Imię i nazwisko: Paulina Busz.
Nr. Indeksu: 186022.
Kierunek: Analityka gospodarcza.

Statystyczna analiza wielowymiarowa

Analiza Dyskryminacyjna

Dane: Analizę przeprowadzono na zbiorze danych „Data_bankrupcy” zawierającym 15 zmiennych i 990 obserwacji.

Zmienne: W analizowanym zbiorze danych zmiennymi są wskaźniki opisujące kondycję finansową firmy. Dla ułatwienia obliczeń ich nazwy zostały skrócone.

Original name	Changed name
X1 - net profit / total assets	X1
X8 - book value of equity / total liabilities	X8
X13 - (gross profit + depreciation) / sales	X13
X27 - profit on operating activities / financial expenses	X27
X28 - working capital / fixed assets	X28
X32 - (current liabilities * 365) / cost of products sold	X32
X33 - operating expenses / short-term liabilities	X33
X37 - (current assets - inventories) / long-term liabilities	X37
X43 - rotation receivables + inventory turnover in days	X43
X55 - working capital	X55

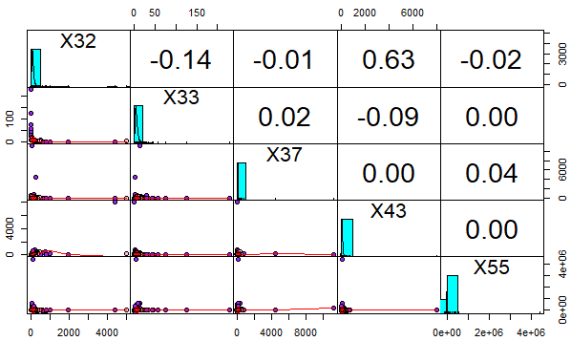
Obserwacje: 990 polskich przedsiębiorstw (Grupa 1 – 897, Grupa 2 – 93)

Zbiór danych został pobrany ze stony http://archive.ics.uci.edu/ml/index.php (arkusz 1), jednakże z podowu współliniowości niektórych zmiennych oraz dużej liczby obserwacji został on zmodyfikowany to postaci przedstawionej powyżej. Oryginalne dane zostały pobrane ze strony https://www.emis.com/. Dane zostały sklasyfikowane na dwie grupy.

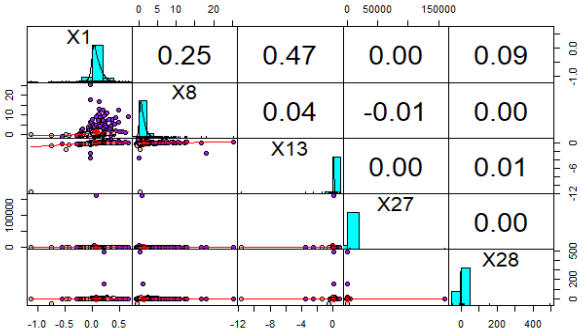
Oryginalne dane: Metodologia doboru danych polegała na określeniu statusu bankructwa firmy a następnie pobraniu jej raportu finansowego sprzed 5 lat od tego momentu.

Grupa Pierwsza zawiera dane finansowe firm, które ogłosiły bankructwo.

Grupa Druga zawiera dane finansowe firm, które nie ogłosiły bankructwa.



Wykres 1. Pairs.panels – zmienne X32,X33,X37,X43,X55.



Wykres 2. Pairs.panels – zmienne X1,X8,X13,X27,X28.

Dokonano podziału zbioru danych (za pomocą komendy „sample()”) na zbiór uczący (training) oraz zbiór testowy (testing). Za wektor prawodpowobieństwa przyjęto [0.6; 0,4], oznacza to, iż każda wartość pochodząca ze zbioru ma 60% szans aby zostać zakwalifikowaną do grupy uczącej oraz 40% aby trafić do grupy testującej.

Training	628 observations
Tesiting	362 observations

Tabela 1. Podział zbioru.

Bankrupted	Not bankrupted
8,24%	91,76%

Tabela 2. Przypisane do grup.

W kolejnym kroku wyznaczono liniową funkcję dyskryminacyjną zawierającą liniowe kombinacje zmiennch. Funkcja ta została ustalona w oparciu o zbiór uczący. Środowisko R umożliwia wyznaczenie liniowej funkcji dyskryminacyjnej za pomocą funkcji „lda()” wchodzącej w skład pakietu „Mass”.

Uzyskano wyniki, na podstawie których stwierdzono, iż 91,76% obserwacji ze zbioru uczącego należy do grupy przedsiębiorstw, które nie ogłosiły bankructwa, natomiast pozostałe 8,24% firm należy do grupy pierwszej. Ponadto obliczono wartości średnich dla każdej zmiennej w danej grupie.

	X1	X8	X13	X27	X28	X32	X33	X37	X43	X55
Bankrupted	0,028	0,470	0,029	4,017	0,127	137,102	3,846	27,003	119,103	1049,688
Not_bankrupted	0,082	1,341	0,075	315,528	0,776	105,903	6,629	27,594	115,354	16572,506

Tabela 3. Wartości średnie.

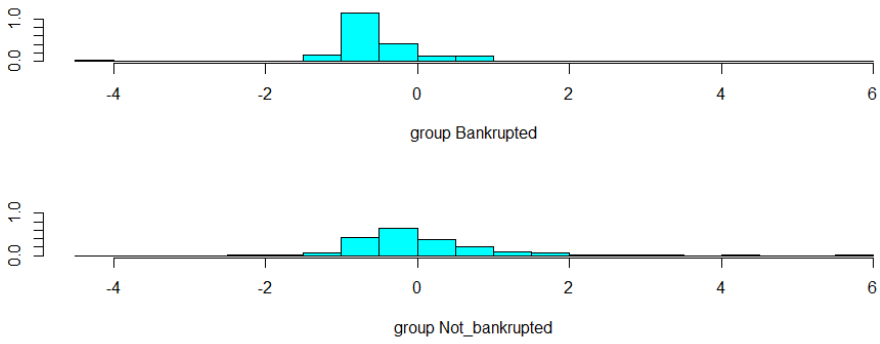
Oszacowano współczynniki liniowej funkcji dyskryminacyjnej. Jest ona liniową kombinacją 10 zmiennych:

*2,1797460*X1+0,3551056*X8+2,7024950*X13+0,0000121*X27+0,0035291*X28+0,0017202*X32+0,0000523*X33+ 0,0000523*X37+0,0022270*X43+0,0000001*X55*

Linear discriminant function									
X1	X8	X13	X27	X28	X32	X33	X37	X43	X55
2,1797460	0,3551056	2,7024950	0,0000121	0,0035291	0,0017202	0,0212080	0,0000523	0,0022270	0,0000001

Tabela 4. Współczynniki liniowej funkcji dyskryminacyjnej.

Kolejnym krokiem analizy było wykonanie predykcji oraz macierz kontyngencji dla próby uczącej oraz porównanie wyników z próbą testową.



Wykres 3. Histogramy.

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	1,000	0,000
	Not bankrupted	48,000	546,000

Tabela 5. Macierz kontyngencji na zbiorze uczącym.

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	2,000	0,000
	Not bankrupted	42,000	351,000

Tabela 6. Macierz kontyngencji na zbiorze testującym.

Trafność dla obliczeń wykonanych na zbiorze uczącym wynosi 91,93% natomiast dla danych ze zbioru testującego wynosi ona 89,37%. Wskaźnik został obliczony jako suma wartości na diagonali macierzy podzielona przez sumę wszystkich wartości w macierzy. Wstępna ocena wartości tego wskaźnika jest bardzo pozytywna, jednakże warto zauważyć, iż tak wysoka jego wartość może być spowodowana faktem, iż firmy, które ogłosiły bankurctwo w badanym okresie stanowią jedynie 10,37% wystkich badanych firm.

Dla porównania wyznaczono również kwadratową funkcję dyskryminacyjną, za pomocą komendy „qda()" zawartej w pakiecie „Mass".

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	49,000	360,000
	Not bankrupted	0,000	186,000

Tabela 7. Macierz kontyngencji na zbiorze uczącym. Kwadratowa funkcja dyskryminacyjna.

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	34,000	226,000
	Not bankrupted	42,000	125,000

Tabela 8. Macierz kontyngencji na zbiorze testującym. Kwadratowa funkcja dyskryminacyjna.

Współczynnik trafności dla obliczeń wykonanych na zbiorze uczącym dla kwadratowej funkcji dyskryminacyjnej jest niewielki. Wynosi on jedynie 39,49%, ten sam współczynnik wykonany na zbiorze testującym dla kwadratowej funkcji wynosi 40,25%. Porównując uzyskane wyniki z liniową funkcją dyskryminacyjną można jednoznacznie stwierdzić nadrzędność funkcji liniowej nad kwadratową.

Dla porównania wykonano macierz kontyngencji dla próby uczącej oraz testującej korzystając z estymatora bayesowskiego, do stworzenia którego użyto komendy „naiveBayes()" z pakietu „e1071".

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	49	353
	Not bankrupted	0	193

Tabela 9. Macierz kontyngencji na zbiorze uczącym. Estymator bayesowski.

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	41	12
	Not bankrupted	3	339

Tabela 10. Macierz kontyngencji na zbiorze testującym. Estymator bayesowski.

Współczynnik trafności dla obliczeń wykonanych na zbiorze uczącym wynosi 40,67% natomiast ten sam współczynnik dla obliczeń wykonanych na zbiorze testującym wynosi aż 97,20%. Można zatem wysunąć wniosek, iż estymacja bayesowska jest bardziej trafna od kwadratowej fukcji dyskryminacyjnej jednakże, w porównaniu z liniową funkcją dyskryminacyjną jest ona gorsza.

Ostatnim krokiem analizy było wykonanie macierzy kontyngencji wykonanych na zbiorze uczącym oraz testującym opartych na wektorach nośnych. Liczba wektorów nośnych dla próby uczącej wynosiła 66 natomiast dla próby testującej wynosił on 60.

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	49	0
	Not bankrupted	0	546

Tabela 11. Macierz kontyngencji na zbiorze uczącym. Wektory nośne.

		Actual	
		Bankrupted	Not bankrupted
Predicted	Bankrupted	44	0
	Not bankrupted	0	351

Tabela 12. Macierz kontyngencji na zbiorze testującym. Wektory nośne.

Analizując wyniki z powyższych tabeli można jednoznacznie stwierdzić 100 % trafności w obydwu grupach. Klasyfikuje on tę metodę jako najlepszą spośród wyżej zbadanych metod pod względem trafności.