PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Beata Jackowska, Ewa Wycinka
Uniwersytet Gdański

MODELOWANIE RYZYKA WYSTĄPIENIA SZKODY UBEZPIECZENIOWEJ: BUDOWA I KRYTERIA OCENY MODELI REGRESJI LOGISTYCZNEJ

Streszczenie: W artykule podjęto próbę zbudowania modelu ryzyka wystąpienia szkody na polisie ubezpieczeniowej, wykorzystując w tym celu regresję logistyczną. Przyjęto hipotezę, że w sytuacji asymetrii informacji na rynku ubezpieczeń ryzyko związane z polisą nie jest addytywne ze względu na ryzyka cząstkowe. Z powodu specyfiki ryzyka ubezpieczeniowego, w tym małego prawdopodobieństwa wystąpienia szkody, model zbudowano na próbie zbilansowanej. Podczas budowy modelu uwzględniono różne sposoby kodowania zmiennych i ich wpływ na postać modelu. Przy wyborze postaci modelu wzięto pod uwagę miary dobroci dopasowania oraz miary zdolności predykcyjnej obliczone dla próby uczącej i testowej. Zaproponowany model pozwala na klasyfikację ubezpieczonych do jednorodnych grup ryzyka na podstawie cech kupowanej polisy.

Słowa kluczowe: regresja logistyczna, dychotomiczna zmienna zależna, ryzyko ubezpieczeniowe.

1. Wstęp

Podstawowym zadaniem zakładu ubezpieczeń jest przyjmowanie do ubezpieczenia ryzyk, ich selekcja i klasyfikacja na jednorodne grupy. Dla tak stworzonych grup ryzyka możliwa jest kalkulacja składki ubezpieczeniowej na podstawie danych statystycznych o realizacji ryzyka w poprzednich okresach. Współcześnie zakłady ubezpieczeń coraz częściej sprzedają ubezpieczenia kompleksowe w formie pakietów oferujących ubezpieczenie ryzyk związanych z osobą i majątkiem ubezpieczonego. Ocena ryzyka związanego z polisą ma więc kluczowe znaczenie dla zakładu ubezpieczeń. Liczba i rodzaj ryzyk objętych pakietem może mieć wpływ na ryzyko realizacji szkody w kompleksowej polisie. W artykule przyjęto hipotezę, że w sytuacji asymetrii informacji na rynku ubezpieczeń ryzyko związane z polisą nie jest addytywne ze względu na ryzyka cząstkowe.

Do szacowania prawdopodobieństwa wystąpienia szkody ubezpieczeniowej, pod warunkiem że zmienne objaśniające przyjmą określone wartości, zastosowanie znajdu-

ją uogólnione modele liniowe (GLM – *generalised linear models*) [Jong, Heller 2008]. Szczególne znaczenie w modelowaniu rozkładu zmiennej dychotomicznej ma model regresji logistycznej ze względu na jego walory interpretacyjne [McCullagh, Nelder, 1989]. W literaturze przedmiotu dotychczas niewiele uwagi poświęcano zastosowaniu uogólnionych modeli liniowych do szacowania ryzyka ubezpieczeniowego, a dla aplikacji ubezpieczeniowych niezbędne stają się podstawy teoretyczne metodologii.

W artykule podjęto próbę zbudowania modelu ryzyka wystąpienia szkody na polisie ubezpieczeniowej, wykorzystując w tym celu funkcję logistyczną. Podczas budowy modelu szczególną uwagę poświęcono specyfice danych ubezpieczeniowych. Sposób kodowania zmiennych i ich dobór ma bowiem wpływ na postać modelu. Przy wyborze postaci modelu uwzględniono miary dobroci dopasowania wartości empirycznych do teoretycznych oraz miary zdolności predykcyjnej.

Analizie poddano roczne polisy ubezpieczeniowe. Z badania wyeliminowano polisy o okresie krótszym niż jeden rok oraz umowy, które wygasły lub zostały zerwane przed upływem roku. Do budowy modelu wykorzystano podstawowe informacje o produkcie oraz dane osobowe dotyczące ubezpieczonych.

2. Przyjęte założenia

Jednostką statystyczną w badaniu jest polisa ubezpieczeniowa pewnego zakładu ubezpieczeń działu II. Z populacji wylosowano próbę zbilansowaną ze względu na proporcję realizacji ryzyka ubezpieczeniowego. Ponieważ odsetek szkód ubezpieczeniowych jest zwykle bardzo niski (poniżej 10%), więc wykorzystanie do budowy modelu próby niezbilansowanej powoduje, że w modelu przywiązywana jest mała waga do grupy jednostek, u których wystąpiła szkoda [Siddiqi 2006]. W celu uniknięcia „przeuczenia" modelu próba zbilansowana została podzielona na próbę uczącą i próbę testową w proporcji 2 : 1. Najlepsze modele zbudowane na próbie uczącej zostały ocenione na danych z próby testowej.

Zmienną zależną w modelu jest zmienna dychotomiczna Y o wartościach:

$$Y = \begin{cases} 1 & \text{szkoda ubezpieczeniowa} \\ 2 & \text{brak szkody ubezpieczeniowej} \end{cases},$$

natomiast zmienne niezależne mogą być ilościowe bądź jakościowe. Zmienne objaśniające zostały skategoryzowane według kryterium maksymalizacji miary *information value IV* (nazywanej także dywergencją Kullbacka-Leiblera). Miara ta informuje o sile predykcyjnej danej zmiennej, a obliczana jest jako suma po wszystkich wariantach cechy:

$$IV_i = \sum_{j=1}^{k_i} \left(p_{ij} - q_{ij} \right) \cdot \ln\left(\frac{p_{ij}}{q_{ij}} \right),$$

gdzie p_{ij} (q_{ij}) opisują rozkład prawdopodobieństwa i-tej zmiennej w grupie jednostek, u których nie wystąpiła szkoda (u których wystąpiła szkoda).

Wyodrębniono kategorie (warianty) cechy jednorodne ze względu na ryzyko. W przypadku cech ilościowych wyodrębniono przedziały wartości cechy, natomiast w przypadku cech jakościowych o wielu wariantach połączono wybrane kategorie (zgodnie ze znaczeniem merytorycznym) [Siddiqi 2006].

W celu wyodrębnienia predyktorów ryzyka wystąpienia szkody wykorzystano miarę siły predykcyjnej zmiennej objaśniającej *IV information value* oraz współczynnik kontyngencji *V* Cramera (tab. 1). Jako predyktory wybrano zmienne, dla których $IV \geq 0,05$ lub $V \geq 0,10$.

Tabela 1. Ranking predyktorów ryzyka wystąpienia szkody według miary *IV information value*

Lp.	Nazwa zmiennej	Opis zmiennej	*IV information value*	*V* Cramera
1	Grupa 3*	liczba produktów z grypy 3 (ubezpieczenia *casco* pojazdów)	0,65	0,39
2	Samochód	liczba produktów dotyczących ryzyka związanego z posiadaniem i użytkowaniem pojazdu	0,40	0,31
3	Pakiet	rodzaj pakietu A, B, …, J	0,31	0,28
4	Grupa 1*	liczba produktów z grupy 1 (ubezpieczenia wypadku)	0,26	0,25
5	Grupa 10*	liczba produktów z grupy 10 (ubezpieczenia odpowiedzialności cywilnej wynikającej z posiadania i użytkowania pojazdów)	0,22	0,23
6	Grupa 18*	liczba produktów z grupy 18 (ubezpieczenia tzw. *assistance*)	0,20	0,21
7	Dom	liczba produktów dotyczących ryzyka związanego z domem/mieszkaniem	0,15	0,20
8	Grupa 8*	liczba produktów z grupy 8 (ubezpieczenia szkód spowodowanych żywiołami)	0,13	0,18
9	Wiek	wiek ubezpieczonego	0,06	0,13
10	Grupa 13*	liczba produktów z grupy 13 (ubezpieczenia odpowiedzialności cywilnej ogólnej nieobjętej grupami 10-12)	0,06	0,13
11	Sprzedaż	kanał dystrybucji usług ubezpieczeniowych: sprzedaż bezpośrednia, sprzedaż poprzez agentów, multiagentów, brokerów	0,06	0,11
12	Grupa 2*	liczba produktów z grupy 2 (ubezpieczenia choroby)	0,03	0,09
13	Płeć	płeć ubezpieczonego	0,01	0,05
14	Grupa 9*	liczba produktów z grupy 9 (ubezpieczenia pozostałych szkód rzeczowych nieobjętych grupami 3-8, np. kradzież)	0,01	0,05

* Grupy ubezpieczeń działu II („Pozostałe ubezpieczenia osobowe oraz ubezpieczenia majątkowe") według: [Ustawa z dnia 22 maja 2003 r.]

Źródło: obliczenia własne za pomocą programu Statistica 8.0.

3. Wyniki estymacji modelu

Następnie zbudowano model regresji logistycznej, pozwalający na oszacowanie prawdopodobieństwa niewystąpienia szkody pod warunkiem, że zmienne niezależne przyjęły wartości $x_1, x_2, \ldots, x_k$,

$$p = P(Y = 0 \mid x_1, x_2, \ldots, x_k) = \frac{\exp\left(b_0 + \sum_{i=1}^{k} b_i x_i\right)}{1 + \exp\left(b_0 + \sum_{i=1}^{k} b_i x_i\right)}.$$

Parametry modelu $b_0, b_1, \ldots, b_k$ oszacowano metodą największej wiarygodności [Allison 2001; Harrell 2001; Kleinbaum 1994]. Model logitowy jest chętnie wykorzystywany przy analizie ryzyka wystąpienia badanego zdarzenia ze względu na możliwość interpretacji współczynników regresji za pomocą ilorazu szans. Szansę definiuje się jako stosunek prawdopodobieństwa danego zdarzenia do prawdopodobieństwa zdarzenia przeciwnego. Dla modelu logitowego szansa wyraża się wzorem:

$$\frac{p}{1-p} = \exp\left(b_0 + \sum_{i=1}^{k} b_i x_i\right).$$

Szansa dla jednostki, dla której wszystkie zmienne objaśniające są równe zeru, wynosi więc $\exp(b_0)$. Z powyższego wzoru wynika także, że przy wzroście wartości zmiennej objaśniającej x_i o jednostkę (przy stałych wartościach pozostałych zmiennych) iloraz szans wynosi $\exp(b_i)$. Fakt ten można wykorzystać przy porównaniu szans dla dwóch jednostek różniących się wartościami tylko jednej zmiennej.

Ze względu na to, że model był szacowany na podstawie próby zbilansowanej, skorygowano wyraz wolny, tak aby otrzymać oszacowanie prawdopodobieństwa w populacji niezbilansowanej [Siddiqi 2006, s. 68-69]. Zmienne objaśniające zostały zakodowane na dwa sposoby:

a) poprzez wprowadzenie zmiennych zero-jedynkowych − zmiennych pozornych (*dummy variables*) [Maddala 1983, s. 6-11],

b) wariantom cechy zostały przyporządkowane wartości miary *weight of evidence* (*WoE* obliczone na etapie kategoryzacji zmiennych), informującej o sile predykcyjnej poszczególnych wariantów cechy [Siddiqi 2006, s. 90-91]

$$WoE_{ij} = \left[\ln\left(\frac{p_{ij}}{q_{ij}}\right)\right] \cdot 100.$$

Do konstrukcji modelu wykorzystano metodę krokową postępującą (*stepwise logistic regression*) [Christensen 1997; Hosmer, Lemeshow 2000]. W wyniku różnego sposobu kodowania zmiennych objaśniających otrzymano dwa modele z różnymi predyktorami (tab. 2):

a) model ze zmiennymi objaśniającymi zero-jedynkowymi (6 zmiennych) oraz

b) model z dwiema zmiennymi ilościowymi o wartościach równych *weight of evidence* (*WoE*).

Tabela 2. Wyniki oszacowania modelu logitowego

Predyktor	Warianty	exp(*b*)	*p-value**
Model logitowy w przypadku zero-jedynkowych zmiennych objaśniających			
Wyraz wolny	*x*	0,6780	
Grupa 1	Grupa referencyjna: 0 produktów		
	1 produkt	0,7836	0,0000
	2 i więcej	1,0507	
Dom	Grupa referencyjna: 0 produktów		
	1 produkt	1,5391	0,0069
	2 i więcej	0,6649	
Grupa 3	Grupa referencyjna: 0 produktów		
	1 produkt	0,4845	0,0000
	2 i więcej	0,6914	
Model logitowy w przypadku zmiennych objaśniających przyjmujących wartości *WoE*			
Wyraz wolny	*x*	0,9771	
Pakiet	Grupa referencyjna: *WoE* = 0		
	x	1,0032	0,0000
Grupa 3	Grupa referencyjna: *WoE* = 0		
	x	1,0091	0,0000

* Wartość *p* dla testu ilorazu wiarygodności modelu zawierającego dany predyktor i predyktory z wierszy poprzedzających oraz wiarygodności tego modelu bez danego predyktora.

Źródło: obliczenia własne za pomocą programu Statistica 8.0.

Tabela 3. Kryteria wyboru modelu: kryteria informacyjne (*AIC*, *BIC*) oraz miary dobroci dopasowania (miary pseudo-R^2)

Modele	*AIC* (Akaike Information Criterion)	*BIC* (Bayesian Information Criterion)	R^2 McFadden	R^2 Cragg-Uhler (Naglekerke)	R^2 Cox-Snell
Model ze zmiennymi zero-jedynkowymi	1033,913	1067,022	0,121	0,206	0,154
Model ze zmiennymi o wartościach *WoE*	1033,893	1048,082	0,114	0,195	0,146

Źródło: obliczenia własne za pomocą programu Statistica 8.0.

Ilorazy szans dla każdej zmiennej pokazują, jak zmieni się szansa niewystąpienia szkody wraz ze wzrostem zmiennej o jednostkę. W modelu pierwszym najsilniejszy wpływ ma cecha „grupa 3". Włączenie do polisy jednego produktu z grupy 3 powoduje spadek szansy niewystąpienia szkody (wzrost ryzyka) o 51,55%. W modelu drugim ilorazy szans należy interpretować w ten sposób, że wzrost wartości *WoE* dla zmiennej „pakiet" o jedną jednostkę powoduje wzrost szansy niewystąpienia szkody o 0,32%, a dla zmiennej „grupa 3" − o 0,91%.

Przy wyborze modeli kierowano się minimalizacją kryterium informacyjnego AIC oraz BIC [Agresti 2002], a także dopasowaniem modeli mierzonym miarami pseudo-R^2 [Maddala 1983]. Miary dobroci dopasowania oraz kryteria informacyjne wskazują na bardzo małe różnice w dobroci dopasowania obu modeli (tab. 3), dlatego można je stosować zamiennie, w zależności od danych posiadanych przez zakład ubezpieczeń.

4. Ocena zdolności predykcyjnej modeli

Do oceny zdolności predykcyjnej modeli wykorzystano miarę *IV information value* (dywergencję Kullbacka−Leiblera), statystykę dywergencji (*divergence statistics*), statystykę Hosmera−Lemeshowa, wskaźniki zbudowane na podstawie krzywej *ROC* (*receiver operating characteristic*): *AUC* (*area under ROC*) pole powierzchni pod krzywą *ROC*, współczynnik Giniego (równy $2 \cdot AUC - 1$) [Hosmer, Lemeshow 2000].

Tabela 4. Porównanie miar zdolności predykcyjnej oszacowanych modeli

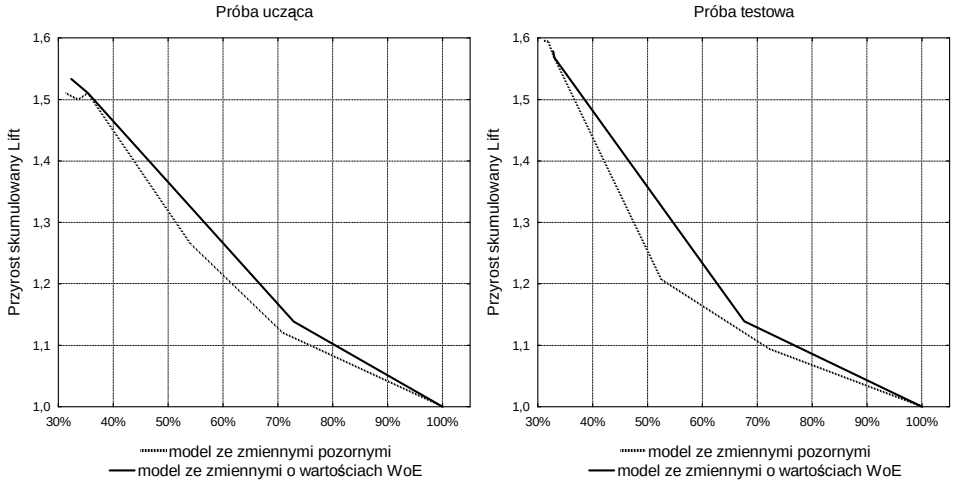
Modele	*IV information value*	Współczynnik Giniego	Statystyka dywergencji	Statystyka Hosmera−Lemeshowa	*AUC*
Wyniki dla próby uczącej					
Model ze zmiennymi o wartościach *WoE*	0,670	0,411	0,715	11,405 (p = 0,18)	0,706
Model ze zmiennymi zero-jedynkowymi	0,700	0,383	0,635	42,832 (p < 0,001)	0,692
Wyniki dla próby testowej					
Model ze zmiennymi o wartościach *WoE*	0,645	0,384	0,726	10,395 (p = 0,238)	0,692
Model ze zmiennymi zero-jedynkowymi	0,681	0,349	0,603	34,532 (p < 0,001)	0,675

Źródło: obliczenia własne za pomocą programu Statistica 8.0.

Miary zdolności predykcyjnej modelu również nie pozwalają na wybór lepszego z dwóch proponowanych modeli. Porównanie wyników dla próby uczącej i testowej wskazuje na stabilność modeli. Potwierdzenie tych wyników otrzymano także na podstawie skumulowanego przyrostu *Lift*, będącego ilorazem skumulowanego odsetka wystąpień szkody oraz skumulowanego odsetka polis ogółem w próbie uporządkowanej rosnąco według prawdopodobieństwa niewystąpienia szkody (porządek od najbardziej ryzykownych polis do najmniej ryzykownych). Przebieg wykresu skumulowanego przyrostu *Lift* (rys. 1) wskazuje nawet na lepsze własności prognostyczne modelu dla próby testowej.

Modele zostały zbudowane na próbie uczącej zbilansowanej, dzięki czemu uniknięto tendencji modelu do preferowania grupy bardziej licznej. Uzyskane modele są stabilne, moc predykcyjna w próbie testowej i uczącej nie różni się istotnie.

Dopasowanie modeli można uznać za zadowalające. Moc predykcyjna modeli badana na próbie rzeczywistej (niezbilansowanej) również nie zmalała, co zapowiada dobrą efektywność modeli w zastosowaniach praktycznych.



Rys. 1. Skumulowany przyrost *Lift*

Źródło: opracowanie własne za pomocą programu Statistica 8.0.

W zbiorze zaproponowanych zmiennych objaśniających znajdowały się zmienne silnie ze sobą skorelowane, jednak zastosowanie algorytmu regresji krokowej postępującej doprowadziło do tego, że w modelu znalazły się tylko zmienne objaśniające, które nie są ze sobą skorelowane. Zastosowanie dwóch różnych sposobów kodowania zmiennych objaśniających nie wpłynęło na dopasowanie i moc predykcyjną modeli, doprowadziło jednak do wyboru innego predyktora z pary zmiennych najsilniej skorelowanych.

5. Wnioski

W procesie underwritingu ubezpieczeniowego następuje ocena, klasyfikacja i selekcja ryzyk zgłaszanych do ubezpieczenia. Ocena ryzyka może być przeprowadzana w oparciu o wiedzę, doświadczenie i subiektywne odczucia underwritera lub przy zastosowaniu metod statystycznych. Metody te umożliwiają automatyzację procesu underwritingu, przyspieszają ocenę ryzyka i eliminują czynnik ludzki w procesie oceny. Mogą być wykorzystane do oceny ryzyk masowych, gdzie w oparciu o historię klientów (*underwriting behawioralny*) oraz dane zewnętrzne możliwe jest zidentyfikowanie ryzyk nieubezpieczalnych oraz zróżnicowanie składki dla ryzyk przyjętych do ubezpieczenia [Jong, Heller 2008].

Zaproponowany model pozwala na klasyfikację ubezpieczonych do jednorodnych grup ryzyka na podstawie cech kupowanej polisy. W przeprowadzonym badaniu wykazano, że poziom ryzyka związanego z polisą nie jest proporcjonalnie związany z liczbą produktów. Istotnymi predyktorami w pierwszym modelu (z zero-jedynkowymi zmiennymi objaśniającymi) okazały się tylko: fakt ubezpieczenia jednego z ryzyk związanych z domem/mieszkaniem, wykupienie AC oraz ubezpieczenia osobowego, a w drugim modelu (z ilościowymi zmiennymi objaśniającymi o wartościach *WoE*) wykupienie AC oraz rodzaj wybranego pakietu ubezpieczeń. Zaproponowane modele mogą posłużyć do konstrukcji systemu zniżek dla ubezpieczonych kupujących pakiety ubezpieczeniowe.

Literatura

Agresti A., *Categorical Data Analysis*, John Wiley & Sons, New Jersey 2002.

Allison P.D., *Logistic Regression Using the SAS System*: *Theory and Application*, SAS Institute and Wiley, Cary 2001.

Christensen R.Ch., *Log-linear Models and Logistic Regression*, Springer, New York 1997.

Harrell F., *Regression Modeling Strategies with Applications to Linear Models*, *Logistic Regression*, and Survival Analysis, Springer-Verlag, New York 2001.

Hosmer D., Lemeshow S., *Applied Logistic Regression*, John Wiley & Sons, New Jersey 2000.

Jong P. de, Heller G.Z., *Generalized Linear Models for Insurance Data*, Cambridge University Press, Cambridge 2008.

Kleinbaum D.G., *Logistic Regression. A Self-Learning Text*, Springer-Verlag, New York 1994.

Maddala G.S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge 1983

McCullagh P., Nelder J.A., *Generalized Linear Models*, Chapman & Hall, London 1989.

Siddiqi N., *Credit Risk Scorecards*. *Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, New Jersey 2006.

Ustawa z dnia 22 maja 2003 r. o działalności ubezpieczeniowej, DzU z 2003, nr 124, poz. 1154.

INSURANCE RISK MODELLING: CONSTRUCTION AND FITTING CRITERIA OF LOGISTIC REGRESSION MODELS

Summary: In the article an attempt to construct a model of insurance risk is made using a logistic regression. It is assumed that due to the information asymmetry on the insurance market the risk on the insurance policy is not additive with respect to partial risks. Owing to the specificity of insurance risk, particularly low probability of damage, the model was constructed on the balanced sample. Different methods of the variable coding and their influence on the model's content were taken into consideration. Optimal models were chosen on the basis of goodness of fit measures and power predictive measures. They were calculated for development and validation samples. The proposed model allows the classification of insured to homogeneous risk's groups with respect to insurance policy characteristics.