Paweł Marcinkowski[1]

Krakowskie Studia Małopolskie
2023, nr 3 (39)

Analiza dotycząca możliwości budowy modelu ryzyka kredytowego

Wstęp

FinAi przeprowadziło projekt budowy innowacyjnego modelu ryzyka kredytowego w oparciu o dane alternatywne takie jak dane z portali społecznościowych (Facebook, Linkedin), czy dane z telefonów komórkowych. Projekt ten był współfinansowany ze środków Europejskiego Funduszu Rozwoju Regionalnego w ramach umowy o dofinansowanie zawartej przez FinAi S.A. z Narodowym Centrum Badań i Rozwoju z siedzibą w Warszawie, pt.: „Modelowanie ryzyka kredytowego w oparciu o dane dostępne w kanałach cyfrowych z wykorzystaniem zaawansowanych rozwiązań teorio-grafowych związanych z sieciami społecznościowymi". Pierwszym etap projektu polegał na zebraniu danych i wykorzystaniu ich do budowy grafu powiązań pomiędzy klientami. Dodatkowo na podstawie danych z BIK i danych zarządzanych przez FinAi zbudowano flagi mające na celu nauczenie pod ich nadzorem modelu predykcyjnego wykorzystujące struktury grafowe. Zbudowany w taki sposób model miał charakteryzować się większą mocą predykcyjną niż standardowe modele stosowane w branży.[2]

Niniejszy artykuł dotyczy dwóch kwestii: możliwości budowy modelu ryzyka kredytowego w oparciu o pozyskane dane zgodnego z powyższymi założeniami oraz oceny jakości danych pochodzących ze źródeł alternatywnych. W ramach procesu pobrano dane na temat ryzyka kredytowego od ok 14 tysięcy klientów. Znacząco odbiega to od deklaro-

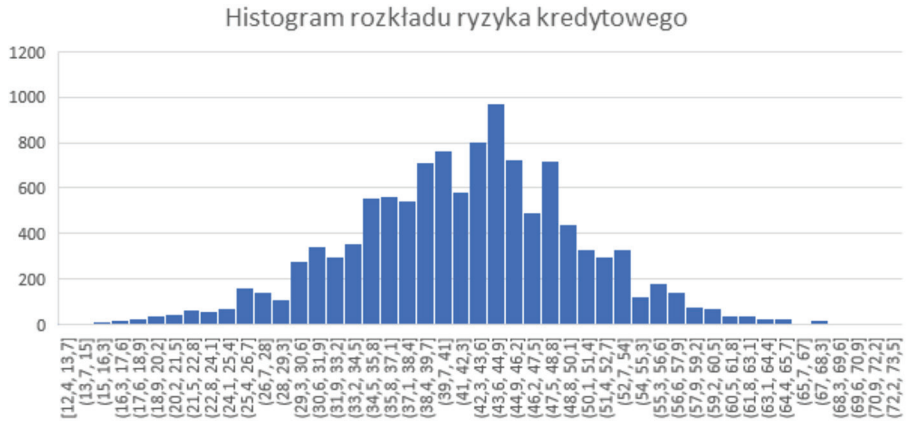
[1] Allegro Pay sp. z o.o.; pawel.marcinkowski@allegro.com; ORCID ID: https://orcid.org/0009-0009-2082-9442.

[2] N. Chen, B. Ribeiro, A. Chen, *Financial credit risk assessment: a recent review*, „Artif Intell. Rev." 2016, 45, s. 1–23.

wanego celu 100 000 rekordów, nie został również osiągnięty obniżony cel zebrania 20 000 rekordów. Taka liczebność próby w przypadku wysokiej jakość zebranych danych, nadal potencjalnie umożliwiłaby zbudowanie stabilnego modelu ryzyka kredytowego. Niestety analiza zbioru zawierającego dane na temat ryzyka kredytowego wskazuje na jego niereprezentatywność.

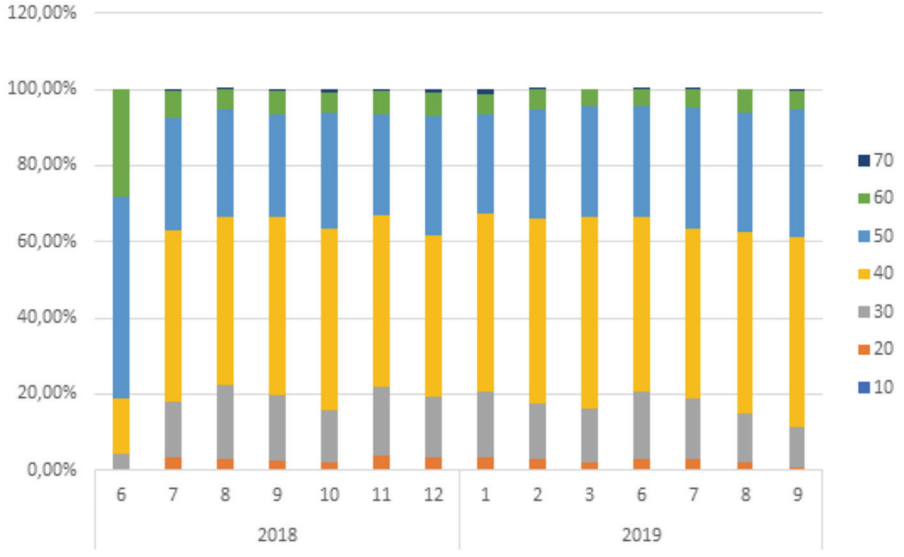
Badając napływające zagregowane dane BIK o ryzyku kredytowym, zauważyliśmy, że średnia ocena kredytowa dla klientów dla których zebrane były dane w ramach procesu (420) znacząco odbiega od średniej dla całej populacji (ok 520, średnia wyliczona i zaraportowana przez BIK). Ponadto, w zależności od metodologii banku, punkt odcięcia, poniżej którego klient nie ma możliwości ubiegania się o kredyt konsumencki, zawiera się w przedziale od 450 do 500 punktów BIK. Jednoznacznie wskazuje to na niereprezentatywność zebranej próby na potrzeby budowy modelu ryzyka. Tak duża rozbieżność uniemożliwia odpowiednią kalibrację modelu. Model ryzyka zbudowany w oparciu o takie dane mógłby działać poprawnie jedynie w zakresie klientów o niskim score, czyli wysokim ryzyku kredytowym. Zbudowany model byłby zatem obciążony niereprezentatywnością próby, co przekładałoby się na nadawanie z zasady niskiego scoringu. Stąd też nie można oczekiwać od takiego model wysokiej mocy predykcyjnej dla klientów bez oceny ryzyka kredytowego.

Poniżej przedstawiamy histogram rozkładu ryzyka kredytowego (score).



Wykres nie wskazuje na to, żeby dane były zaburzone. Histogram dość wyraźnie koncentruje się wokół średniej, wokół której jest symetryczny. Oznacza to, że dane zebrane w ramach projektu w sposób systematyczny pochodziły od niereprezentatywnej populacji, klientów o wysokim ryzyku kredytowym, co jak opisano wyżej utrudnia budowę wiarygodnego modelu.

Poniżej prezentujemy wykres rozkładu score w czasie. Jest on wyraźnie stabilny, co wzmacnia tezę o systematycznym pobieraniu danych głównie dla klientów o wysokim ryzyku kredytowym



Poniższa tabela przedstawia procentowy udział klientów w stanie default w kolejnych miesiącach. Jak widać udział ten utrzymuje się na poziomie 10% lub wyższym przez cały okres zbierania danych. Ponownie wskazuje to na nadreprezentatywność klientów o wyższym ryzyku kredytowym.

rok	miesiąc	% default
2018	6	8,7%
2018	7	17,0%

rok	miesiąc	% default
2018	8	12,9%
2018	9	12,1%
2018	10	9,3%
2018	11	11,2%
2018	12	17,4%
2019	1	18,7%
2019	2	9,9%
2019	3	7,4%
2019	4	13,2%
2019	5	12,7%
2019	6	15,3%
2019	7	13,6%
2019	8	7,6%
2019	9	7,2%

Dla części użytkowników niedostępne były dane dotyczące ryzyka kredytowego poniższa tabela przedstawia miesięczne zestawienie wszystkich klientów i klientów bez scoringu.

rok	miesiac	null_count	count
2018	4	2	18
2018	5	2	19
2018	6	9	115
2018	7	103	595
2018	8	133	728
2018	9	125	605
2018	10	230	986
2018	11	120	749
2018	12	166	759
2019	1	129	668
2019	2	165	746

rok	miesiac	null_count	count
2019	3	167	880
2019	4	347	1835
2019	5	462	2518
2019	6	443	1891
2019	7	533	2907
2019	8	95	726
2019	9	27	278
2019	10	0	2

Źródła danych

Pozostałe źródła danych obejmowały w początkowej fazie projektu, dane z Facebooka (jedynie imię, nazwisko i adres e-mail, które były odpowiednio hashowane), dane z portalu LinkedIn (również podstawowe informacje o profilu), listę połączeń i smsów, odczyty GPS[3], dane dotyczące zatrudnienia. Niestety w początkowej fazie projektu Google ograniczył dostęp do listy połączeń i smsów. Literatura wskazuje, że dane takie mają duży potencjał do budowy grafów, które zawierałyby informacje istotne ze względu na modelowanie ryzyka kredytowego. W naturalny sposób grafy te zawierałyby komponentę czasową, umożliwiającą wykorzystanie grafów dynamicznych w analizie danych.

Poniżej przedstawiamy liczności użytkowników, dla których zebrane zostały dane w pierwszej fazie konkursu.

Rok	Facebook	Pracodawca	Połączenia	Kontakty	SMS	GPS	Score	LinkedIn
2018	4 108	1 813	164	174	144	185	501	45
2019	4 012	1 803	187	205	176	215	505	25
Suma	8 120	3 616	351	379	320	400	1006	70

Jak widać, ze względu na niską konwersję użytkowników (niewielki odsetek klientów udostępniał dane na temat połączeń, SMS i GPS) ze-

[3] S. Zamore, L.A. Beisland, R. Mersland, *Geographic diversification and credit risk in microfinance*, „Journal of Banking & Finance" 2019, Vol. 109.

brano niewystarczającą liczbę rekordów do tego, aby móc skorzystać z tych informacji do budowy grafów.

W kolejnej fazie zbierano dane w ramach procesu kredytowego FinAi, ponieważ dane z Facebooka oraz dane na temat połączeń nie były dostępne, nie zbierano ich, zakres danych został natomiast poszerzony o dane zbierane w ramach wniosku kredytowego oraz wyciągu z konta bankowego. Odpowiednie liczności są zaprezentowane w poniższej tabeli. Dane zawierają również rok 2018, ponieważ zdecydowano się skorzystać ze wszystkich danych zebranych w ramach działalności FinAi.

Rok	Wniosek	Wyciąg	Score	Kontakty	GPS
2018	11 227	2 937	3 721	0	0
2019	34 786	9 088	9 615	3 798	1 220
Suma końcowa	46 013	12 025	13 336	3 798	1 220

Z całą pewnością dane pochodzące z wniosku i wyciągu bankowego niosą istotną informację o ryzyku kredytowym klienta i można na ich podstawie zbudować predykcyjny model. Jednakże dane te są standardowo wykorzystywane przez banki oraz inne instytucje finansowe w ocenie ryzyka kredytowego. W związku z tym model budowany wyłącznie w oparciu o te dane nie byłby innowacyjny, nie mógłby też konkurować z modelami działającymi w bankach, które posiadają dużo większe zbiory danych, oraz posiadają dokładniejszą informację na temat ryzyka kredytowego swoich klientów. W związku z taką strukturą pozyskanego zbioru danych, jedyną możliwością budowy innowacyjnego modelu ryzyka kredytowego w oparciu o teorię grafów, była budowa grafów na podstawie książki adresowej z telefonu (Kontakty) oraz ewentualne wykorzystanie danych na temat lokalizacji i danych adresowych. Ponieważ dane z GPS dostępne są dla niewielkiego odsetka klientów, nie pozwalają na wykorzystanie ich w budowie modelu. Aby zbadać możliwość budowy innowacyjnego modelu ryzyka kredytowego, poniżej zbadany został graf zbudowany na podstawie kontaktów z książki adresowej oraz możliwość wykorzystania danych o adresie zamieszkania.

Budowa grafu i zmiennych grafowych

Jedną z głównych innowacyjnych cech niniejszego projektu jest wykorzystanie grafów dynamicznych[4] (zmiennych w czasie) oraz zmiennych opartych o strukturę grafów społecznościowych do modelowania ryzyka kredytowego. Główna hipoteza badawcza postawiona na początku projektu brzmiała następująco: istnieją grafy dynamiczne oparte o sieci społecznościowe, odzwierciedlające ryzyko kredytowe.

Hipoteza taka pojawia się obecnie w literaturze statystycznej i ekonomicznej (10)[5] i wiąże się z coraz bardziej powszechnym przekonaniem o wartości danych z alternatywnych źródeł w modelowaniu ryzyka kredytowego, szczególnie dla klientów detalicznych. Celem projektu jest identyfikacja takich grafów z możliwie szerokiego spektrum różnych sieci powiązań. Pozyskanie odpowiedniego zbioru danych okazało się dużo trudniejsze niż pierwotnie zakładano (zaprzestanie udostępniania danych przez Facebooka i Google w związku z aferą Cambridge Analytica, niechęć klientów do udostępniania swoich danych), możliwości budowy potencjalnie wartościowych grafów okazały się istotnie ograniczone. W szczególności: nie była możliwa budowa grafów znajomości w oparciu o znajomość w serwisach społecznościowych; oraz nie była możliwa budowa grafów opartych o tzw. homofilię, czyli podobieństwo na podstawie np. zainteresowań/polubień.

Wyniki literaturowe wskazują, że potencjalnie dużą wartość informacyjną może nieść graf zbudowany na podstawie połączeń telefonicznych oraz SMS, jednak w trakcie trwania projektu takie dane również przestały być udostępniane na systemie Android.

---

[4] J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins, *Microscopic evolution of social networks*, [w:] In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, 2008, s. 462–470; R. Muñoz-Cancino, C. Bravo, S.A. Ríos, M. Graña, *On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance*, „Expert Systems with Applications" 2023, Vol. 218.

[5] M. Óskarsdóttir et al., *The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics*, „Applied Soft Computing" 2019.

W celu oceny możliwości wykorzystania oraz użyteczności zmiennych grafowych zbudowanych na podstawie dostępnych danych, na potrzeby niniejszego Raportu zbudowano graf nieskierowany na podstawie danych o kontaktach z książek adresowych. Dokładniej: wierzchołki reprezentują osoby, które udostępniły dane z książki adresowej, a krawędź pomiędzy wierzchołkami (osobami) jest obecna wtedy i tylko wtedy, gdy istnieje wspólny kontakt w ich książkach adresowych.

W celu ograniczenia występowania krawędzi wynikających np. z faktu posiadania numeru skrzynki pocztowej lub numeru call center operatora w kontaktach, zdecydowano się usunąć z książek adresowych wszystkie numery telefonów, których liczba wystąpień przekraczała liczbę 10 (wartość ustalona ekspercko). W ocenie autorów raportu jest to jedyny graf którego wykorzystanie jest uzasadnione merytorycznie, który może być utworzony na podstawie dostępnego zbioru danych.

Pierwotnie rozpatrywano graf, w którym krawędź pomiędzy dwoma wierzchołkami/osobami była obecne tylko w przypadku, gdy jedna z tych osób posiadała drugą w swojej książce adresowej. Tak skonstruowany graf zawierał jednak zbyt mało krawędzi (kilkaset krawędzi przy kilku tysiącach wierzchołków) by przeprowadzona na jego podstawie analiza mogła mieć wartość merytoryczną.

Utworzono graf w trzech konfiguracjach:

- G1 – zbiór wierzchołków zawierał wszystkie osoby, które udostępniły dane z książki adresowej,
- G2 – graf indukowany (ograniczony) przez wierzchołki/osoby, dla których jest dostępna informacja o ryzyku kredytowym,
- G3 – podgraf grafu G1, w którym zachowane są tylko krawędzie, dla których u co najmniej jednego z wierzchołków sąsiadujących dostępna jest informacja o ryzyku kredytowym, a wierzchołki izolowane bez danych kredytowych są usunięte.

Tak skonstruowane grafy niestety nie posiadają wymiaru czasowego, nie są więc grafami dynamicznymi. Wymiar ten mógłby być uwzględniony np. poprzez wykorzystanie daty dodania kontaktu do książki adresowej, jednak informacja ta nie jest dostępna w systemie Android. Dodatkowo w naszej ocenie, data dodania kontaktu nie jest zmienną niosącą istotną informację.

Poniżej przedstawione są definicje statystyk grafów oraz ich wartości dla grafów G1, G2 i G3.

- Średnica grafu – największa długość najkrótszej ścieżki pomiędzy wierzchołkami w grafie,
- Średnia odległość w grafie – średnia z długości najkrótszych ścieżek pomiędzy wierzchołkami w grafie,
- Stopień wierzchołka grafu – liczba krawędzi wychodzących z danego wierzchołka,
- Średni stopień w grafie – średnia ze stopni wszystkich wierzchołków w grafie,
- Składowa grafu – (spójny) podgraf wyjściowego grafu, w którym istnieje ścieżka pomiędzy każdą parą wierzchołków.

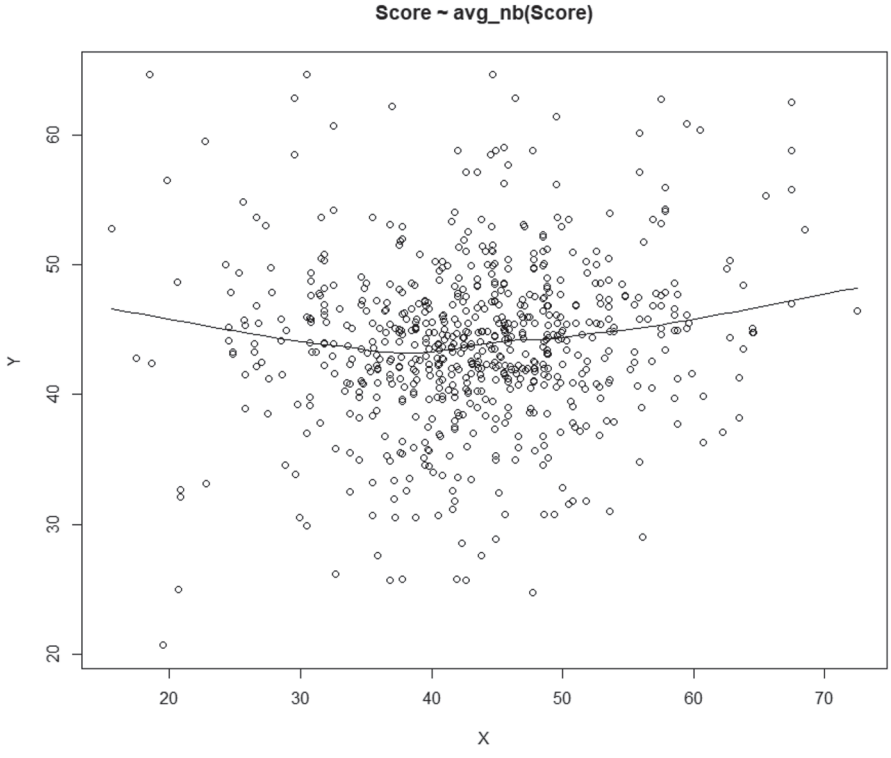
Graf	G1	**G2**	**G3**
Liczba wierzchołków	3 781	864	3 191
Liczba krawędzi	21 452	2 254	11 519
Średnica grafu	9	11	10
Średnia odległość w grafie	3,68	4,03	3,95
Średni stopień w grafie	11,35	5,22	7,22
Składowe grafu	14 składowych, z czego jedna duża zawierająca 3755 wierzchołków oraz 13 składowych dwuwierzchołkowych.	26 składowych, z czego jedna duża zawierająca 806 wierzchołków oraz 25 składowych dwu – lub trzy-wierzchołkowych	8 składowych, z czego jedna duża zawierająca 3176 wierzchołki oraz 7 składowych dwu – lub trzy-wierzchołkowych

Wszystkie skonstruowane grafy można uznać za niemalże spójne. Po ograniczeniu się do wierzchołków z informacją kredytową, rozmiar grafu istotnie się zmniejsza, przy czym zdecydowana większość wierzchołków bez informacji kredytowej jest połączona z co najmniej jednym z wierzchołków zawierających taką informację. Powyższych statystyki świadczą o tym, że udało się uzyskać pewną złożoną strukturę grafową. Nie ma podstaw do twierdzenia, że tak otrzymane grafy nie mogą być wykorzystane do budowy modelu ryzyka kredytowego.

W celu oceny hipotezy, że tak otrzymany graf ma wartość informacyjną w problemie oceny ryzyka kredytowego, tzn. czy niesie ze sobą infor-

macje o korelacji ryzyka kredytowego, przeprowadzono analizę regresji liniowej[6] , gdzie jest stanowi Score wierzchołka, a jest średnią ze scorów wszystkich jego sąsiadów. Jeśli współczynnik będzie statystycznie różny od zera, będzie to przesłanka za hipotezą, że graf ten niesie ważne informacje na temat ryzyka kredytowego. Na potrzeby analizy ograniczono się do podgrafu grafu G2, z którego usunięto wierzchołki znajdujące się w stanie niewypłacalności (*default*). Dla wierzchołków w stanie niewypłacalności została przeprowadzona osobna analiza opisana w dalszej części raportu.

Poniżej prezentujemy wykres , na którym ciężko doszukać się istotnej zależności.



[6] D.C. Montgomery, E.A. Peck, G.G. Vining, *Introduction to linear regression analysis*, 2021.

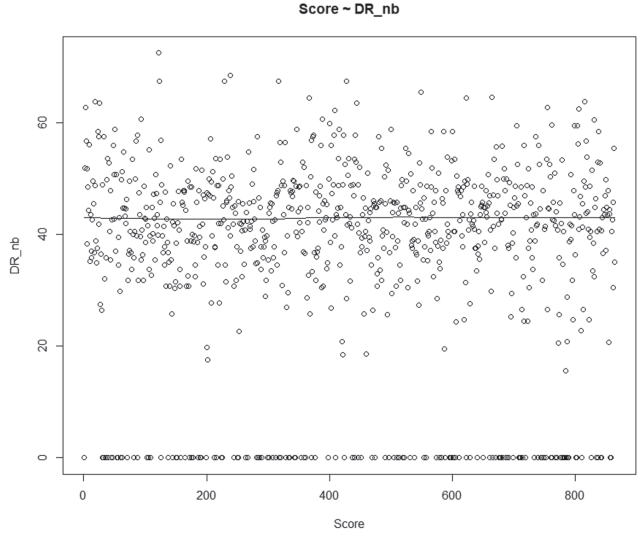
Wyniki regresji liniowej:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.79666 2.36646 15.127 < 2e-16 ***
AvgScore 0.16917 0.05325 3.177 0.00156 **

Residual standard error: 8.859 on 684 degrees of freedom
Multiple R-squared: 0.01454,    Adjusted R-squared: 0.0131
F-statistic: 10.09 on 1 and 684 DF, p-value: 0.001555
```

Otrzymano równanie regresji Y = 0.16917 * X + 35.79666 oraz oba współczynniki są istotne statystycznie. Jednak z uwagi na charakter wykresu, bardzo niski współczynnik korelacji (R^2) oraz niewielki współczynnik (patrz powyżej), hipotezę jakoby graf zbudowany w oparciu o dane z książki adresowej posiadał informację o ryzyku kredytowym należy odrzucić.

Wykonano również podobną analizę, gdzie zmienną wyznaczono jako *default rate* dla danego wierzchołka, czyli jako % sąsiadujących wierzchołków, dla których stwierdzono przesłankę niewywiązania się z zobowiązania. Na potrzeby tej analizy wykorzystano pełny graf G2. Otrzymany wykres (patrz poniżej) ponownie wskazuje na brak realnej zależności. Potwierdza to również wartość współczynnika korelacji.



Wyniki regresji również jednoznacznie wskazują na konieczność odrzucenia stawianej hipotezy.

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.3978 0.7402 49.174 <2e-16 ***
DR – 2.1726 2.4612 – 0.883 0.378

Residual standard error: 17.95 on 862 degrees of freedom
Multiple R-squared: 0.0009032,    Adjusted R-squared: – 0.0002559
F-statistic: 0.7792 on 1 and 862 DF, p-value: 0.3776
```

Dane geolokalizacyjne

W poniższej tabeli przedstawiony został rozkład ocen kredytowych w rozbiciu na poszczególne okręgi geograficzne. Rozbicia dokonano w oparciu o pierwszą cyfrę kodu pocztowego, który każdy klient był zobowiązany podać przy aplikowaniu o kredyt przez platformę FinAi.

Pierwsza cyfra kodu pocztowego	Okrąg	Liczba obserwacji	Średni FinAiRiskScore	DefaultRate	Procent rekordów bez oceny
Brak informacji		499	42,3	15,2%	21,6%
0	**warszawski**	1 970	43,3	12,1%	17,2%
1	**olsztyński**	893	42,1	11,6%	21,4%
2	**lubelski**	1 179	42,0	9,0%	23,6%
3	**krakowski**	1 534	42,4	11,0%	20,1%
4	**katowicki**	2 755	41,5	13,1%	18,3%
5	**wrocławski**	1 657	41,7	14,8%	20,6%
6	**poznański**	2 336	41,2	14,4%	15,5%
7	**szczeciński**	1 162	41,3	13,1%	20,3%
8	**gdański**	2 066	42,0	13,3%	20,0%
9	łódzki	974	42,1	11,4%	18,2%
Razem		17 025	41,9	12,8%	19,1%

Można zauważyć, że średni score przyjmuje porównywalne wartości we wszystkich okręgach. Można na tej podstawie wnioskować, że nadreprezentatywność niskich ocen jest podobna dla wszystkich obszarów. Ponadto, wskazuje to też, że hipoteza mówiąca, że dane geo dyskryminują klientów pod względem ryzyka kredytowego powinna zostać odrzucona (przynajmniej na poziomie powyższych agregatów).

Powyższa informacja oraz informacja o miejscu zarejestrowania pracodawcy, są jedynymi danymi geograficznymi w dostępnym zbiorze. W ramach konkursu, oraz później w procesie, zbierano informacje o miejscach logowania się telefonu komórkowego w trakcie korzystania z aplikacji FinAi. Takie dane niosłyby potencjalnie dużo więcej informacji o zachowaniu klienta. Zgoda na udostępnienie tych danych nie była jednak obligatoryjna, a dostępne są dane od zaledwie 10% wszystkich klientów (patrz poprzedni rozdział). Ze względu na niski odsetek klientów, dla których dane o GPS z telefonów komórkowych są dostępne, wykorzystanie ich nie spowodowałoby znacznego wzrostu predykcyjności modelu. Należy też tutaj podkreślić, że niezagregowane dane geolokalizacyjne (np. GPS) charakteryzują się wielowymiarowością/dużą liczbą stopni swobody. Budowa modeli w oparciu o niewielkie zbiory takich danych może skutkować przeuczeniem modelu i w konsekwencji błędnymi wnioskami. Na podstawie modelu segmentacji geograficznej Polski, do każdego odczytu GPS możemy przypisać jego atrybuty (co do zasady, wektor długości ok 1000 liczb) wyznaczone w oparciu m.in. o bliskość tzw. POI (Point Of Interest), średnie zarobki czy stopień bezrobocia. Tak duże zbiory zmiennych wymagają jednak wielokrotnie większej próby do modelowania niż jest obecnie dostępna.

Podsumowanie

Z uwagi na niewielki zakres informacji typu grafowego, w ocenie autorów raportu jedynym możliwym do wykorzystania grafem jest graf zbudowany w oparciu o dane z książki adresowej. Tak otrzymany graf wydaje się mieć dobre własności, jest dostatecznie gęsty oraz zróżnicowany. Jednakże, głębsza analiza statystyczna wskazuje na brak realnej zależności/korelacji pomiędzy ryzykiem kredytowym sąsiadujących wierzchołków. Oznacza to, że zbudowany graf nie powinien być wyko-

rzystany przy budowie modelu ryzyka. Literatura wskazuje jednak na inne grafy (np. zbudowane w oparciu o połączenia telefoniczne), które niosą za sobą istotną informację.[7] Budowa takich, merytorycznie uzasadnionych grafów niestety nie jest możliwa w oparciu o dostępne dane.

Najważniejsze wnioski z analizy

1) Dane dot. ryzyka kredytowego są niereprezentatywne dla populacji. Średni score BIK w pozyskanym zbiorze danych jest istotnie niższy niż w populacji. Nadreprezentatywność danych od klientów z wysokim ryzykiem kredytowym istotnie utrudniłaby budowę wiarygodnego modelu oraz zmniejszyła jego predykcyjność. Ponad 50% klientów od których pozyskano dane posiada score BIK poniżej punktów odcięcia dla kredytowania standardowo stosowanych w bankach.
2) Alternatywne źródła danych są wypełnione dla niewielkiego odsetka populacji, dodatkowo jakość tych danych jest niezadowalająca. Co do zasady, dane z alternatywnych źródeł (big data) wymagają dużo większych zbiorów, by odkryć w nich ukryte zależności.
3) Zebrane dane nie pozwalają skonstruować grafów czasowych lub chociażby grafów, które niosłyby za sobą istotne informacje o korelacjach ryzyka kredytowego.
4) Wielkość zbioru nie jest wystarczająca do budowy wiarygodnego modelu ryzyka kredytowego oraz zdecydowanie odbiega od oczekiwanego rozmiaru danych dla modeli big data. W szczególności budowa grafów wymaga dużego zbioru danych, tak aby powstały graf odzwierciedlał w sposób wiarygodny prawdziwą sieć społecznościową. Dla porównania, w ramach pracy nad artykułem Óskarsdóttir et al., *The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics*, Applied Soft Computing (2019) cytowanego we wcześniej części artykułu, zbiór danych będący podstawą analiz ryzyka kredytowego składał się z 2 milionów wyciągów bankowych o 1,5 rocznej historii oraz prawie 90 milionów unikalnych numerów telefonów o 5 miesięcznej aktywności telefonicznej.

[7] P. Giudici, B. Hadji-Misheva, A. Spelta, *Network based credit risk models*, „Quality Engineering" 2020, nr 32(2), s. 199–211.

5) Brak możliwości osiągnięcia przewagi w stosunku do modeli funkcjonujących w bankach. FinAi dysponuje wyciągami z kont bankowych, o których wiadomo, że niosą istotne informacje dot. ryzyka kredytowego klienta. Takie same dane są jednak wykorzystywane w bankach, co nie pozwala oczekiwać by model zbudowany w oparciu o dane FinAi miałby istotnie większą predykcyjność lub żeby był innowacyjny w stosunku do istniejących modeli.

Bibliografia

Chen N., Ribeiro B., Chen A., *Financial credit risk assessment: a recent review*, „Artif Intell. Rev." 2016, nr 45.

Emmert-Streib F., Dehmer M., *Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference*, „Mach. Learn. Knowl. Extr." 2019, nr 1.

Giudici P., Hadji-Misheva B., Spelta A., *Network based credit risk models*, „Quality Engineering" 2020, nr 32(2).

James G., Witten D., Hastie T., Tibshirani R., *An Introduction to Statistical Learning*, [w:] *Springer Texts in Statistics*, 2021.

Leskovec J., Backstrom L., Kumar R., Tomkins A., *Microscopic evolution of social networks*, [w:] *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, 2008.

Lopez J.A., Saidenberg M.R., *Evaluating credit risk models*, „Journal of Banking & Finance" 2000, Vol. 24, Issues 1–2.

Montgomery D.C., Peck E.A., Vining G.G., *Introduction to linear regression analysis*, 2021.

Muñoz-Cancino R., Bravo C., Ríos S.A., Graña M., *On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance*, „Expert Systems with Applications" 2023, Vol. 218.

Nguyen T., Wu B., *Fundamentals of Statistics with Fuzzy Data*, „Studies in Fuzziness and Soft Computing" 2006, Vol. 198.

Óskarsdóttir M., Bravo C., Sarraute C., Vanthienen J., Baesens B., *The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics*, „Applied Soft Computing" 2019.

Webber J., *A programmatic introduction to Neo4j*, [w:] *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity (SPLASH '12)*, 2012.
Zamore S., Beisland L.A., Mersland R., *Geographic diversification and credit risk in microfinance*, „Journal of Banking & Finance" 2019, Vol. 109.

Abstrakt

FinAi przeprowadziło projekt budowy innowacyjnego modelu ryzyka kredytowego w oparciu o dane alternatywne takie jak dane z portali społecznościowych (Facebook, Linkedin), czy dane z telefonów komórkowych. Projekt ten był współfinansowany ze środków Europejskiego Funduszu Rozwoju Regionalnego w ramach umowy o dofinansowanie zawartej przez FinAi S.A. z Narodowym Centrum Badań i Rozwoju z siedzibą w Warszawie. Pierwszym etap projektu polegał na zebraniu danych i wykorzystaniu ich do budowy grafu powiązań pomiędzy klientami. Dodatkowo, na podstawie danych zewnętrznych i danych zarządzanych przez FinAi, zbudowano flagi mające na celu nauczenie pod ich nadzorem modelu predykcyjnego wykorzystujące struktury grafowe.

Zbudowany w taki sposób model miał charakteryzować się większą mocą predykcyjną niż standardowe modele stosowane w branży. Badano możliwości budowy modelu ryzyka kredytowego w oparciu o pozyskane dane oraz jakość danych pochodzących ze źródeł alternatywnych. Wykazano m.in., że alternatywne źródła danych są wypełnione dla niewielkiego odsetka populacji, a ich jakość jest niezadowalająca. Wielkość zbioru okazała się niewystarczająca do budowy wiarygodnego modelu ryzyka kredytowego czy osiągnięcia przewagi w stosunku do modeli funkcjonujących w bankach.

Słowa kluczowe: kredyt, ryzyko, dane, modele, wykresy, finanse

Analysis of the possibility of building a credit risk model

Abstract

FinAi has undertaken a project focused on the development of an innovative credit risk model utilizing alternative data sources, such as data from social media platforms (Facebook, LinkedIn) and mobile phone records. This project was co-financed through the European Regional Development Fund under a funding agreement between FinAi S.A. and the National Centre for Research and Development (NCBiR), headquartered in Warsaw. The initial phase of the project

involved the collection of data and their utilization in constructing a network graph of customer relationships. Furthermore, leveraging external data as well as data managed by FinAi, specific indicators were formulated. These indicators were employed under the supervision of experts to train a predictive model that incorporated graph structures.

The model thus constructed was to exhibit a higher predictive capability compared to conventional models commonly employed within the industry. The study explored the feasibility of creating a credit risk model based on the acquired data and assessed the quality of data originating from alternative sources. It was demonstrated that alternative data sources were populated for a small fraction of the population, and their quality has proven unsatisfactory. The scale of the dataset proved inadequate for establishing a robust credit risk model or attaining a competitive advantage over the models in use within banking institutions.

Keywords: credit, risk, data, models, graphs, finance