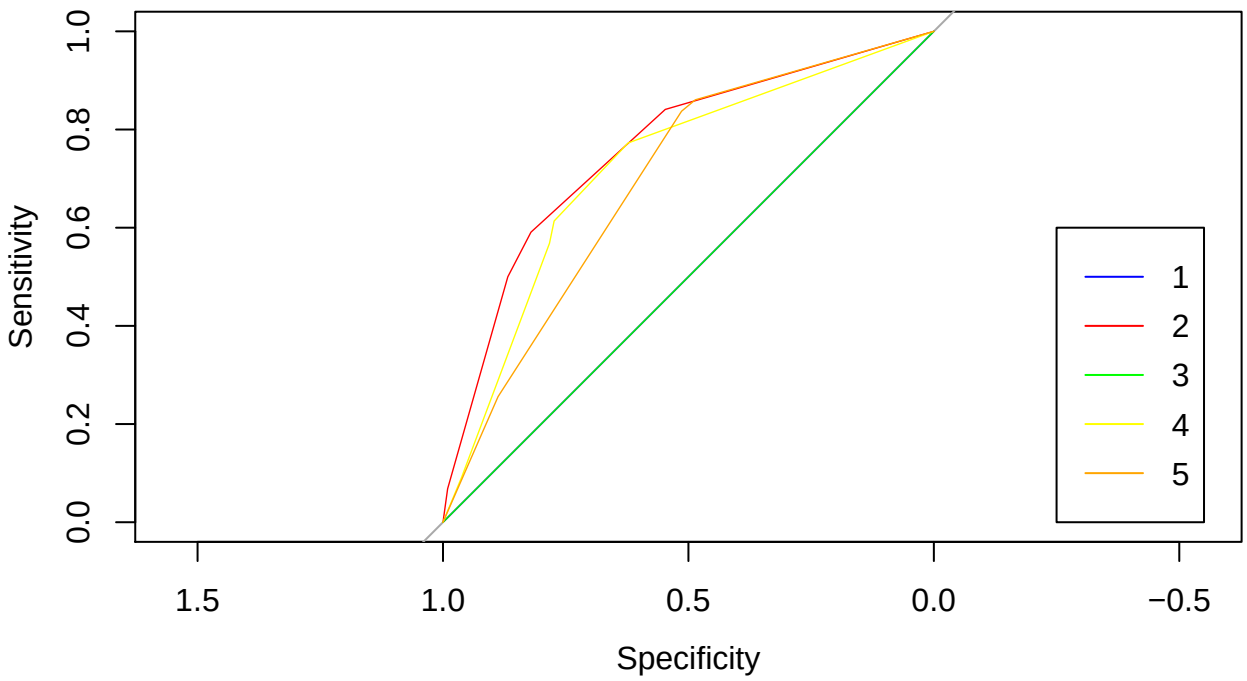
Tabela 19: Tabela wskaźników dla 5-fold cross-validation w modelu QDA

	ACC	TPR	TNR	F1	MMC
1	0.70	0.74	0.68	0.45	0.59
2	0.67	0.84	0.60	0.54	0.51
3	0.59	0.79	0.51	0.60	0.65
4	0.65	0.73	0.61	0.51	0.67
5	0.64	0.81	0.57	0.56	0.57
średnia	0.65	0.78	0.60	0.53	0.60

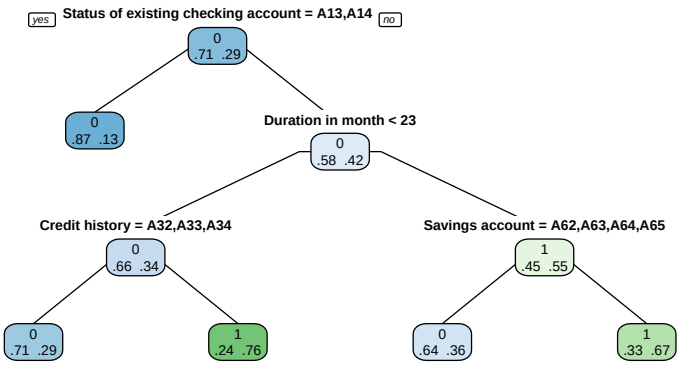
Tabela 20: Tabela wskaźników dla 5-fold cross-validation w modelu QDA z p=1/6

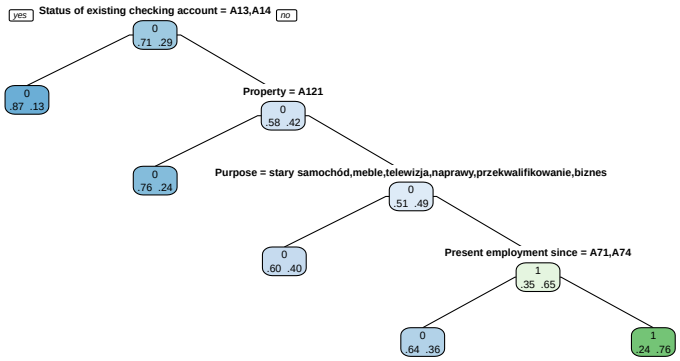
5.7 Drzewa decyzyjne

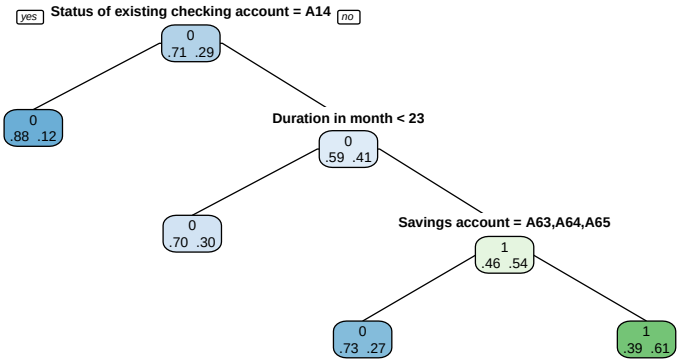
Kolejnym modelem jest drzewo decyzyjne. Do wyboru drzewa w danej iteracji walidacji skorzystamy z metody pruningu opartej na kryterium kosztu złożności. Optymalne drzewo wybierzemy na podstawie reguły 1SE, tzn. naszym wyborem będzie drzewo o najmniejszym rozmiarze, dla którego ułamek błędnych klasyfikacji jest odległy o nie więcej niż jedno odchylenie standardowe od minimum ułamka błędnych klasyfikacji.



Rysunek 24: Krzywe ROC uzyskane dla każdej walidacji (Drzewa decyzyjne)







Rysunek 25: Drzewa decyzyjne dla każdej walidacji

	ACC	TPR	TNR	F1	MMC
1	0.71	0.00	1.00		1.43
2	0.76	0.50	0.87	0.21	0.83
3	0.71	0.00	1.00		1.43
4	0.71	0.09	0.96	0.05	1.36
5	0.71	0.26	0.89	0.16	1.15
średnia	0.72	0.17	0.94		1.24

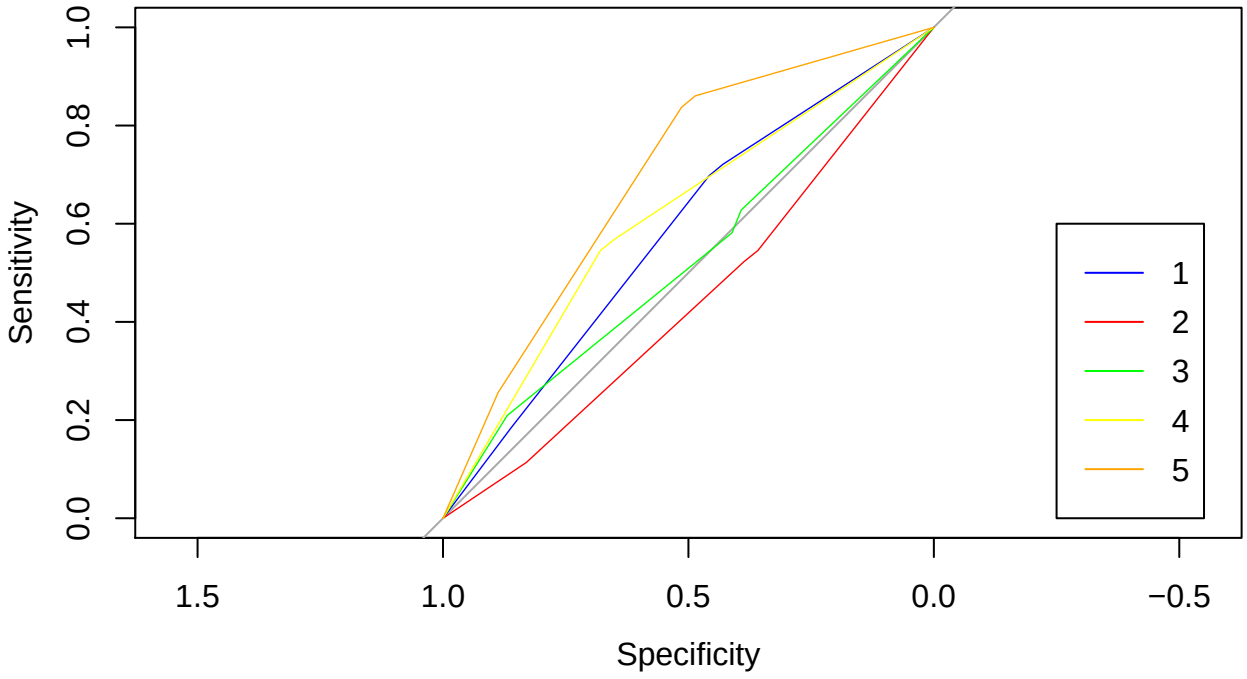
Tabela 21: Tabela wskaźników dla 5-fold cross-validation dla drzew decyzyjnych

	ACC	TPR	TNR	F1	MMC
1	0.29	1.00	0.00	1.00	0.71
2	0.63	0.84	0.55	0.59	0.55
3	0.29	1.00	0.00	1.00	0.71
4	0.67	0.77	0.62	0.51	0.60
5	0.59	0.86	0.49	0.64	0.57
średnia	0.49	0.89	0.33	0.75	0.63

Tabela 22: Tabela wskaźników dla 5-fold cross-validation dla drzew decyzyjnych z p=1/6

5.8 kNN zmienne ciągłe

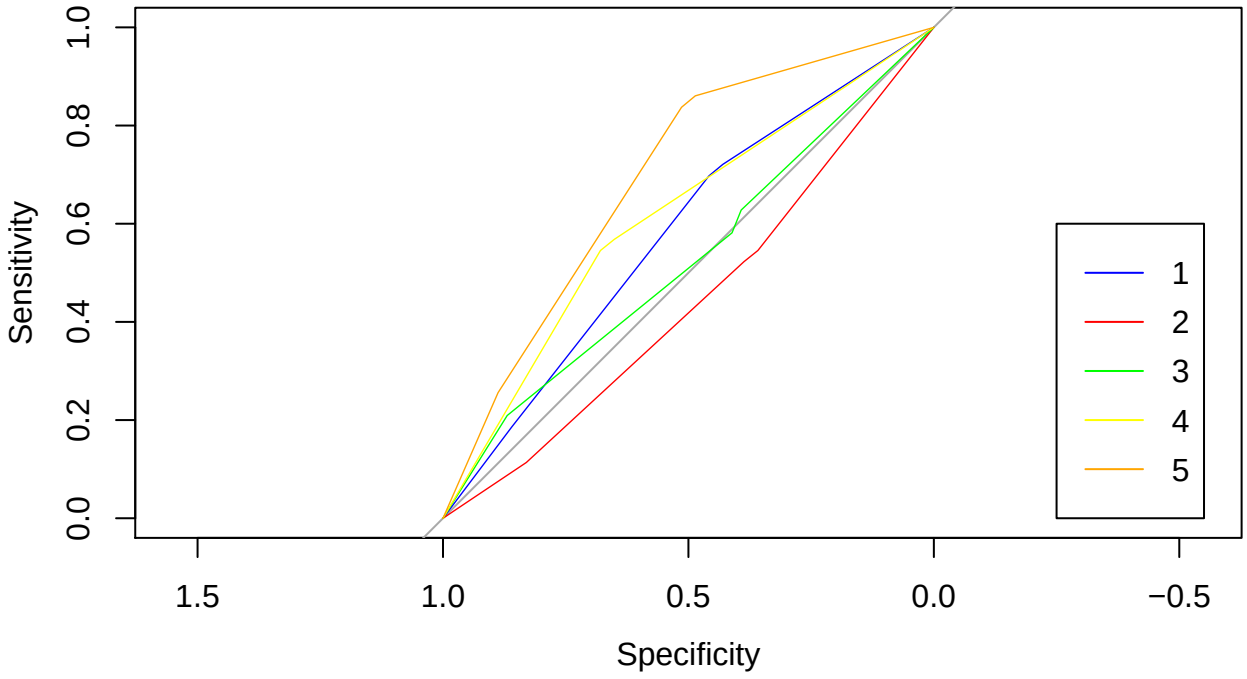
Skorzystamy teraz z metody k- najbliższych sąsiadów. Do stworzenia modelu użyjemy jedynie zmiennych typu integer. Robimy tak dlatego, że algorytm kNN może nie działać prawidłowo dla zmiennych kategorycznych nieuporządkowanych. Zmienne które wykorzystujemy do utworzenia modelu na początku normalizujemy, tzn. odejmujemy minimium i dzielimy przez różnicę między maksimum i minimum. W tworzonych przez nas modelach będziemy rozpatrywać różne wartości k. W związku z tym, że model budujemy jedynie dla 7 zmiennych objaśniających, k będzie przyjmowało wartości 1, 3 oraz 5.



Rysunek 26: Krzywe ROC uzyskane dla każdej walidacji (kNN(1))

	ACC	TPR	TNR	F1	MMC
1	0.65	0.33	0.78	0.27	1.13
2	0.62	0.32	0.75	0.28	1.18
3	0.59	0.26	0.72	0.27	1.27
4	0.66	0.43	0.75	0.31	1.01
5	0.65	0.28	0.79	0.24	1.18
średnia	0.63	0.32	0.76	0.27	1.15

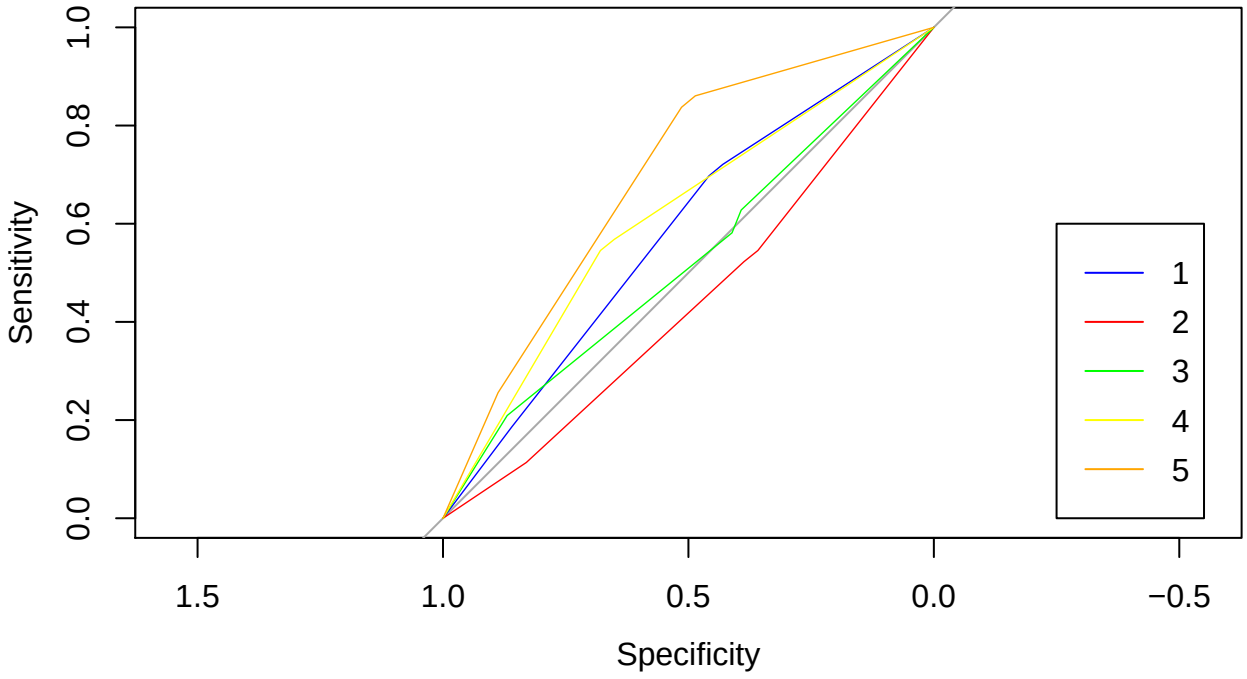
Tabela 23: Tabela wskaźników dla 5-fold cross-validation w modelu kNN(1)



Rysunek 27: Krzywe ROC uzyskane dla każdej walidacji kNN(3)

	ACC	TPR	TNR	F1	MMC
1	0.67	0.16	0.87	0.15	1.29
2	0.63	0.11	0.85	0.13	1.41
3	0.65	0.19	0.83	0.18	1.29
4	0.65	0.16	0.86	0.15	1.33
5	0.68	0.19	0.88	0.15	1.25
średnia	0.66	0.16	0.86	0.15	1.31

Tabela 24: Tabela wskaźników dla 5-fold cross-validation w modelu (kNN(3))



Rysunek 28: Krzywe ROC uzyskane dla każdej walidacji kNN(5)

	ACC	TPR	TNR	F1	MMC
1	0.65	0.05	0.90	0.06	1.44
2	0.67	0.18	0.87	0.15	1.29
3	0.65	0.09	0.88	0.11	1.39
4	0.69	0.16	0.92	0.11	1.29
5	0.71	0.19	0.93	0.11	1.22
średnia	0.68	0.13	0.90	0.11	1.33

Tabela 25: Tabela wskaźników dla 5-fold cross-validation w modelu (kNN(5))

Na podstawie uzyskanych wyników wnioskujemy, że model uzyskany z użyciem metody k- najbliższych sąsiadów nie jest dobry dla naszych danych. Otrzymujemy bardzo niskie wartości dla czułości oraz F_1. Mamy także wysokie koszty błędnej klasyfikacji. Porównując rezultaty uzyskane dla k=1, k=3 oraz k=5 okazuje się, że najlepszy jest model w którym narzuciliśmy k=1. Może to być spowodowane małą ilością zmiennych objaśniających.

6 Porównanie modeli

Tabela 26 zawiera porównanie średnich wartości wskaźników dla wszystkich zbudowanych modeli.

	ACC	TPR	TNR	F1	MMC
LR	0.75	0.48	0.86	0.21	0.84
LR(step)	0.74	0.43	0.87	0.20	0.92
LR(category)	0.73	0.38	0.86	0.20	0.99
LR(WoE)	0.73	0.37	0.88	0.18	0.99
LDA	0.72	0.36	0.87	0.19	1.02
QDA	0.70	0.65	0.72	0.39	0.70
Drzewa decyzyjne	0.72	0.17	0.94		1.24
kNN(1)	0.63	0.32	0.76	0.27	1.15
kNN(3)	0.66	0.16	0.86	0.15	1.31
kNN(5)	0.68	0.13	0.90	0.11	1.33
LR p=1/6	0.64	0.82	0.56	0.57	0.58
LR(step) p=1/6	0.63	0.82	0.55	0.58	0.58
LR(category) p=1/6	0.61	0.81	0.53	0.59	0.61
LR(WoE) p=1/6	0.62	0.83	0.54	0.60	0.57
LDA p=1/6	0.59	0.86	0.48	0.64	0.57
QDA p=1/6	0.65	0.78	0.60	0.53	0.60
Drzewa decyzyjne p=1/6	0.49	0.89	0.33	0.75	0.63

Tabela 26: Tabela średnich wskaźników dla wszystkich modeli

Widzimy, że większość modeli 'cost-insensitive' ma większą dokładność od modeli 'cost-sensitive'. Jednak dla wszystkich modeli 'cost-insensitive' średni koszt złej klasyfikacji jest wyższy niż dla modeli z thresholdem zmienionym na 1/6. W dalszej analizie uwzględnimy: regresję logistyczną, regresję logistyczną z kategoryzacją, QDA, regresję logistyczną ze stepem i p=1/6, LDA z p=1/6, drzewo decyzyjne z p=1/6.

7 Modele z najlepszymi rezultatami

Najlepiej rokujące modele zbudujemy ponownie, tym razem dla całego zbioru treningowego i porównamy ich skuteczność na zbiorze testowym. Aby skontrolować, czy modele nie są przeuczone, sprawdzimy również jak wygląda ich predykcja na zbiorze treningowym.

7.1 Regresja logistyczna

Jednym z modeli, przy którym uzyskaliśmy dobre wyniki jest model regresji logistycznej, zbudowany na wszystkich zmiennych zawartych w danych.

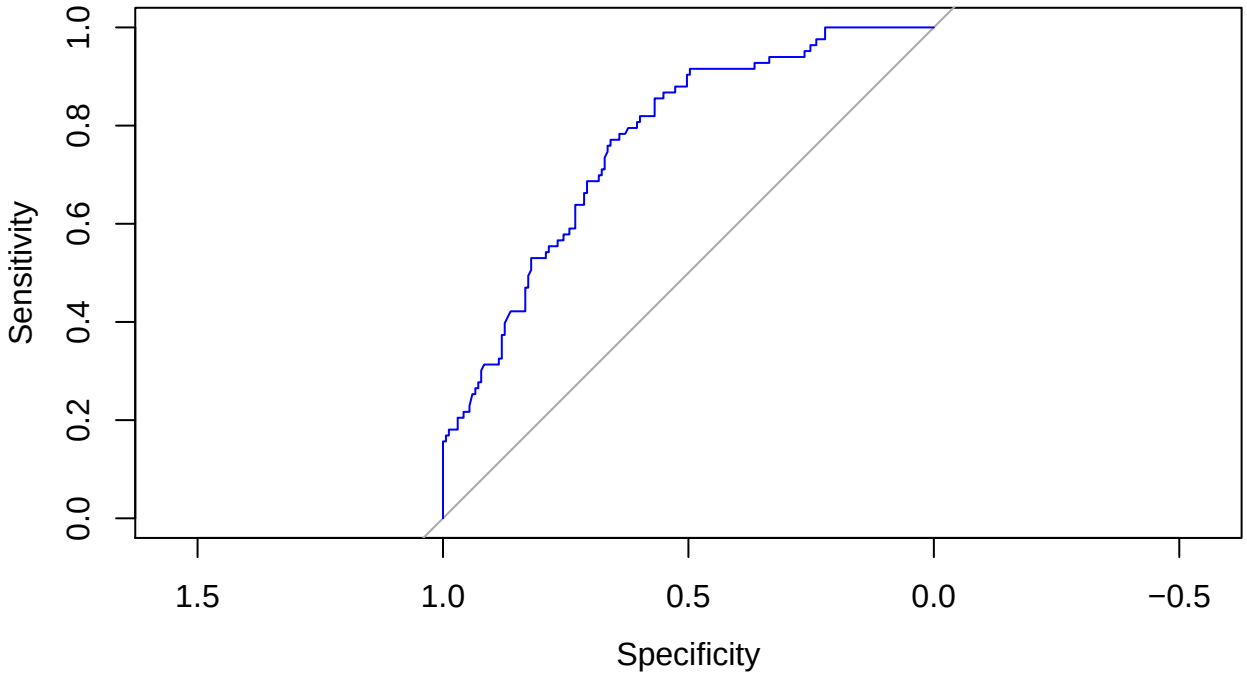
	ACC	TPR	TNR	F1	MMC
p=0.5 test	0.73	0.47	0.86	0.22	0.98
p=1/6 test	0.61	0.77	0.53	0.58	0.69
p=0.5 train	0.79	0.51	0.90	0.17	0.78
p=1/6 train	0.68	0.90	0.59	0.56	0.43

Tabela 27: Tabela wskaźników dla LR

W powyższej tabeli widzimy, że dokładność dla danych testowych różni się od tej uzyskanej dla danych treningowych na drugim miejscu po przecinku. Oznacza to, że nasz model nie jest przeuczony. W przypadku, gdy threshold ustawiony został na 1/6 otrzymujemy satysfakcjonujące wyniki. Dla zbioru testowego czułość wynosi 0.77, wartość F_1 to 0.58 a średni koszt błędnej klasyfikacji wynosi 0.69.

7.2 Regresja logistyczna z kategoryzacją

Następnym modelem, który zbudujemy dla całego zbioru testowego jest regresja logistyczna z kategoryzacją.



Rysunek 29: Krzywa ROC dla LR(category)

	ACC	TPR	TNR	F1	MMC
p=0.5 test	0.71	0.37	0.87	0.19	1.12
p=1/6 test	0.65	0.87	0.54	0.60	0.53
p=0.5 train	0.78	0.48	0.91	0.16	0.82
p=1/6 train	0.65	0.89	0.55	0.60	0.47

Tabela 28: Tabela wskaźników dla LR(category)

Na podstawie powyższej tabeli stwierdzamy, że podobnie jak w przypadku poprzedniego modelu, nie obserwujemy przeuczenia. Dla thresholdu równego 1/6 otrzymujemy satysfakcjonujące wyniki. Szczególną uwagę możemy zwrócić na niski średni koszt, który nie różni się znacznie na zbiorze treningowym i testowym - wynosi odpowiednio 0,47 i 0,53. Wysoka jest również czułość tego modelu: dla zbioru testowego jest to 0,87, zaś dla zbioru treningowego 0,89. Dla obu zbiorów F_1 wynosi 0,6.

7.3 QDA

Kolejnym modelem, który dał dobre wyniki był model kwadratowej analizy dyskryminacyjnej.

	ACC	TPR	TNR	F1	MMC
p=0.5 test	0.68	0.61	0.72	0.39	0.83
p=1/6 test	0.67	0.77	0.62	0.51	0.64
p=0.5 train	0.76	0.75	0.76	0.36	0.54
p=1/6 train	0.68	0.84	0.61	0.53	0.50

Tabela 29: Tabela wskaźników dla QDA

Po raz kolejny, tym razem korzystając z metody kwadratowej analizy dyskryminacyjnej otrzymujemy dość dobre wartości wskaźników. Jak widać, warto zmienić punkt odcięcia. Zmniejszamy wówczas średnie koszty błędnej klasyfikacji z 0.83 na 0.64 (dla zbioru testowego). W przypadku p*=1/6 dostajemy czułość na poziomie 77%.

7.4 Regresja logistyczna ze stepem i p=1/6

	ACC	TPR	TNR	F1	MMC
p=1/6 test	0.66	0.81	0.59	0.55	0.60
p=1/6 train	0.65	0.89	0.56	0.59	0.48

Tabela 30: Tabela wskaźników dla LR(step)

Tabela 30 zawiera wartości wskaźników, które uzyskaliśmy dla modelu regresji logistycznej przy ustawionym p=1/6. Ponownie możemy stwierdzić, że nasz model nie został przeuczony, gdyż wartości które otrzymaliśmy dla zbioru treningowego oraz testowego są bardzo zbliżone.

7.5 LDA z p = 1/6

Kolejnym modelem, który dał dobre wyniki był model liniowej analizy dyskryminacyjnej z p=1/6.

	ACC	TPR	TNR	F1	MMC
p=1/6 test	0.61	0.86	0.49	0.64	0.58
p=1/6 train	0.61	0.87	0.50	0.63	0.55

Tabela 31: Tabela wskaźników dla LDA

Dla modelu opartego na LDA na obu zbiorach otrzymujemy bardzo zbliżone wyniki. Dla zbioru testowego wartość TPR wynosi 0.86, współczynnik F_1 ma wartość 0.64.

7.6 Drzewo decyzyjne z p=1/6

	ACC	TPR	TNR	F1	MMC
p=1/6 test	0.33	1.00	0.00	1.00	0.67
p=1/6 train	0.29	1.00	0.00	1.00	0.71

Tabela 32: Tabela wskaźników dla drzewa decyzyjnego

W Tabeli 32 widzimy, że dokładność tego modelu jest bardzo niska zarówno na zbiorze testowym, jak i treningowym. Czułość wynosi aż 1, za to specyficzność wynosi 0. Oznacza to, że model klasyfikuje wszystkie obserwacje do jednej klasy, co jest dla niego dyskwalifikujące.

8 Podsumowanie

Porównanie wskaźników dla wszystkich modeli, które uczyniliśmy 'cost-sensitive' za pomocą zastosowania thresholdu $p = 1/6$, zostało przedstawione w Tabeli 33.

	ACC	TPR	TNR	F1	MMC
LR	0.61	0.77	0.53	0.58	0.69
LR(category)	0.65	0.87	0.54	0.60	0.53
LR(step)	0.66	0.81	0.59	0.55	0.60
LDA	0.61	0.86	0.49	0.64	0.58
QDA	0.67	0.77	0.62	0.51	0.64
Drzewa decyzyjne	0.33	1.00	0.00	1.00	0.67

Tabela 33: Tabela wskaźników dla wszystkich modeli na zbiorze testowym z p =1/6

Najniższy średni koszt złej klasyfikacji obeserwujemy dla modelu regresji logistycznej z kategoryzacją. Ten model ma także wysoką dokładność, czułość i specyficzność. Gdybyśmy mieli wybrać najlepszy ze zbudowanych modeli, byłby to właśnie ten. Jednak niektóre inne modele dają bardzo zbliżone wyniki. Niewiele gorzej prezentuje się model LDA, ma on nawet wyższe F_1 niż poprzedni model. Być może zbudowanie modeli z innym wyborem zmiennych dałoby lepsze wyniki, więc prawdopodobnie otrzymane modele nie są optymalne. Poprawę wyników moglibyśmy uzyskać stosując bardziej zaawansowane metody, takie jak np. bagging.