
Czynniki zwiększające prawdopodobieństwo bankructwa firm w Polsce

Krzysztof Karabon

Spis treści

Wstęp	3
Zmienne w modelu	3
Metodyka przygotowania danych i wyboru zmiennych	5
Obliczenia	6
Bayesowskie łączenie wiedzy	6
Regresja logistyczna	9
Bayesowska regresja logistyczna	10
Podsumowanie	12

Ekonometria bayesowska 2016/2017

Wstęp

Celem projektu jest sprawdzenie jakie charakterystyki dostępne w sprawozdaniu finansowym i ich kombinacje istotnie wpływają na prawdopodobieństwo bankructwa firmy w ciągu roku. Zbiór danych obejmuje polskie firmy na podstawie serwisu Emerging Markets Information (EMIS), który jest bazą danych zawierającą informacje na temat rynków wschodzących na całym świecie. Spółki, które zbankrutowały zanalizowano w okresie 2000-2012, podczas gdy firmy nadal działające oceniano od 2007 do 2013 roku.

Zbiór danych zawierał 5 tabel danych, zawierających dane o bankructwie firmy w przeciągu innego horyzontu czasowego. Do badania skorzystano z tabel zawierających informacje o upadku firmy w czasie 1, 2 i 3 lat. Zbiory te zawierają odpowiednio 5910, 9792 oraz 10503 rekordy i 410, 515 oraz 495 upadków.

Dane pochodzą ze strony <http://archive.ics.uci.edu/>, dostępne dzięki Sebastianowi i Jakubowi Tomczakom oraz Maciejowi Ziębie z Politechniki Wrocławskiej.

Zmienne w modelu

Zmienne w modelu zawierają informację finansowe o spółkach polskich. Dostępne zmienne wymienione zostały poniżej:

- Attr 1 – zysk netto / aktywa ogółem
- Attr 2 – zobowiązania ogółem / aktywa ogółem
- Attr 3 – kapitał obrotowy / aktywa ogółem
- Attr 4 – zobowiązania krótkoterminowe / aktywa ogółem
- Attr 5 – $[(\text{środkie pieniężne} + \text{krótkoterminowe papiery wartościowe} + \text{należności} - \text{zobowiązania krótkoterminowe}) / (\text{koszty operacyjne} - \text{amortyzacja})] * 365$
- Attr 6 – zyski zatrzymane / aktywa ogółem
- Attr 7 – EBIT / aktywa ogółem
- Attr 8 – wartość księgowa kapitału własnego / pasywów ogółem
- Attr 9 – sprzedaż / aktywa ogółem
- Attr 10 – kapitał własny / aktywa ogółem
- Attr 11 - $(\text{zysk brutto} + \text{pozycje nadzwyczajne} + \text{koszty finansowe}) / \text{aktywa ogółem}$
- Attr 12 - $-\text{zysk brutto} / \text{zobowiązania krótkoterminowe}$
- Attr 13 - $-\text{zysk brutto} + \text{amortyzacja} / \text{przychody ze sprzedaży}$
- Attr 14 - $(\text{zysk brutto} + \text{odsetki}) / \text{aktywa ogółem}$
- Attr 15 - $(\text{zobowiązania ogółem} * 365) / (\text{zysk brutto} + \text{amortyzacja})$
- Attr 16 - $(\text{zysk brutto} + \text{amortyzacja}) / \text{zobowiązania ogółem}$
- Attr 17 - $\text{aktywa ogółem} / \text{pasywa ogółem}$
- Attr 18 - $\text{zysk brutto} / \text{aktywa ogółem}$

- Attr 19 – zysk brutto / przychody ze sprzedaży
- Attr 20 – (zapasy * 365) / przychody ze sprzedaży
- Attr 21 – przychody ze sprzedaży (okres N) / przychody ze sprzedaży (okres N-1)
- Attr 22 – zysk na działalności operacyjnej / aktywa ogółem
- Attr 23 – zysk netto / przychody ze sprzedaży
- Attr 24 – Zysk brutto (w ciągu 3 lat) / aktywa ogółem
- Attr 25 – (kapitał własny - kapitał akcyjny) / aktywa ogółem
- Attr 26 – (zysk netto + amortyzacja) / zobowiązania ogółem
- Attr 27 – zysk na działalności operacyjnej / koszty finansowe
- Attr 28 – kapitał obrotowy / aktywa trwałe
- Attr 29 – logarytm sumy aktywów
- Attr 30 – (Zobowiązania Ogółem - środki pieniężne) / przychody ze sprzedaży
- Attr 31 – (zysk brutto + odsetki) / przychody ze sprzedaży
- Attr 32 – (zobowiązania krótkoterminowe * 365) / koszty sprzedanych produktów
- Attr 33 – koszty operacyjne / zobowiązania krótkoterminowe
- Attr 34 – koszty operacyjne / zobowiązania ogółem
- Attr 35 – zysk ze sprzedaży / aktywa ogółem
- Attr 36 – sprzedaży ogółem / aktywa ogółem
- Attr 37 – (aktywa obrotowe - zapasy) / zobowiązania długoterminowe
- Attr 38 – kapitał / aktywa ogółem
- Attr 39 – zysk ze sprzedaży / przychody ze sprzedaży
- Attr 40 – (aktywa obrotowe - zapasy - należności) / zobowiązania krótkoterminowe
- Attr 41 – zobowiązania ogółem / ((zysk z działalności operacyjnej + amortyzacja) * (12/365))
- Attr 42 – Zysk na działalności operacyjnej / przychody ze sprzedaży
- Attr 43 – Rotacja należności + obroty zapasów w dniach
- Attr 44 – (należności * 365) / przychody ze sprzedaży
- Attr 45 – zysk netto / zapasy
- Attr 46 – (aktywa obrotowe - zapasy) / zobowiązania krótkoterminowe
- Attr 47 – (zapasy * 365) / koszty sprzedanych produktów
- Attr 48 – EBITDA (zysk z działalności operacyjnej - amortyzacja) / aktywa ogółem
- Attr 49 – EBITDA (zysk z działalności operacyjnej - amortyzacja) / przychody ze sprzedaży
- Attr 50 – aktywa obrotowe / zobowiązania ogółem
- Attr 51 – Zobowiązania krótkoterminowe / aktywa ogółem
- Attr 52 – (zobowiązania krótkoterminowe * 365) / koszty sprzedanych produktów
- Attr 53 – kapitał / aktywa trwałe
- Attr 54 – kapitał stały / aktywa trwałe
- Attr 55 – kapitału obrotowy

- Attr 56 – (przychody ze sprzedaży - koszt wytworzenia sprzedanych produktów) / przychody ze sprzedaży
- Attr 57 – (aktywa obrotowe - zapasy - zobowiązania krótkoterminowe) / (zysk brutto na sprzedaży - amortyzacja)
- Attr 58 – całkowite koszty / łączna sprzedaż
- Attr 59 – zobowiązania długoterminowe / kapitał własny
- Attr 60 – przychody ze sprzedaży / zapasy
- Attr 61 – przychody ze sprzedaży / należności
- Attr 62 – (zobowiązania krótkoterminowe * 365) / przychody ze sprzedaży
- Attr 63 – przychody ze sprzedaży / zobowiązania krótkoterminowe
- Attr 64 – przychody ze sprzedaży / środki trwałe

Metodyka przygotowania danych i wyboru zmiennych

Pierwszym etapem przygotowania danych było zastąpienie braków danych średnimi wartościami danej zmiennej. W kolejnym kroku odrzucono zmienne o większej bezwzględnej korelacji niż 0,95.

W dalszym etapie doboru zmiennych zbudowano modele regresji liniowej na podstawie bayesowskiego łączenia wiedzy. Dziesięć zmiennych z największym prawdopodobieństwem obecności w najlepszych modelach zostało wybranych do krótkiej listy potencjalnych zmiennych.

Następnie sprawdzono istotność zmiennych z krótkiej listy szacując model regresji logistycznej i odrzucono zmienne nieistotne statystycznie na poziomie ufności 0,1.

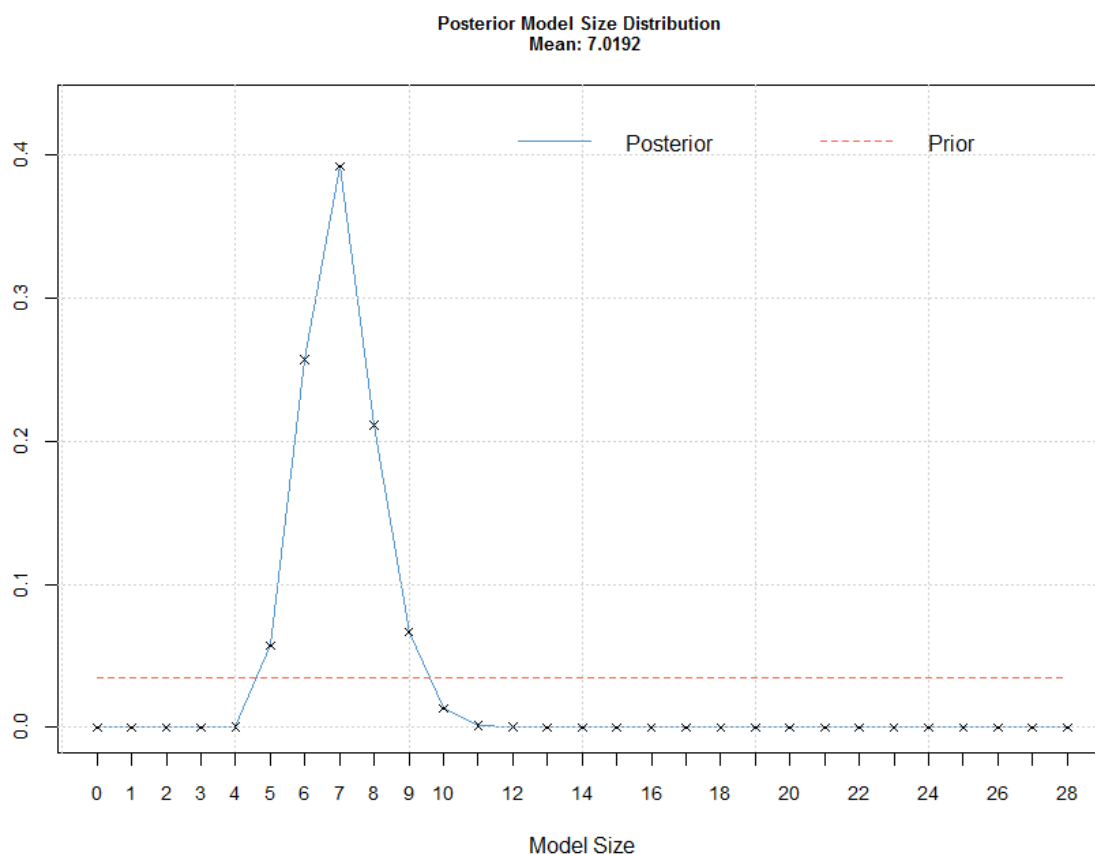
Obliczenia

Wszystkie obliczenia dokonane na potrzeby projektu zostały dokonane w języku programowania R, skrypt dodany został w osobnym pliku.

Bayesowskie łączenie wiedzy

W celu wyznaczenia krótkiej listy 10 najbardziej istotnych zmiennych skorzystano z tabeli danych zawierającej informację o bankructwie firm w horyzoncie 3 lat. Na podstawie bayesowskiego łączenia wiedzy zbudowano 1 000 000 modeli, z czego odrzucono 10 000 początkowych a zachowano 5 000 najlepszych. Algorytmem doboru zmiennych do modelu był algorytm birth-death z losowym parametrem theta.

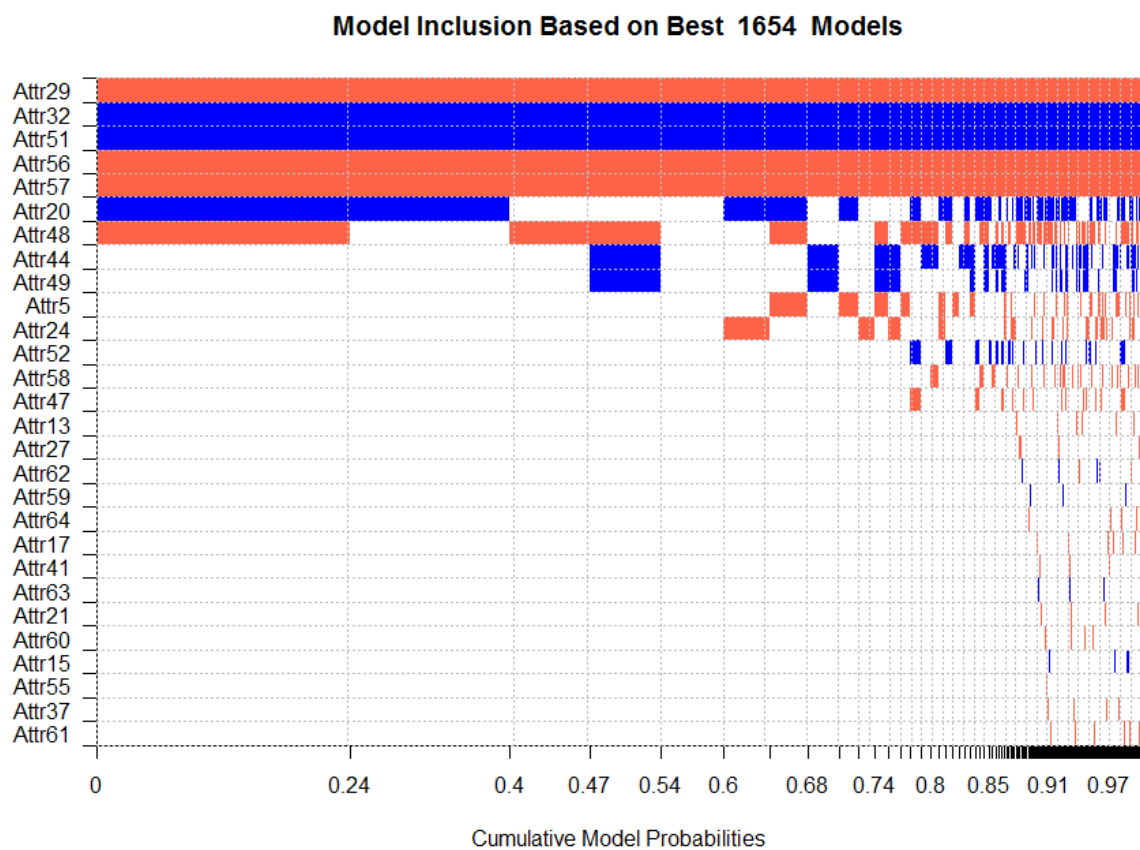
Wykres 1 Wykres rozkładu ilości zmiennych w 1654 najlepszych modeli



Średnio zmiennych w modelach a posteriori było 7 z głównym ciężarem prawdopodobieństwa między 4 a 11 zmiennymi w modelu.

Poniżej przedstawiono obecność poszczególnych zmiennych w najlepszych modelach z zaznaczeniem prawdopodobieństwa aposteriori modeli oraz znaki parametrów zmiennych.

Wykres 2 Wykres prawdopodobieństwa obecności zmiennej w najlepszych 1654 modelach i znaku parametru a posteriori



Łatwo można zauważyć, iż najbardziej istotnymi są zmienne Attr29, Attr32, Attr51, Attr56, Attr57 i obecne są w 100% lub prawie 100% modeli. Dokładne dane prezentuje tabela poniżej:

Tabela 1 Parametry zmiennych w 1654 najlepszych modeli stworzonych na podstawie bayesowskiego łączenia wiedzy

Zmienna	PIP	Post Mean	Post SD	Cond.Pos.Sign	Idx
Attr29	1,00	-0,03	0,00	0	9
Attr32	1,00	0,02	0,00	1	10
Attr51	1,00	0,03	0,00	1	17
Attr56	1,00	-0,02	0,00	0	20
Attr57	1,00	-0,02	0,00	0	21
Attr20	0,61	0,01	0,01	1	5
Attr48	0,57	-0,01	0,01	0	15
Attr44	0,24	0,00	0,01	1	13
Attr49	0,19	0,00	0,01	1	16
Attr5	0,12	0,00	0,00	0	1
Attr24	0,10	0,00	0,00	0	7
Attr52	0,05	0,00	0,00	1	18
Attr47	0,03	0,00	0,00	0	14
Attr58	0,03	0,00	0,00	0	22
Attr27	0,01	0,00	0,00	0	8
Attr62	0,01	0,00	0,00	0,79	26
Attr13	0,01	0,00	0,00	0	2
Attr59	0,01	0,00	0,00	1	23
Attr55	0,01	0,00	0,00	0	19
Attr41	0,01	0,00	0,00	0	12
Attr64	0,01	0,00	0,00	0	28
Attr15	0,01	0,00	0,00	1	3
Attr21	0,01	0,00	0,00	0	6
Attr61	0,00	0,00	0,00	0,34	25
Attr60	0,00	0,00	0,00	0	24
Attr17	0,00	0,00	0,00	0	4
Attr63	0,00	0,00	0,00	1	27
Attr37	0,00	0,00	0,00	0	11

Postanowiono wykorzystać 10 zmiennych z największym prawdopodobieństwem obecności w najlepszych modelach, ponieważ kolejne zmienne i modele niewiele wносиły do objaśniania bankructwa firm.

Regresja logistyczna

Kolejnym etapem wyboru zmiennych było zbudowanie regresji logistycznej wykorzystując tabelę danych zawierającą informację o bankructwie firm w horyzoncie 2 lat, na podstawie zmiennych wybranych w punkcie poprzednim. Tabela poniżej przedstawia parametry wyestymowanego modelu i ich istotność:

Tabela 2 Parametry modelu regresji logistycznej na zmiennych z krótkiej listy

Zmienna	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2,894	0,057	-51,110	< 0,0001	***
Attr29	-0,255	0,045	-5,617	0,000	***
Attr32	0,001	0,062	0,021	0,983	
Attr51	0,010	0,048	0,212	0,832	
Attr56	-6,020	5,085	-1,184	0,236	
Attr57	0,016	0,062	0,262	0,794	
Attr20	0,018	0,029	0,621	0,535	
Attr48	-0,133	0,044	-3,031	0,002	**
Attr44	-0,152	0,479	-0,318	0,751	
Attr49	1,382	1,217	1,136	0,256	
Attr5	0,002	0,052	0,042	0,967	

Jedynie dwie zmienne – Attr 29 oraz Attr 48 okazały się istotnie statystycznie, dlatego też w dalszej pracy jedynie one zostały uwzględnione. Wyestymowany został na tabeli danych zawierające informację o bankructwie firm w horyzoncie 2 lat model regresji liniowej, który posłuży za wiedzę a priori w bayesowskim modelu regresji logistycznej. Wyniki jej zamieszczone są w tabeli poniżej:

Tabela 3 Parametry zmiennych w regresji logistycznej

Zmienna	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.92811	0.04671	-62.681	<0.00002	***
Attr29	-0.25776	0.04519	-5.704	<0.00001	***
Attr48	-0.11926	0.04211	-2.832	0.00462	**

Interpretacja zmiennych jest zgodna z biznesowym rozumieniem działań rynkowych. Im logarytm sumy aktywów lub EBITDA (zysk z działalności operacyjnej - amortyzacja) przez aktywa ogółem jest mniejszy tym większe jest prawdopodobieństwo upadku firmy.

Bayesowska regresja logistyczna

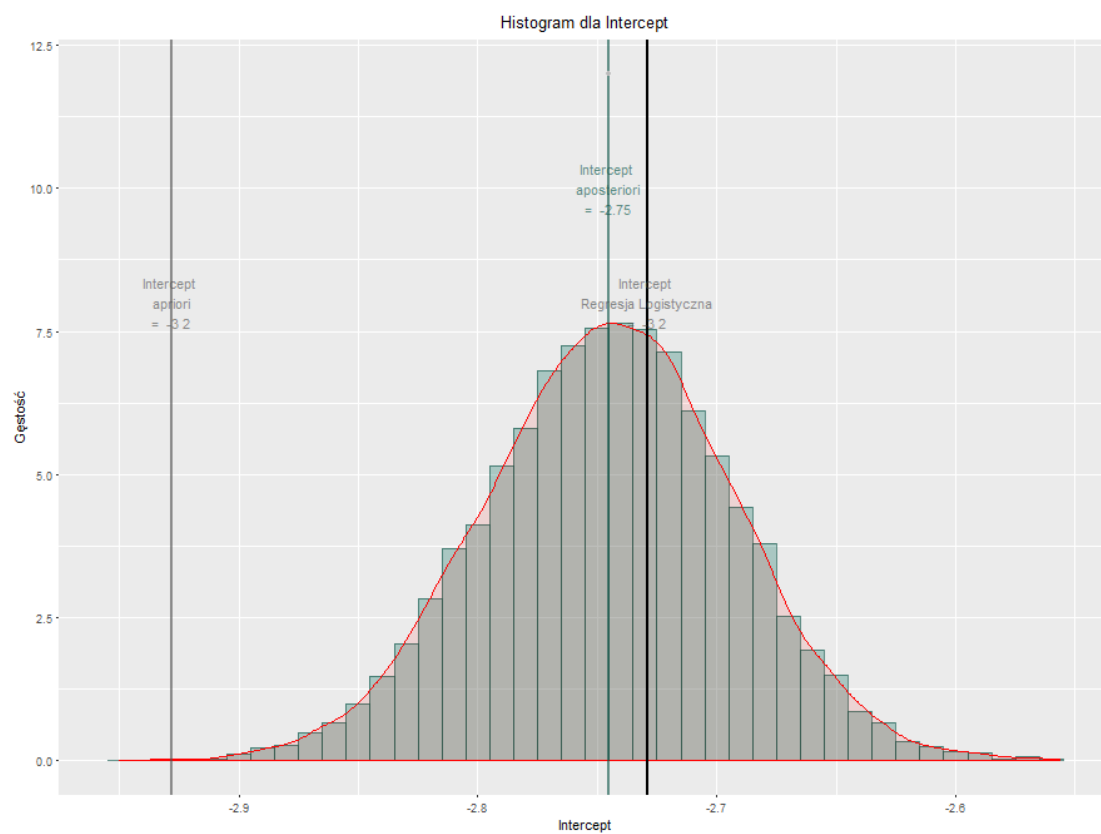
Ostatnim etapem stworzenia modelu objaśniającego bankructwo firm w horyzoncie jednego roku jest zbudowanie regresji logistycznej z danymi a priori. Za dane a priori odnośnie wielkości współczynników posłużyły parametry wyestymowane w poprzednim punkcie. Z powodu lekko rozbieżnego charakteru danych postanowiono zmniejszyć wagę informacji a priori poprzez podzielenie wielkości próbki przez 100 dla zmiennych (97.92) a dla stałej precyzje ustanowiono na poziomie 50.

W ten sposób zbudowano 5 szeregów markowa po 10 000 obserwacji, z każdego odrzucono pierwsze 5 000 rekordów i wybrano co piątą z pozostałych. Na ich podstawie określono histogramy dla każdej zmiennej osobno przedstawione na wykresach 3, 4 i 5. Czarną pionową kreską naznaczono wartość oszacowania parametry klasyczną regresją logistyczną, kolorem szarym wartość a priori, a zielonym wartość oczekiwaną.

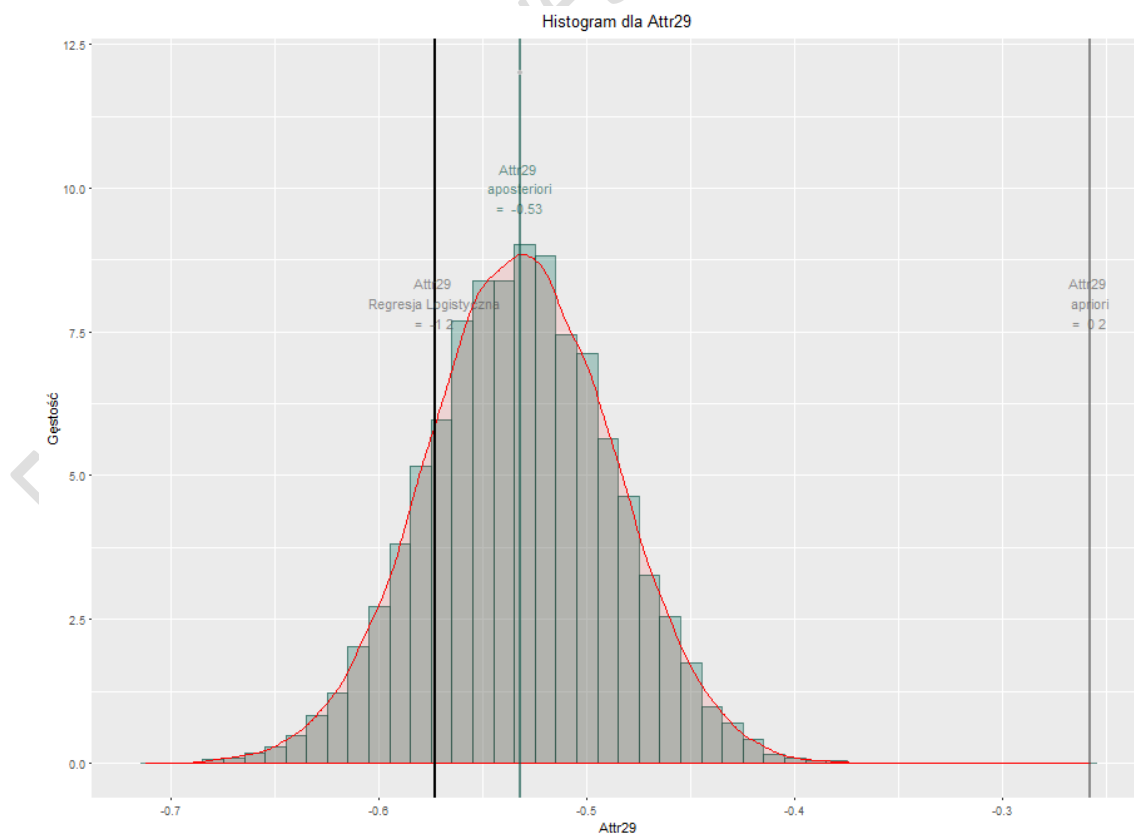
Wartością oczekiwaną dla stałej jest -2.75, dla Attr29 -0.53 a dla Attr48 0.19. Przedziały ufności plasują się następująco:

- 90-procentowy przedział ufności dla stałej wynosi: -2.83 – -2.66, z odchyleniem standardowym równym 0.05,
- 90-procentowy przedział ufności dla Attr29 wynosi: -0.61 – -0.46, z odchyleniem standardowym równym 0.04,
- 90-procentowy przedział ufności dla Attr48 wynosi: -0.34 – -0.05, z odchyleniem standardowym równym 0.09.

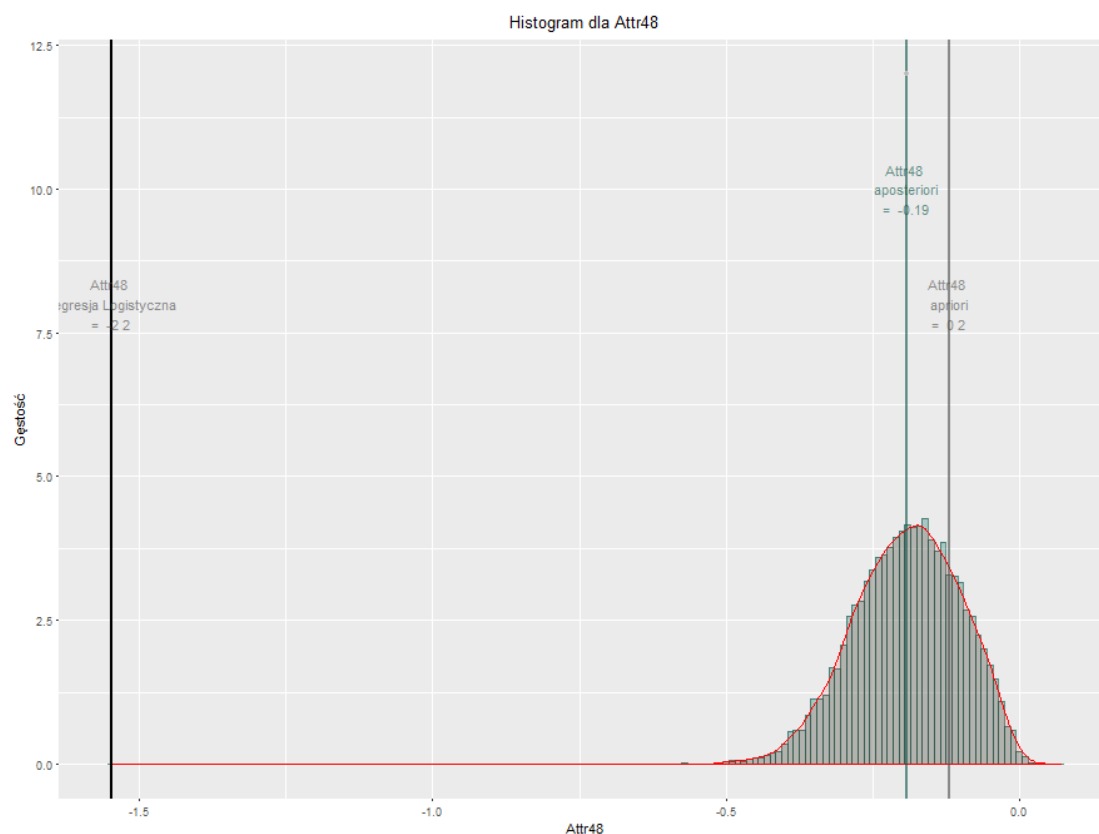
Wykres 3 histogram rozkładu dla stałej w modelu



Wykres 4 histogram rozkładu współczynnika stojącego przy zmiennej Attr29 – logarytm sumy aktywów



Wykres 5 histogram rozkładu współczynnika stojącego przy zmiennej Attr48 – EBITDA / aktywa ogółem



Na podstawie histogramów zauważyć można że oszacowania a priori i a posteriori dla wszystkich parametrów były zgodne co do znaku i podobnej sile. Jedynie znaczna różnica nastąpiła przy zmiennej Attr48 oznaczającej EBITDA / aktywa ogółem.

Podsumowanie

W niniejszej pracy postarano się sprawdzić przy pomocy podstawowych narzędzi ekonometrii bayesowskiej jakie zmienne i z jaką siłą wpływają na bankructwo Polskich firm w horyzoncie jednego roku. Na podstawie bayesowskiego łączenia wiedzy wybrano krótką listę zmiennych, która następnie została zweryfikowana w modelu regresji logistycznej. W ostatnim etapie dokonano bayesowskiej regresji logistycznej wykorzystując dane z a priori z estymacji na innym zbiorze danych.

Zmiennymi istotnymi jest logarytm sumy aktywów oraz EBITDA podzielone przez aktywa ogółem. Wzrost obu zmiennych ma pozytywny wpływ na przetrwanie firmy w ciągu roku.