Karol Przanowski[1], Jolanta Mamczarz[2]

[1]Szkoła Główna Handlowa w Warszawie, Instytut Statystyki i Demografii, Zakład Analizy Historii Zdarzeń i Analiz Wielopoziomowych

[2] Studentka - Szkoła Główna Handlowa w Warszawie (kierunek: Metody ilościowe w ekonomii i systemy informacyjne); Uniwersytet Warszawski (kierunek: Matematyka)

GENERATOR DANYCH CONSUMER FINANCE – METODY PORÓWNYWANIA TECHNIK MODELOWYCH W CREDIT SCORING

Streszczenie: Przedmiotem artykułu jest porównanie technik budowy kart skoringowych, które można otrzymać na wiele sposobów, wykorzystując różne metody transformacji zmiennych. Porównywanie i wykazanie wyższości jednej metody nad drugą jest dość dużym wyzwaniem. Relatywnie łatwo wykonać to na podstawie konkretnych danych. Wykazanie różnicy na podstawie jednych danych nie gwarantuje otrzymania tego samego wyniku na podstawie danych pochodzących z innego źródła. Pojawia się zatem pytanie, w jaki sposób otrzymać wnioski bardziej ogólne? Pomocne w tym celu są dane losowe otrzymane przy użyciu generatora liczb losowych. Generator wykorzystuje macierze migracji i skoringi. W artykule przedstawione są wnioski wynikające z porównania kilku technik budowy kart skoringowych. Przed budową karty skoringowej można stwierdzić, jaka technika budowy dla określonych danych jest najlepsza. Mierniki jakości modelu predykcyjnego, takie jak: Gini mierzący moc predykcyjną, Delta Gini wykorzystywany do oceny stabilności, VIF – stosowany do pomiaru współliniowości i Max *p-value* używany do oceny istotności zmiennych, wskazują na wielokryterialną naturę „dobrego modelu". Proponowany sposób postępowania może być wykorzystany w procesie budowy modelu, gdy zdefiniowanych jest wiele nietrywialnych kryteriów, np. stabilność modelu w czasie całego cyklu koniunkturalnego (ang. TTC - through the cycle).

Słowa kluczowe: credit scoring, analizy kryzysu, generator danych bankowych, portfel detaliczny, karty skoringowe, modele skoringowe, modelowanie, modele predykcyjne.

Wstęp

Techniki credit scoring są dziś powszechnie używane w biznesie. Znalazły one zastosowanie zarówno w bankowości [Huang E., 2007] do optymalizacji procesów

akceptacji produktów kredytowych, jak i modeli PD (*ang. probability of default*) stosowanych w rekomendacjach Basel II i III do wyznaczania RWA (*ang. Risk Weighted Assets*) [Basel Committee on Banking Supervision. A Revised Framework, 2005]. Bezpośredni związek z procesami biznesowymi, z jednej strony, czyni credit scoring dziedziną powszechnie znaną, ale z drugiej, utrudnia jej pełny rozwój w oderwaniu od wpływu dużych korporacji i firm konsultingowych. Istnienie sił czerpiących zyski ze stosowania credit scoring jest przyczyną pojawiania się różnych stwierdzeń nie popartych naukowymi badaniami. Przepisy chroniące dane praktycznie uniemożliwiają pełne i rzetelne studia z wykorzystaniem wielu zbiorów danych, co mocno hamuje rozwój tej dziedziny.

Obecny kryzys ekonomiczny skłania wielu badaczy do poszukiwania lepszych modeli predykcyjnych, bardziej stabilnych w czasie [Mays E., 2009].

Nasuwają się zatem następujące pytania:

- Czy możliwe są badania credit scoring bez dostępu do rzeczywistych danych?
- Czy możliwe są sposoby dowodzenia wyższości jednej techniki nad drugą w oderwaniu od rzeczywistych danych?
- Czy możliwe jest stworzenie ogólnego repozytorium danych credit scoring dostępnego dla wszystkich zainteresowanych i zawierającego wystarczająco dużo danych dotyczących szczególnych przypadków, by nazwać go OGÓLNYM i wykorzystywać do porównywania technik credit scoring?

1. Dane wykorzystane do analiz

Do badania zostały wykorzystane dwa zestawy danych rzeczywistych pochodzące z odległych dziedzin: bankowej i medycznej.

1.1. Dane bankowe

Dane bankowe pochodzą z jednego z polskich banków sektora Consumer Finance. Zawierają 50 000 wierszy i 134 kolumny. Nazwy kolumn są utajnione a zmienna celu zawiera informację o klientach, którzy posiadali opóźnienia w spłatach rat kredytowych więcej niż 60 dni w ciągu 6 miesięcy od daty obserwacji.

1.2. Dane medyczne

Dane medyczne opisują przypadki zachorowania na raka w Stanach Zjednoczonych [Kadam A., Delen D., Walker G., 2005]. Zawierają 1 343 646 wierszy i 40 kolumn. Zmienna celu opisuje przeżycie lub zgon pacjenta z powodu nowotworu danego typu w ciągu pięciu lat od diagnozy.

1.3. Generator danych losowych

Generator danych losowych Consumer Finance został obszernie opisany [Przanowski K., 2011]. Główna idea generowania danych oparta jest na łańcuchu Markowa z podaną macierzą przejść pomiędzy kategoriami opóźnień w spłacaniu rat kredytowych. Macierz ulega niewielkim zmianom w czasie na skutek wpływu makroekonomicznej zmiennej, dzięki czemu zmienia się ryzyko w czasie. Dane dla każdego nowego miesiąca są tworzone przez wyznaczenie skoringu dla wszystkich istniejących kredytów. Kredyty, które w danym miesiącu nie mają spłaconej raty znajdują się w grupie kredytów o najgorszych wartościach skoru, a ich udział w miesięcznej populacji jest określony za pomocą odpowiedniego współczynnika w macierzy przejść. Oczywiście, skoringi otrzymane przy różnych funkcjach celu mogą być różne.

Mechanizm generowania danych jest dość prosty. Można rozbudować generator w ten sposób, aby otrzymać zestawy danych dla portfeli kredytów: z małym, średnim i dużym ryzykiem, w zależności od macierzy przejść, z małą, średnią i dużą cyklicznością oraz z różną zmiennością w czasie reguł skoringowych. Jest to zatem dość elastyczny sposób generowania danych. Zbiór zawiera 2 694 377 wierszy i 56 kolumn.

2. Ogólny proces budowy modelu skoringowego

Na podstawie każdego zestawu danych uruchomiono procesy tworzenia modeli predykcyjnych. Wszystkich obliczeń dokonano w systemie SAS 9.2 [SAS Institute Inc.] wykorzystując moduły Base SAS, SAS/STAT i SAS/GRAPH. Proces tworzenia i szacowania modeli składał się z następujących etapów:

- **Próby losowe**. Tworzone są dwa zbiory danych: treningowy i walidacyjny. Są one przesunięte w czasie: zbiór walidacyjny zawiera obserwacje z późniejszego okresu niż zbiór treningowy. Taka metoda, w języku angielskim określana jako *time sampling*, pozwala badać stabilność modeli w czasie.

- **Tworzenie atrybutów** – kategoryzacja zmiennych lub grupowanie. Na podstawie statystyki entropii każda zmienna ciągła jest kategoryzowana do kilku grup zwanych atrybutami po to, aby uzyskać najbardziej jednorodne grupy pod względem ryzyka kredytowego oraz by zachować monotoniczność kategoryzacji, np. wraz ze wzrostem przedziału wieku klienta ryzyko kredytowe maleje. Jest to metoda stosowana przy drzewach decyzyjnych.
- **Wstępna selekcja zmiennych**. Chodzi o odrzucenie zmiennych o niskiej wartości predykcyjnej lub niestabilnych w czasie na podstawie prostych jednowymiarowych kryteriów, mierzących wpływ danej zmiennej na funkcję celu bez powiązania z innymi zmiennymi objaśnianymi.
- **Wielowymiarowa selekcja zmiennych** – generator modeli. W procedurze Logistic istnieje metoda selekcji zmiennych oparta na heurystyce *branch and band* [Furnival G. M., Wilson R. W., 1974]. W metodzie tej wstępnie rozważa się wszystkie możliwe kombinacje zmiennych objaśniających w następujący sposób: wszystkie modele z jedną zmienną, wszystkie z dwiema zmiennymi, wszystkie z trzema, aż do modelu pełnego ze wszystkimi zmiennymi. Analizowanie wszystkich modeli, jest oczywiście problemem NP-zupełnym. Heurystyka metodami *branch and band* pozwala wyeliminować bardzo wiele kombinacji zmiennych objaśniających pozostawiając ich na tyle mało, że możliwe jest wykonanie obliczeń w dość krótkim czasie. W wyniku otrzymuje się modele uporządkowane od najlepszego do najgorszego ze względu na ustaloną statystykę dopasowania.
- **Ocena modeli**. Nie ma jednego kryterium oceny modeli. Stosuje się wiele kryteriów, głównie związanych z mocą predykcyjną *AR* inaczej *Gini* [Thomas L. C., Edelman D. B., Crook J.N., 2002] i [Basel Committee on Banking Supervision. Working paper no. 14, 2005], stabilnością AR_{diff} inaczej *Delta Gini* (różnica względna predykcji pomiędzy zbiorami: treningowym i walidacyjnym), miary współliniowości: MAX_{VIF} - maksymalny współczynnik podbicia wariancji [Koronacki J., Mielniczuk J., 2001], $MAX_{Pearson}$ - maksymalny współczynnik korelacji Pearsona dla par zmiennych i $MAX_{ConIndex}$ - maksymalny indeks warunkowy [Welfe A., 2003] oraz miara istotności: $MAX_{ProbChiSquare}$ – maksymalna *p-value* dla zmiennych.

3. Kodowania i selekcje zmiennych

Model skoringowy, choć oparty na tym samym zestawie zmiennych, może być estymowany na różne sposoby, zależnie od metody kodowania zmiennych.

Pierwszy sposób, oznaczany przez REG, to estymacja modeli bez transformacji zmiennych. W tym przypadku potrzebna jest metoda uzupełniania brakujących danych *(ang. Missing Imputation)*; wybrano najprostszą: uzupełnianie przez średnią.

Drugi sposób, LOG, oparty jest na transformacji logitowej: każdej kategorii (atrybutowi) zmiennej przypisany jest jej *logit* (jest to metoda podobna do metody *WoE, ang. Weight of Evidence*, stosowanej w SAS Credit Scoring Solution [Siddiqi N., 2005]).

Trzeci, GRP, związany jest z binarnym kodowaniem referencyjnym zwanym *reference* lub *dummy*, zob. tabela 1. W tym przypadku poziomem referencyjnym jest atrybut o najwyższym numerze, czyli o najmniejszym ryzyku. Tak kodowanych zmiennych nie można selekcjonować metodą *branch and band*, gdyż liczba zmiennych staje się tak duża, że czas obliczeń znacząco się wydłuża. Firma Score Plus [Scallan G., 2011] proponuje, aby selekcję zmiennych binarnych przeprowadzić przy innym kodowaniu kumulatywnym [Frątczak, E. red., 2012], zwanym *ordinal* lub *nested*, zob. tabele 3, 4 i 5.

W przypadku GRP nie była możliwa taka selekcja zmiennych, jak dla REG i LOG. Wszystkie modele otrzymane sposobem REG i LOG zostały ocenione metodą GRP. Ze względu na brak selekcji zmiennych w metodzie GRP zbadano kilka algorytmów korekty liczby zmiennych w GRP. Wiele zmiennych w zetknięciu się z innymi zmiennymi w modelu staje się nie istotnych, potrzeba zatem metody ich wykrywania. Wybrano dwie metody: wstecznej eliminacji zmiennych *(ang. backward)* i krokowej *(ang. stepwise)*. Parametry finalne modelu można estymować zarówno przy kodowaniu *dummy*, jak i *nested*. Ostatecznie otrzymano 12 metod korekt modeli GRP, zob. tabela 6.

Wszystkie modele, z wyjątkiem REG, są modelami karty skoringowej, zob. tabela 2. Modele REG były rozważane jako dodatkowy punkt odniesienia.

Tabela 1. Kodowanie referencyjne – *dummy (reference)*

Numer atrybutu	Zmienna1	Zmienna2	Zmienna3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Źródło: SAS Institute Inc. 2002-2010. *SAS/STAT® 9.2: Proc Logistic - User's Guide, Other Parameterizations.*

Tabela 2. Przykładowa karta skoringowa

Zmienna	Warunek (atrybut)	Skor cząstkowy
Wiek	≤ 20	10
	≤ 35	20
	≤ 60	40
	> 60	50
Wynagrodzenie	≤ 1500	15
	≤ 3500	26
	≤ 6000	49
	> 6000	52

Źródło: Opracowanie własne.

Tabela 3. Kodowanie kumulatywne malejące – *nested descending (ordinal)*

Numer atrybutu	Zmienna1	Zmienna2	Zmienna3
1	0	0	0
2	1	0	0
3	1	1	0
4	1	1	1

Źródło: SAS Institute Inc. 2002-2010. *SAS/STAT® 9.2: Proc Logistic - User's Guide, Other Parameterizations.*

Tabela 4. Kodowanie kumulatywne rosnące – *nested ascending*

Numer atrybutu	Zmienna1	Zmienna2	Zmienna3
1	1	1	1
2	0	1	1
3	0	0	1
4	0	0	0

Źródło: Opracowanie własne.

Tabela 5. Kodowanie kumulatywne monotoniczne – *nested monotonic*

Numer atrybutu	Zmienna1	Zmienna2	Zmienna3
1	1	1	1
2	1	1	0
3	1	0	0
4	0	0	0

Źródło: Opracowanie własne.

Tabela 6. Metody korekty modeli GRP

Nawa metody	Estymacja	Selekcja	Kodowanie
NBA	nested	backward	ascending nested
NBD	nested	backward	descending nested
NBM	nested	backward	monotonic nested
NSA	nested	stepwise	ascending nested
NSD	nested	stepwise	descending nested
NSM	nested	stepwise	monotonic nested
DBA	dummy	backward	ascending nested
DBD	dummy	backward	descending nested
DBM	dummy	backward	monotonic nested
DSA	dummy	stepwise	ascending nested
DSD	dummy	stepwise	descending nested
DSM	dummy	stepwise	monotonic nested

Źródło: Opracowanie własne.

Tabela 7. Liczebności prób

Typ danych	Treningowy	Walidacyjny	Liczba wybranych zmiennych
Bankowe	27 325	12 435	60
Medyczne	29 893	17 056	23
Losowe	66 998	38 199	33

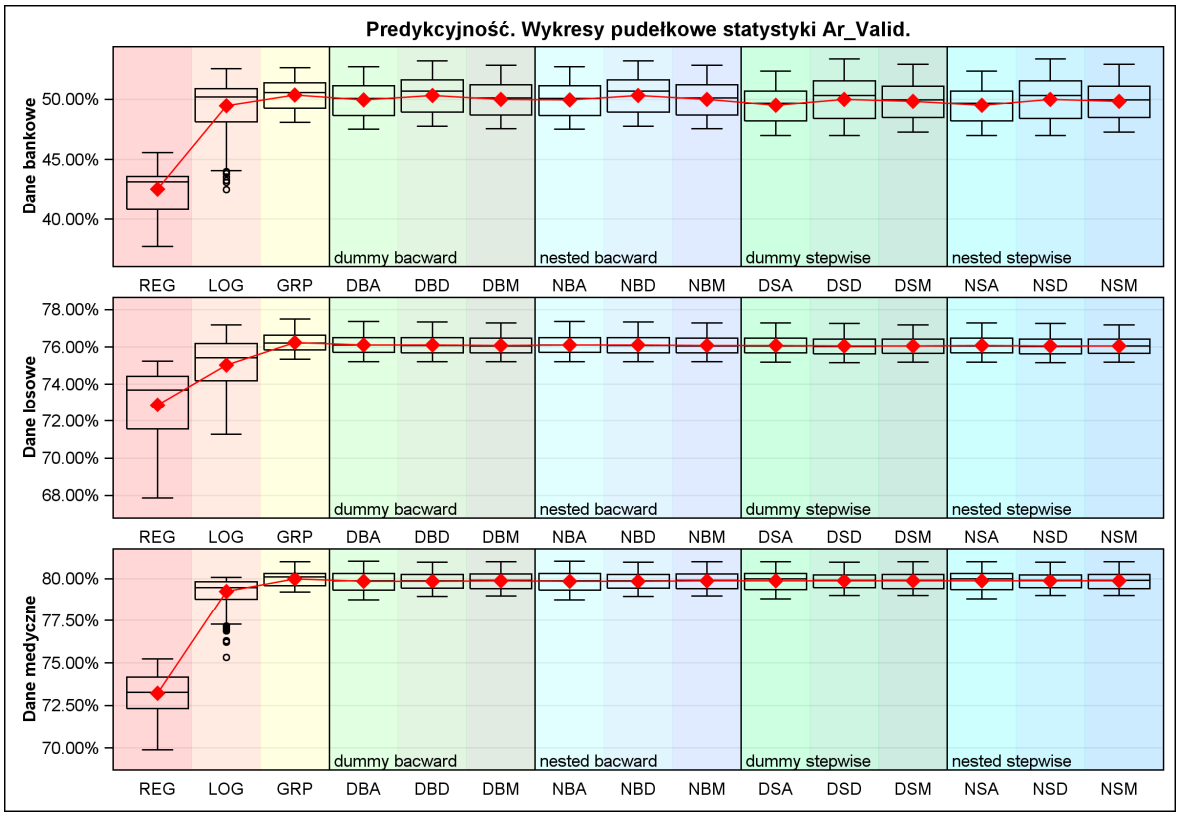
Źródło: Opracowanie własne.

4. Wyniki

Dla każdych z trzech danych: bankowych, losowych i medycznych wylosowano dane do zbiorów: treningowego i walidacyjnego, oraz przeprowadzono wstępną selekcję zmiennych, zob. tabela 7.

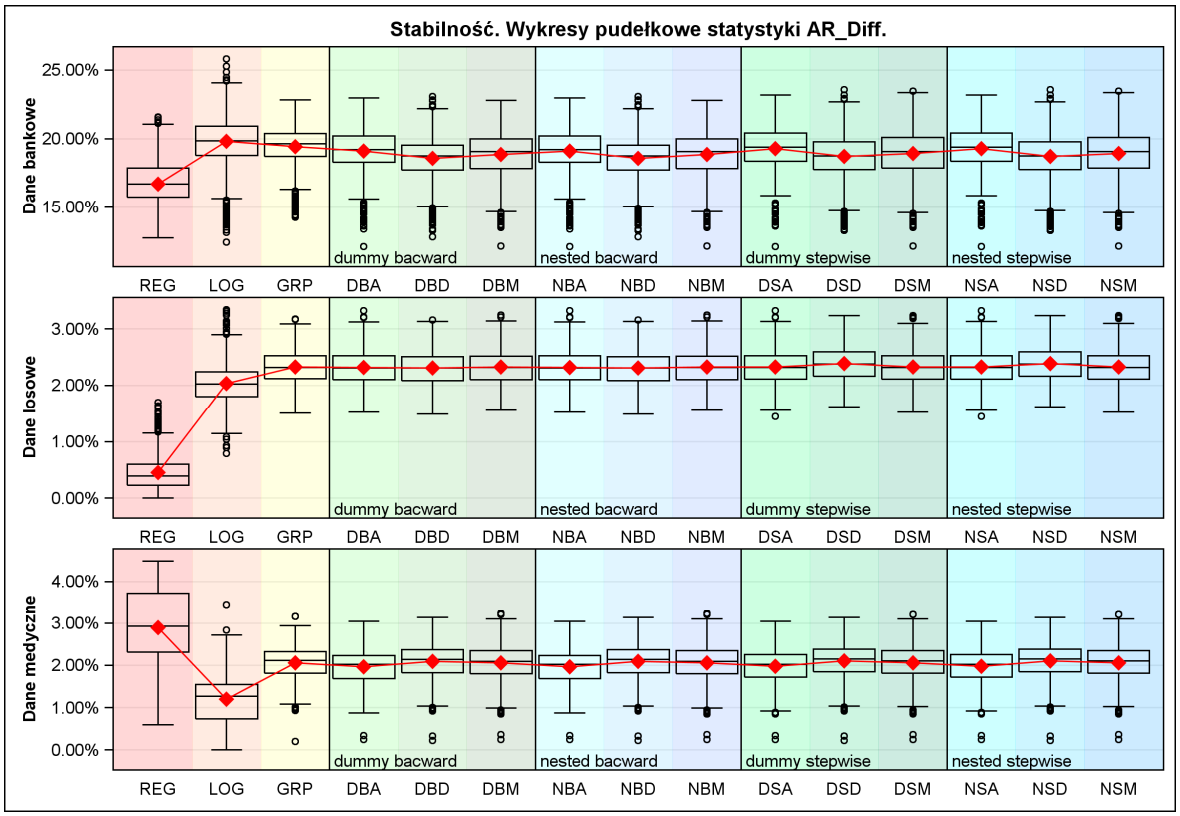
Następnie oszacowano 700 modeli, oddzielnie, dla kodowania REG i LOG. Następnie oszacowano 1 400 modeli dla kodowania GRP, które w kolejnym kroku korygowano 12 metodami. W sumie zbudowano i oceniono około 19 600 modeli dla trzech zbiorów danych oddzielnie (czyli w sumie 58 800 modeli). Tak duża liczba estymowanych modeli daje możliwość badania rozkładów współczynników wykorzystanych do ich oceny.

Wszystkie obliczenia wykonano na Laptopie Core Duo 1,67GHz. Obliczenia trwały 2 miesiące.



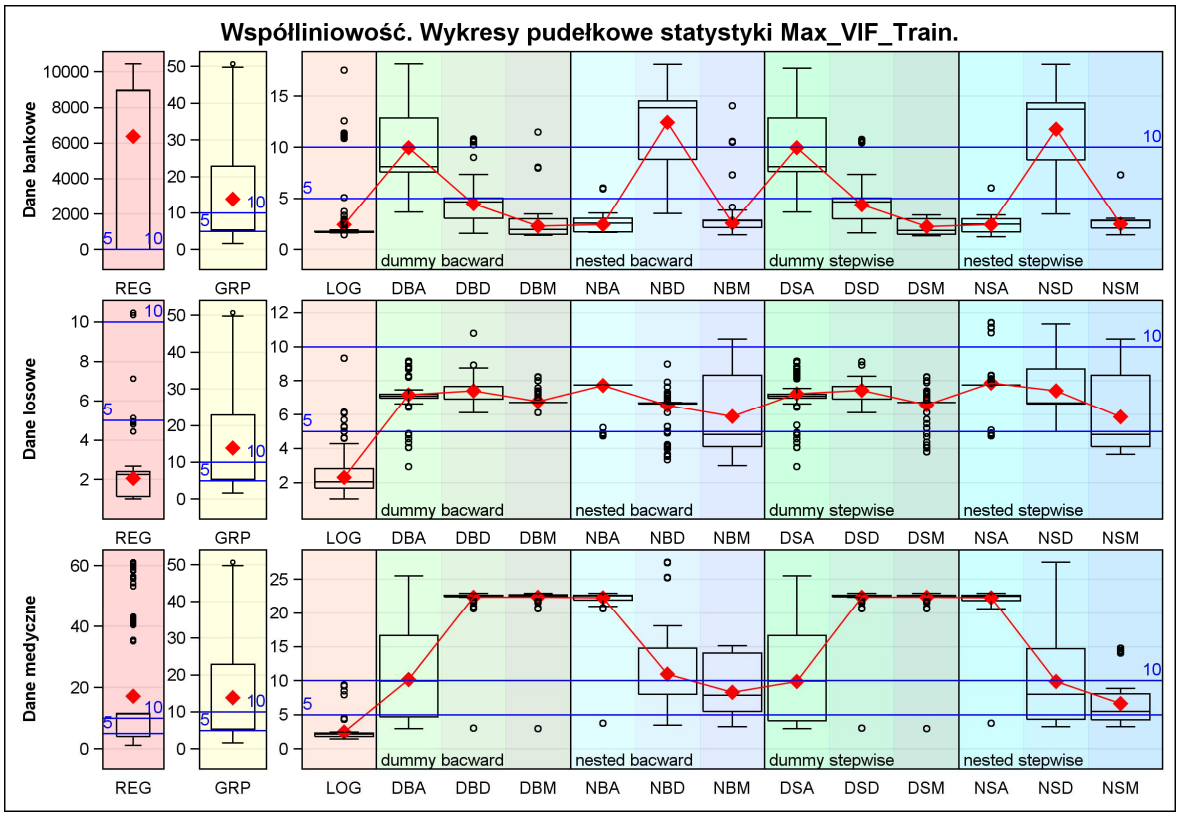
Rysunek 1. Rozkłady jednowymiarowe. Wykresy pudełkowe dla statystyki predykcyjności

Źródło: Opracowanie własne



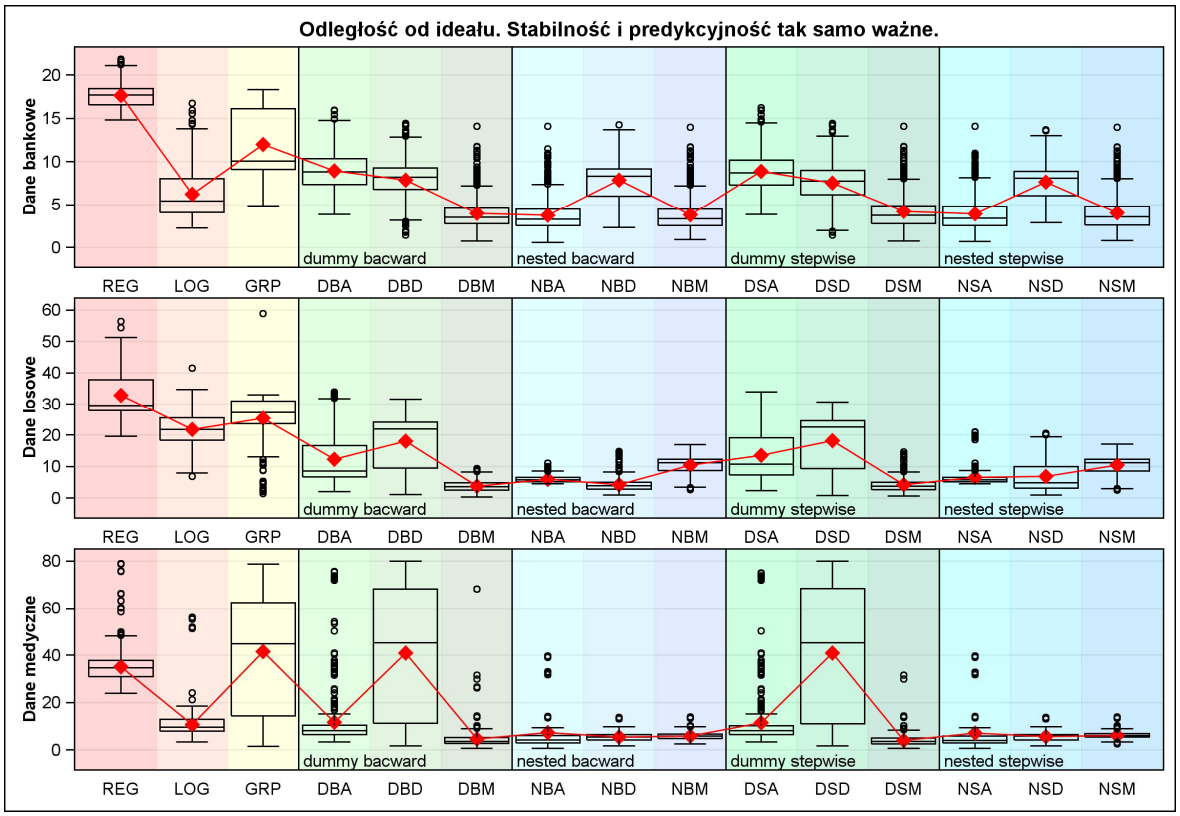
Rysunek 2. Rozkłady jednowymiarowe. Wykresy pudełkowe dla statystyki stabilności

Źródło: Opracowanie własne



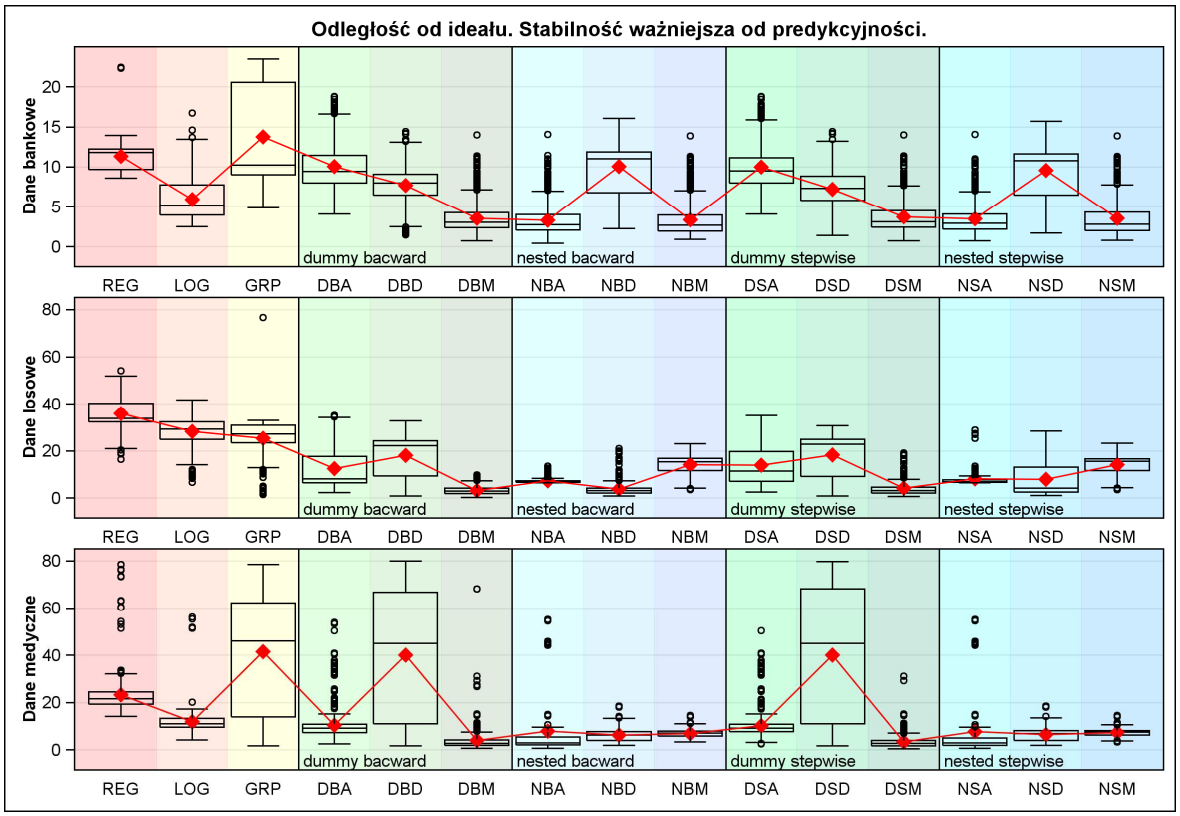
Rysunek 3. Rozkłady jednowymiarowe. Wykresy pudełkowe dla statystyki współliniowości

Źródło: Opracowanie własne



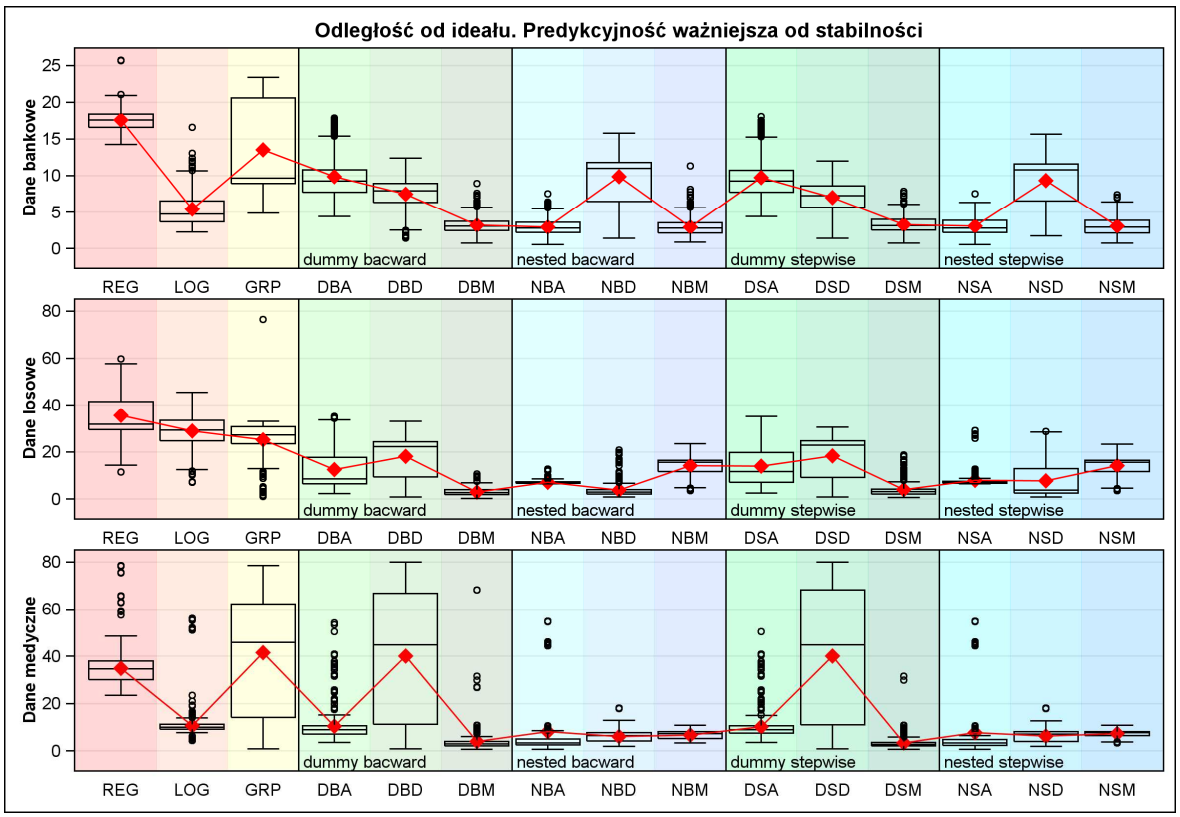
Rysunek 4. Ujęcie wielowymiarowe. Wykresy pudełkowe dla odległości od ideału. Stabilność i predykcyjność tak samo ważne

Źródło: Opracowanie własne



Rysunek 5. Ujęcie wielowymiarowe. Wykresy pudełkowe dla odległości od ideału. Stabilność ważniejsza od predykcyjności

Źródło: Opracowanie własne



Rysunek 6. Ujęcie wielowymiarowe. Wykresy pudełkowe dla odległości od ideału. Predykcyjność ważniejsza od stabilności

Źródło: Opracowanie własne

5. Interpretacja wyników

W celu zastosowania tej samej skali dla wszystkich sposobów postępowania z każdego zestawu modeli wybrano 700 najlepszych modeli ze względu na moc predykcyjną *(ang. Gini)* na zbiorze walidacyjnym.

Na rysunkach 1, 2 i 3 przedstawiono jednowymiarowe rozkłady wartości podstawowych kryteriów oceny modeli. Wyraźnie widać, że modele REG, LOG i GRP są zróżnicowane ze względu na moc predykcyjną. Wartości współczynników dla wszystkich modeli GRP i różne ich korekty mają podobne rozkłady. Analogicznie jest w przypadku kryterium stabilności AR_{diff}. Tylko rozkłady modeli REG, LOG i GRP różnią się między sobą. Modele GRP i wszystkie korekty mają bardzo podobne rozkłady stabilności. W przypadku współliniowości występują duże rozbieżności. Każdy rodzaj modeli ma inny rozkład współczynników. Warto podkreślić, że korekty modeli GRP mają zdecydowanie lepsze współczynniki MAX_{VIF}; dla modeli LOG wartości współczynników skupiają się w obrębie wartości akceptowalnych. W literaturze podaje się nierówność VIF > 5 jako granicę poziomu współliniowości [O'Brien R. M., 2007].

Znacznie lepszym sposobem porównywania modeli jest ich odległość od ideału. Tego typu porównanie przeprowadza się ze względu na wszystkie kryteria. Wprowadza się wagi, by zróżnicować moc kryteriów, zob. rysunki 4, 5 i 6. Im niżej położone są punkty na układzie współrzędnych, tym modele bliższe są ideałowi. Widać, że modele REG odbiegają od ideału dla każdego zestawu danych. Modele GRP charakteryzują się dużym rozrzutem. Modelom LOG odpowiadają pożądane rozkłady współczynników, choć nie zawsze wartości współczynników są najlepsze. Niektóre korekty modeli GRP mają najlepsze własności, szczególnie modele estymowane z wykorzystaniem kodowania kumulatywnego *(ang. nested).*

Zakończenie

Wyniki otrzymane na podstawie trzech różnych typów danych są podobne. Można więc analizować własności różnych sposobów postępowania na podstawie jednego wybranego zestawu danych, np. wygenerowanych losowo. Jedynym problemem jest czas obliczeń. Oszczędność czasu skłania do wykorzystania wygenerowanych danych losowych, gdyż zawsze są one dostępne. Możemy je też dowolnie modyfikować.

Można zatem odpowiedzieć na pytania postawione na początku artykułu. Choć zestawy danych różnią się od siebie, to można stworzyć repozytorium danych Credit Scoring na bazie generatora losowego i testować na tym zbiorze wiele różnych technik.

Bibliografia

Basel Committee on Banking Supervision, A Revised Framework, 2005, *International convergence of capital measurement and capital standards*, Updated November 2005, http://www.bis.org [dostęp: 01.05.2012].

Basel Committee on Banking Supervision, Working paper no. 14, 2005. *Studies on the validation of internal rating systems. Bank for International Settlements.*

Frątczak, E. red., 2012, *ZAAWANSOWANE METODY ANALIZ STATYSTYCZNYCH*, SGH.

Furnival, G.M., Wilson, R.W., 1974, *Regression by leaps and bounds*. Technometrics, 16 s.499-511.

Huang, E., 2007, *Scorecard specification, validation and user acceptance: A lesson for modellers and risk managers*, Credit Scoring Conference CRC, Edinburgh, http://www.crc.man.ed.ac.uk/publications/papers/ [dostęp: 01.05.2012].

Kadam, A., Delen, D., Walker, G., 2005, *Predicting breast cancer survivability: a comparison of three data mining methods*. Artificial Intelligence in Medicine, 34 s.113-127, http://seer.cancer.gov [dostęp: 01.05.2012].

Koronacki J., Mielniczuk J., 2001, *Statystyka dla studentów kierunków technicznych i przyrodniczych*, Wydawnictwo Naukowo-Techniczne.

Mays, E., 2009, *Systematic risk effects on consumer lending products*, Credit Scoring Conference CRC, Edinburgh, http://www.crc.man.ed.ac.uk/publications/papers/ [dostęp: 01.05.2012].

O'Brien R. M., 2007, *A Caution Regarding Rules of Thumb for Variance Inflation Factors*, Quality and Quantity 41(5) ss.673-690.

Przanowski, K., 2011, *Banking retail consumer finance data generator - credit scoring data repository*, Eprint-ArXiv, Q-Fin, w trakcie recenzji w czasopiśmie PAN FINANSE, http://arxiv.org/abs/1105.2968 [dostęp: 01.05.2012].

SAS Institute Inc., http://www.sas.com [dostęp: 01.05.2012].

Scallan, G., 2011, *Selecting characteristics and attributes in logistic regression*. Credit Scoring Conference CRC, Edinburgh, http://www.scoreplus.com/resources/reference/index.php [dostęp: 01.05.2012].

Siddiqi, N., 2005, *Credit risk scorecards: Developing and implementing intelligent credit scoring*, Wiley and SAS Business Series.

Thomas L. C., Edelman D. B., Crook J.N., 2002, *Credit Scoring and Its Applications*, Society for Industrial and Applied Mathematics, Philadelfia.

Welfe A., 2003, *Ekonometria*, Polskie Wydawnictwo Ekonomiczne.

Consumer finance data generator – method of Credit Scoring technique comparison

Summary: This paper presents a general idea of the comparison of different methods used for Credit Scoring techniques. Every scorecard can be constructed using various methods based on variable transformations in the logistic regression model, but to arrive at a comparison, with the corresponding proof that one technique is better than another presents a significant challenge due to the assessment being based on particular data. Similar results cannot be guaranteed in case of using new data from a new source. The following research hypothesis can therefore be formulated: how should a comparison be managed in order to get general results, not biased by particular data? A possible solution may be by the use of various random data generators. The data generator uses two approaches: transition matrix and scorings. Here both the results of the comparison methods and the methodology of the creation of these comparison techniques are presented. Before building a new model a modeler can conduct a comparison exercise to identify the best method in the case of the particular data. Here the various measures of predictive models such as: Gini – predictive power, Delta Gini – stability measure, VIF - collinearity and Max p-value - measure of variables significance, emphasizing the multi-criteria problem of “Good model” are presented. The suggested method can be especially useful in the model building process where there are defined complex criteria trying to cover the important problems of model stability through the cycle (TTC).