Uniwersytet Warszawski

Wydział Nauk Ekonomicznych

Krzysztof Szafrański

Nr albumu: 386 118

Determinanty niewypłacalności kredytowej klientów indywidualnych

Warszawa, czerwiec 2022

Wstęp

W dobie rosnących stóp procentowych wiele osób jest zagrożonych niewypłacalnością kredytową. Temat predykcji bankructwa czy niewypłacalności przedsiębiorstw cieszy się w literaturze dużą popularnością, jednak na poziomie klientów indywidualnych zagadnienie to jest rzadziej podejmowane. W relacji między bankami a kredytobiorcami nieunikniona jest silna asymetria informacji, która utrudnia bankom poprawne określenie ryzyka kredytowego. Zatem istotne wydaje się być pytanie jakie czynniki najlepiej mogą określać prawdopodobieństwo niewypłacalności.

Celem niniejszej pracy jest właśnie znalezienie zmiennych mających największe znaczenie dla predykcji niewypłacalności kredytowej klientów banków. Do analizy tego zagadnienia wykorzystano dane na temat posiadaczy kart kredytowych z Tajwanu. Ze względu na binarność analizowanej zmiennej zależnej w badaniu posłużono się modelem logit. Na podstawie przeglądu literatury spodziewa się, że zmienne finansowe będą miały większe znaczenie niż czynniki demograficzne.

Na początku pracy dokonano krótkiego przeglądu literatury. Następnie przedstawiono opis danych i sformułowano hipotezę. W ostatniej części przedstawiony zostanie wykonany model oraz interpretacja wyników. Pracę zakończono krótkim podsumowaniem.

Przegląd literatury

Ze względu na binarność zmiennej zależnej, w analizowanym temacie często używa się regresji logistycznej. Özdemir i Boran (2004) wykorzystują ją by zbadać przyczyny niewypłacalności klientów banków w Turcji. Wyniki ich pracy wskazują, że zmienne finansowe mają większe znaczenie dla oceny ryzyka kredytowego niż zmienne demograficzne. W szczególności istotna była stopa procentowa oraz długość okresu spłaty. Obie zmienne pozytywnie wpływały na ryzyko niespłacenia kredytu na czas.

Natomiast Jacobson i Roszbach (2004) korzystali z modelu Probit na przykładzie danych ze Szwecji. Wyniki ich badania wskazywały, że wraz z wiekiem maleje prawdopodobieństwo niespłacenia kredytu. Co zaskakujące, prawdopodobieństwo to rośnie wraz z większym dochodem. Autorzy zwracają uwagę, że być może osoby o wyższych dochodach posiadają inne cechy, które wiążą się z wyższym ryzykiem niewykonania zobowiązania. Ryzyko rośnie również w przypadku osób będących po rozwodzie, natomiast sama wysokość pożyczki okazała się być nieistotna, choć trzeba zwrócić uwagę, że w próbie dominowały głównie niewysokie pożyczki. Ponadto badanie sugeruje, że sposób w jaki banki przyznają kredyty nie pokrywa się ze strategią minimalizującą ryzyko niewypłacalności klientów.

Balina i Idasz-Balina (2021) podjęli ten temat na przykładzie klientów polskich banków spółdzielczych, również wykorzystując regresję logistyczną. Istotność zmiennych różniła się między analizowanymi bankami, jednak najważniejsze okazały się być zmienne charakteryzujące sytuację finansową kredytobiorców. Ponadto istotna była również historia współpracy między klientem a danym bankiem. To prowadzi autorów do wniosku, że bardziej lokalna działalność banków spółdzielczych może dawać im przewagę w ocenie ryzyka kredytowego, w porównaniu do większych banków komercyjnych.

Pomimo częstego wykorzystywania modelu logit i probit do analizowania tego tematu, warto zwrócić uwagę, że niektórzy badacze sugerowali, że czasem inne metody mogą spisywać się lepiej. Na przykład Kruppa, Schwarz i Arminger (2013) badali skuteczność predykcyjną różnych metod uczenia maszynowego. Wyniki ich analizy wskazywały, że metoda lasów losowych (*random forests)* może dawać lepsze wyniki niż klasyczny model logitowy. Z kolei Yu i in. (2010) sugerowali, że najlepszy może być algorytm SVM *support vector machine*, choć w ich badaniu regresja logistyczna wypadała niewiele gorzej.

Opis danych

Dane użyte do analizy został zaczerpnięte ze strony https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#. Zawierają one 30 000 obserwacji na temat klientów posiadających karty kredytowe w Tajwanie. Wśród dostępnych informacji są np. zmienne demograficzne, historia płatności czy wyciąg z rachunku. Zmienna zależna DEFAULT jest dwuwartościowa i wskazuje czy dany klient spłacił swoje zobowiązania (1 - niespłacenie zobowiązań, 0 - zobowiązania spłacone w terminie). Poniżej przedstawiono opis zmiennych objaśniających:

- LIMIT_BAL - wysokość przyznanego kredytu w dolarach tajwańskich
- SEX - płeć (1 - mężczyzna, 2 - kobieta)
- EDUCATION - wykształcenie (1 - podyplomowe, 2 - wyższe, 3 - średnie, 4 - inne)
- MARRIAGE - stan cywilny (1 - żonaty/zamężna, 2 - kawaler/panna, 3 - inne)
- AGE - wiek
- BILL_ATM1, … ,BILL_ATM6 - zmienne wskazujące wartość na wyciągu z karty kredytowej od października (BILL_ATM1) do kwietnia (BILL_ATM6)
- PAY_ATM1, …. ,PAY_ATM6 - miesięczna zapłacona opłata za korzystanie z karty kredytowej od października (PAY_ATM1) do kwietnia (PAY_ATM6)

Do analizy powyższych danych wykorzystana zostanie regresja logistyczna. Na podstawie literatury sformułowano następującą hipotezę badawczą:

Hipoteza: Zmienne finansowe (LIMIT_BAL, BILL_ATMX, PAY_ATMX) są istotniejsze w predykcji niewypłacalności niż zmienne demograficzne (SEX, EDUCATION, MARRIAGE, AGE).

Badanie empiryczne

W poniższej tabeli (Tabela 1.) przedstawiono wyniki przeprowadzonej regresji logistycznej. Okazuje się, że wraz ze wzrostem wysokości przyznanego kredytu maleje prawdopodobieństwo jego niespłacenia. Może się to wydawać nieintuicyjne, ponieważ wyższe kwoty powinny być cięższe do spłacenia. Jednak należy zwrócić uwagę, że większe kredyty są przyznawane osobom o lepszej zdolności kredytowej, co może tłumaczyć ujemny znak przy parametrze tej zmiennej.

Parametr jest także ujemny przy zmiennej SEX2, co oznacza, że kobiety średnio rzadziej mają problem z niewypłacalnością niż mężczyźni. W przypadku wykształcenia jedyną istotną kategorią jest ta wskazujące na wykształcenie "inne". Osoby z wykształceniem z tej kategorii mają mniejsze prawdopodobieństwo na niespłacenie kredytu niż osoby z wykształceniem podyplomowym, które jest poziomem bazowym.

Tabela 1. Wyniki regresji logistycznej

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.020e-01  8.164e-02  -7.374 1.66e-13 ***
LIMIT_BAL   -2.868e-06  1.497e-07 -19.159  < 2e-16 ***
SEX2        -1.783e-01  2.914e-02  -6.118 9.47e-10 ***
EDUCATION2   6.052e-03  3.383e-02   0.179 0.858023
EDUCATION3  -1.127e-02  4.508e-02  -0.250 0.802618
EDUCATION4  -1.253e+00  1.843e-01  -6.800 1.05e-11 ***
MARRIAGE2   -2.005e-01  3.293e-02  -6.088 1.14e-09 ***
MARRIAGE3   -1.951e-01  1.255e-01  -1.554 0.120120
AGE          3.516e-03  1.765e-03   1.992 0.046343 *
BILL_AMT1   -7.721e-06  1.206e-06  -6.402 1.53e-10 ***
BILL_AMT2    4.918e-06  1.534e-06   3.205 0.001349 **
BILL_AMT3    1.881e-06  1.333e-06   1.411 0.158192
BILL_AMT4    7.212e-07  1.347e-06   0.535 0.592464
BILL_AMT5    3.098e-06  1.475e-06   2.101 0.035648 *
BILL_AMT6    1.675e-06  1.152e-06   1.454 0.145872
PAY_AMT1    -2.596e-05  2.766e-06  -9.385  < 2e-16 ***
PAY_AMT2    -1.767e-05  2.449e-06  -7.218 5.26e-13 ***
PAY_AMT3    -7.174e-06  1.880e-06  -3.817 0.000135 ***
PAY_AMT4    -8.248e-06  1.893e-06  -4.358 1.31e-05 ***
PAY_AMT5    -4.953e-06  1.815e-06  -2.729 0.006355 **
PAY_AMT6    -1.771e-06  1.274e-06  -1.390 0.164639
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Źródło: opracowanie własne.

W przypadku zmiennej wskazującej na stan cywilny poziomem bazowym jest związek małżeński. Zatem osoby klasyfikowane jako kawaler/panna z mniejszym

prawdopodobieństwem okażą się niewypłacalne. Ponownie może się to wydawać nieintuicyjne, ale prawdopodobnie ludzie będący w związku małżeńskim mają też większe potrzeby kredytowe niż single, co może prowadzić do większego zadłużania się.

Ostatnią zmienną demograficzną jest wiek, który jest istotny na poziomie 1%. Pozytywny parametr sugeruje, że wraz z wiekiem rośnie prawdopodobieństwo niespłacenia kredytu.

Wyższe wartości wyciągu z karty kredytowej w październiku (BILL_AMT1) wiążą się z mniejszym prawdopodobieństwem niewypłacalności w następnym miesiącu. Może to sugerować, że osoby bardziej przekonane o swojej wypłacalności pozwalały sobie na większe wydatki. Jednak wyższy wyciąg z sierpnia (BILL_ATM2) wiązał się z wyższym prawdopodobieństwem niespłacenia kredytu. Wyciągi z bardziej odległych miesięcy okazały się być nieistotne.

Prawie wszystkie zmienne wskazujące na płatności za korzystanie z karty kredytowej z poprzednich miesięcy są istotne, a parametry przy nich są ujemne. Zatem wyższe płatności w poprzednich miesiącach są powiązane z mniejszym prawdopodobieństwem niewypłacalności. Do wyników przy powyższych zmiennych finansowych należy jednak podchodzić z rezerwą, ponieważ ich wartości są w pewnym stopniu skorelowane.

Wszystkie zmienne demograficzne okazały się być istotne w modelu, zatem nie można powiedzieć, że mają one mniejsze znaczenie niż zmienne finansowe, co prowadzi do odrzucenia postawionej wcześniej hipotezy.

Niestety jakość predykcyjna modelu nie jest dobra. Zakładając punkt odcięcia równy 0,5 wszystkie obserwacje są klasyfikowane do kategorii “0”, czyli wypłacalny. Może to wynikać ze znacznego niezbalansowania danych, ponieważ obserwacje, w których doszło do niewypłacalności stanowią tylko 20% wszystkich obserwacji.

Zdecydowano się więc przeprowadzić oversampling, aby wyrównać liczebności obu klas zmiennej zależnej. Dzięki temu predykcja modelu nieco się poprawiła, choć wciąż nie jest ona zadowalająca, co zostało przedstawione w poniższej macierzy błędów (Tabela 2.). Tylko 61% obserwacji zostało zaklasyfikowanych poprawnie. Wrażliwość (prawdopodobieństwo, że osoba, która spłaci swoje zobowiązania zostanie zaklasyfikowana poprawnie) wyniosła 0,65, natomiast specyficzność (prawdopodobieństwo, że osoba, która będzie niewypłacalna zostanie zaklasyfikowana poprawnie) jest równa 0,59.

Tabela 2. Macierz błędów klasyfikacji modelu

	zaobserwowane		
prog nozo wane		0	1
	0	11 816	11 548
	1	6 353	16 964

Źródło: opracowanie własne

Podsumowanie

Celem niniejszej pracy było określenie jakie czynniki w największym stopniu wpływają na prawdopodobieństwo niewypłacalności kredytowej klientów indywidualnych. Wszystkie zmienne użyte w badaniu okazały się być istotne, zatem nie można powiedzieć, że zmienne finansowe mają większe znaczenie niż te demograficzne, jak sugerowano w niektórych badaniach (Özdemir i Boran, 2004; Balina i Idasz-Balina, 2021).

Ponadto jakość predykcyjna modelu logit okazała się być dość słaba, ponieważ tylko 61% obserwacji zostało zaklasyfikowanych do poprawnej grupy. Może to być spójne z wnioskami niektórych autorów (Kruppa, Schwarz i Arminger, 2013; Yu i in., 2010), którzy wskazywali, że w badaniu predykcji niewypłacalności efektywniejsze mogą być inne modele, na przykład metody uczenia maszynowego.

Literatura

Balina, R., & Idasz-Balina, M. (2021). Drivers of Individual Credit Risk of Retail Customers—A Case Study on the Example of the Polish Cooperative Banking Sector. *Risks*, *9*(12), 219.

Jacobson, T., & Roszbach, K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of banking & finance*, *27*(4), 615-633.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, *40*(13), 5125-5131.

Özdemir, Ö., & Boran, L. (2004). *An empirical investigation on consumer credit default risk* (No. 2004/20). Discussion Paper.

Yu, H., Huang, X., Hu, X., & Cai, H. (2010, October). A comparative study on data mining algorithms for individual credit risk evaluation. In *2010 International Conference on Management of e-Commerce and e-Government* (pp. 35-38). IEEE.